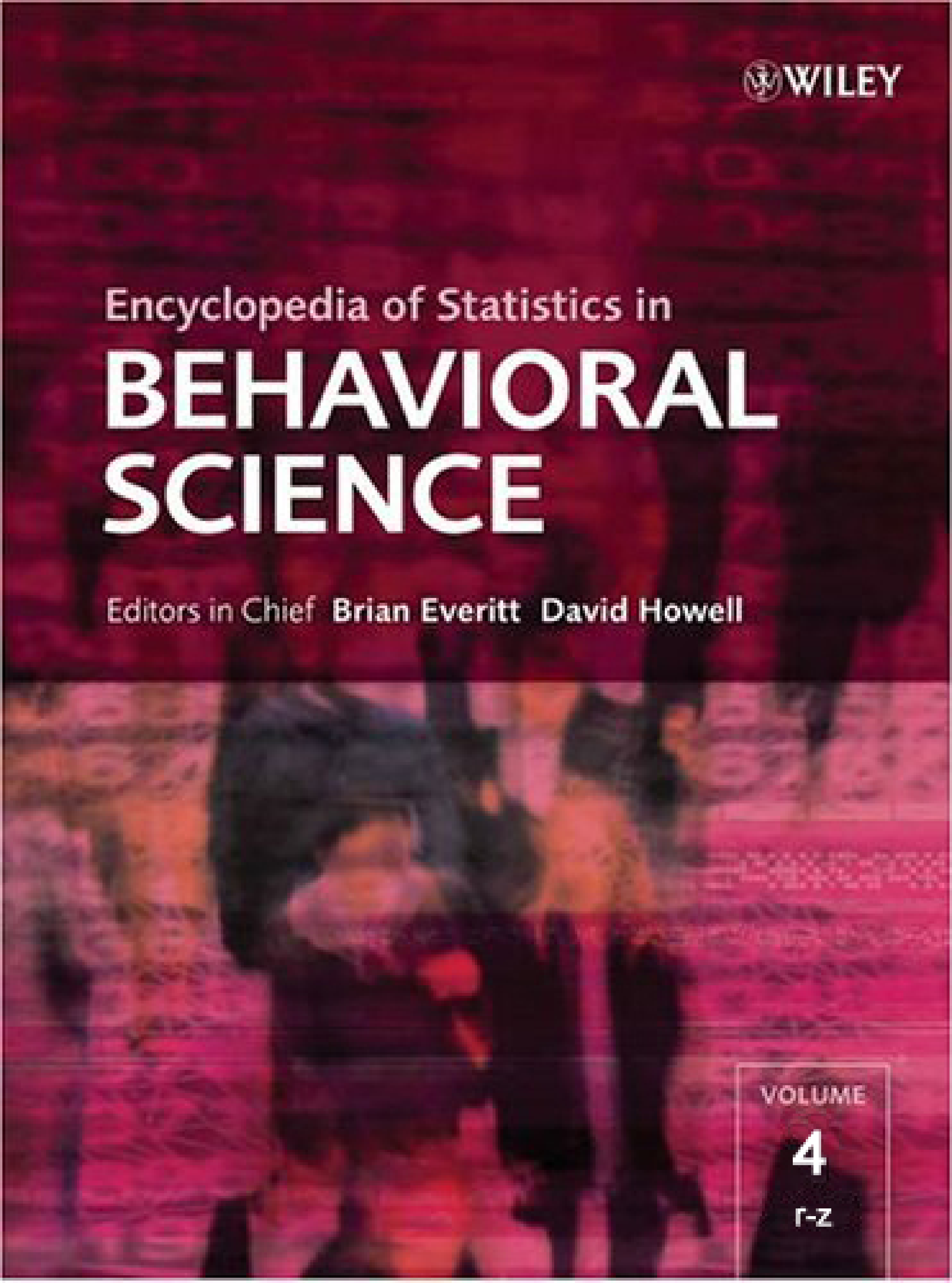


WILEY

Encyclopedia of Statistics in

BEHAVIORAL SCIENCE

Editors in Chief **Brian Everitt** **David Howell**

VOLUME

4

r–z

VOLUME 4

R & Q Analysis

J. Edward Jackson

Volume 4, pp. 1653–1655

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

R & Q Analysis

A number of criteria exist for performing **principal component analysis** on a data matrix. These criteria are referred to as *R*-, *Q*-, *N*-, *M*- and *P-analysis*. The first four of these criteria all involve deviations about column means, row means, or both. The properties of these criteria were given by Okamoto [7]. P-analysis involves the raw data directly.

In principal component analysis, one generally begins with an $n \times p$ data matrix $\boldsymbol{X}$ representing n observations on p variables. Some of the criteria will be illustrated by numerical examples, all using the following data matrix of $n = 4$ observations (rows) on $p = 3$ variables (columns):

$$\begin{array}{rc} & \text{(variables)} \\ \boldsymbol{X} = & \begin{bmatrix} -2 & -2 & -1 \\ 0 & 1 & 0 \\ 2 & 2 & 2 \\ 0 & -1 & -1 \end{bmatrix} \text{(observations)} \\ \text{variable means} & \begin{matrix} 0 & \;\;0 & \;\;0 \end{matrix} \end{array} \tag{1}$$

This matrix will have a rank of three. The variable means have been subtracted to simplify the computations. This example is taken from [6, Chapter 11], which includes more detail in the operations that are to follow.

R-analysis

In principal component analysis, one generally forms some type of $p \times p$ dispersion matrix of the variables from the $n \times p$ data matrix $\boldsymbol{X}$, usually a covariance matrix (*see* **Correlation and Covariance Matrices**) or its related correlation matrix. A set of linear transformations, utilizing the **eigenvectors** of this matrix, is found which will transform the original correlated variables into a new set of variables. These new variables are uncorrelated and are called *principal components*. The values of the transformed data are called *principal component scores*. Further analysis may be carried out on these scores (*see* **Principal Components and Extensions**). (A subset of these transformed variables associated with the larger **eigenvalues** is often retained for this analysis.) This procedure is sometimes referred to as *R-analysis*, and is the most common application of principal component analysis. Similar procedures may also be carried out in some **factor analysis** models.

For example, consider an R-analysis of matrix $\boldsymbol{X}$. Rather than use either covariance or correlation matrices, which would require different divisors for the different examples, the examples will use sums of squares and cross-products matrices to keep the units the same. Then, for R-analysis, this matrix for the variables becomes:

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 8 & 8 & 6 \\ 8 & 10 & 7 \\ 6 & 7 & 6 \end{bmatrix} \tag{2}$$

whose eigenvalues are $l_1 = 22.282$, $l_2 = 1.000$ and $l_3 = 0.718$. The fact that there are three positive eigenvalues indicates that $\boldsymbol{X}'\boldsymbol{X}$ has a rank of three. The unit (i.e., $\boldsymbol{U}'\boldsymbol{U} = \mathbf{I}$) eigenvectors for $\boldsymbol{X}'\boldsymbol{X}$ are:

$$\boldsymbol{U} = \begin{bmatrix} -0.574 & 0.816 & -0.066 \\ -0.654 & -0.408 & 0.636 \\ -0.493 & -0.408 & -0.768 \end{bmatrix} \tag{3}$$

Making a diagonal matrix of the reciprocals of the square roots of the eigenvalues, we have:

$$\boldsymbol{L}^{-0.5} = \begin{bmatrix} 0.212 & 0 & 0 \\ 0 & 1.000 & 0 \\ 0 & 0 & 1.180 \end{bmatrix} \tag{4}$$

and the principal component scores $\boldsymbol{Y} = \boldsymbol{XUL}^{-0.5}$ become:

$$\boldsymbol{Y} = \begin{bmatrix} 0.625 & -0.408 & -0.439 \\ -0.139 & -0.408 & 0.751 \\ -0.729 & 0 & -0.467 \\ 0.243 & 0.816 & 0.156 \end{bmatrix} \tag{5}$$

where each row of this matrix gives the three principal component scores for the corresponding data row in $\boldsymbol{X}$.

Q-analysis

In *Q-analysis*, this process is reversed, and one studies the relationships among the observations rather than the variables. Uses of Q-analysis include the clustering of the individuals in the data set (*see* **Hierarchical Clustering**). Some **multidimensional scaling** techniques are an extension of Q-analysis, and are often used where the data are not homogeneous and require segmentation [4].

In Q-analysis, an $n \times n$ covariance or correlation matrix will be formed for the observations and the eigenvectors, and principal component scores obtained from these. Generally, $n > p$ so that covariance or correlation matrices will not have full rank, and there will be a minimum of $n - p$ zero eigenvalues.

Using the same data matrix from the preceding section, the corresponding sums of squares and cross-products matrix become:

$$\boldsymbol{XX'} = \begin{bmatrix} 9 & -2 & -10 & 3 \\ -2 & 1 & 2 & -1 \\ -10 & 2 & 12 & -4 \\ 3 & -1 & -4 & 2 \end{bmatrix} \qquad (6)$$

with eigenvalues $l_1 = 22.282, l_2 = 1.000, l_3 = 0.718$, and $l_4 = 0$. The first three eigenvalues are identical to those in the Q-analysis. The significance of the fourth eigenvalue being zero is because $\boldsymbol{XX}'$ contains no more information than does $\boldsymbol{X}'\boldsymbol{X}$, and, hence, only has a rank of three.

Although one can obtain four eigenvectors from this matrix, the fourth one is not used as it has no length. The first three eigenvectors are:

$$\boldsymbol{U}* = \begin{bmatrix} 0.625 & -0.408 & -0.439 \\ -0.139 & -0.408 & 0.751 \\ -0.729 & 0 & -0.467 \\ 0.243 & 0.816 & 0.156 \end{bmatrix} \qquad (7)$$

Note that this is the same as the matrix $\boldsymbol{Y}$ of principal scores obtained in the R-analysis above. If one obtains the principal component scores using these eigenvectors, (i.e., $\boldsymbol{Y}^* = \boldsymbol{X}'\boldsymbol{U}^*\boldsymbol{L}^{-0.5}$), it will be found that these principal component scores will be equal to the eigenvectors $\boldsymbol{U}$ of the R-analysis. Therefore, $\boldsymbol{Y}^* = \boldsymbol{U}$, and $\boldsymbol{Y} = \boldsymbol{U}^*$.

N-analysis (Singular Value Decomposition)

With proper scaling or normalization, as has been used in these examples, the eigenvectors of R-analysis become the principal component scores of Q-analysis, and vice versa. These relationships can be extended to *N-analysis* or the *singular value decomposition* [1, 5]. Here, the eigenvalues and vectors as well as the principal component scores for either R- or Q-analysis may be determined directly from the data matrix, namely:

$$\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{L}^{0.5}\boldsymbol{U}' = \boldsymbol{U}^*\boldsymbol{L}^{0.5}\boldsymbol{Y}^{*\prime} \qquad (8)$$

The practical implication of these relationships is that the eigenvalues, eigenvectors, and principal component scores can all be obtained from the data matrix directly in a single operation. In addition, using the relationships above,

$$\boldsymbol{X} = \boldsymbol{U}^*\boldsymbol{L}^{0.5}\boldsymbol{U}' \qquad (9)$$

This relationship is employed in *dual-scaling* techniques, where both variables and observations are being presented simultaneously. Examples of such a technique are the **biplot** [2] and MDPREF [4], which was designed for use with preference data (*see* **Scaling of Preferential Choice**). The graphical presentation of both of these techniques portrays both the variables and the observations on the same plot, one as vectors and the other as points projected against these vectors. These are not to be confused with the so-called 'point–point' plots, which use a different algorithm [6, Section 10.7].

Related Techniques

In addition to R-, Q-, and N-analysis, there are two more criteria, which, though more specialized, should be included for completeness. One of these, *M-analysis*, is used for a data matrix that has been corrected for both its column and row means (so-called *double-centering*). This technique has been used for the two-way **analysis of variance** where there is no estimate of error other than that included in the interaction term. The interaction sum of squares may be obtained directly from double-centered data. M-analysis may then be employed on these data to detect instances of nonadditivity, and/or obtain a better estimate of the true inherent variability [6, Section 13.7]. A version of M-analysis used in multidimensional scaling is a method known as *principal coordinates* [3, 8, 9].

In the antithesis of M-analysis, the original data are not corrected for either variable or observation means. This is referred to as *P-analysis*. In this case, the covariance or correlation matrix is replaced by a matrix made up of the raw sums of squares and cross-products of the data. This is referred to as a *product* or *second moment* matrix and, does not involve deviations about either row or column means. The method of principal components may be carried out on this matrix as well, but some of the usual properties such as rank require slight modifications. This technique is useful for certain additive models, and, for this reason, many of the published applications appear to be in the field of chemistry, particularly with regard to Beer's Law. For some examples, see [6, Section 3.4].

References

[1] Eckart, C. & Young, G. (1936). The approximation of one matrix by another of lower rank, *Psychometrika* **1**, 211–218.

[2] Gabriel, K.R. (1971). The biplot-graphic display of matrices with application to principal component analysis, *Biometrika* **58**, 453–467.

[3] Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis, *Biometrika* **53**, 325–338.

[4] Green, P.E., Carmone Jr, F.J. & Smith, S.M. (1989). *Multidimensional Scaling*, Allyn and Bacon, Needam Heights.

[5] Householder, A.S. & Young, G. (1938). Matrix approximations and latent roots, *American Mathematical Monthly* **45**, 165–171.

[6] Jackson, J.E. (1991). *A User's Guide to Principal Components*, John Wiley & Sons, New York.

[7] Okamoto, M. (1972). Four techniques of principal component analysis, *Journal of the Japanese Statistical Society* **2**, 63–69.

[8] Torgerson, W.S. (1952). Multidimensional scaling I: theory and method, *Psychometrika* **17**, 401–419.

[9] Torgerson, W.S. (1958). *Theory and Methods of Scaling*, John Wiley & Sons, New York.

(*See also* **Multivariate Analysis: Overview**)

J. EDWARD JACKSON

R-squared, Adjusted *R*-squared

Jeremy Miles

Volume 4, pp. 1655–1657

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

R-squared, Adjusted *R*-squared

Many statistical techniques are carried out in order to predict or explain variation in a measure – these include univariate techniques such as linear regression (*see* **Multiple Linear Regression**) and **analysis of variance**, and multivariate techniques, such as multilevel models (*see* **Linear Multilevel Models**), **factor analysis**, and **structural equation models**. A measure of the proportion of variance accounted for in a variable is given by R-squared (*see* **Effect Size Measures**).

The variation in an outcome variable (y) is represented by the sum of squared deviations from the mean, referred to as the total sum of squares (SS_{total}):

$$SS_{\text{total}} = \sum (y - \bar{y})^2 \tag{1}$$

(Note that dividing this value by $N - 1$ gives the variance.)

General linear models (which include regression and ANOVA) work by using least squares estimators; that is, they find parameter estimates and thereby predicted values that account for as much of the variance in the outcome variable as possible – the difference between the predicted value and the actual score for each individual is the **residual**. The sum of squared residuals is the error sum of squares, also known as the within groups sum of squares or residual sum of squares (SS_{error}, SS_{within}, or SS_{residual}). The variation that has been explained by the model is the difference between the total sum of squares and the residual sum of squares, and is called the *between groups sum of squares* or the *regression sum of squares* (SS_{between} or $SS_{\text{regression}}$).

R-squared is given by:

$$R^2 = \frac{SS_{\text{between}}}{SS_{\text{total}}} \tag{2}$$

In a standardized regression equation, where the correlations between variables are known, R^2 is given by:

$$R^2 = b_1 r_{yx_1} + b_2 r_{yx_2} + \cdots + b_k r_{yx_k}, \tag{3}$$

where b_1 represents the standardized regression of y on x, and r_{yx} represents the correlation between y and x.

Where the correlation matrix is known, the formula:

$$R^2_{i.123..k} = 1 - \frac{1}{\mathbf{R}^{-1}_{ii}} \tag{4}$$

may be used, although this involves the inversion of the matrix $\mathbf{R}$, and should really only be attempted by computer (or by those with considerable time on their hands). $\mathbf{R}^{-1}$ is the inverse of the correlation matrix of all variables.

In the simple case of a regression with one predictor, the square of the correlation coefficient (*see* **Pearson Product Moment Correlation**) is equal to R-squared. However, this interpretation of R does not generalize to the case of multiple regression. A second way of considering R is to consider it as the correlation between the values of the outcome predicted by the regression equation and the actual values of the outcome. For this reason, R is sometimes considered to indicate the 'fit' of the model.

Cohen [1] has provided conventional descriptions of effect sizes for R-squared (as well as for other effect size statistics). He defines a small effect as being R^2 equal to 0.02, a medium effect as $R^2 = 0.13$, and a large effect as being $R^2 = 0.26$.

R^2 is a sample estimate of the proportion of variance explained in the outcome variables, and is biased upwards, relative to the population proportion of variance explained. To explain this, imagine we are in the unfortunate situation of having collected random numbers rather than real data (fortunately, we do not need to actually collect any data because we can generate these with a computer). The true (population) correlation of each variable with the outcome is equal to zero; however, thanks to sampling variation, it is very unlikely that any one correlation in our sample will equal zero – although the correlations will be distributed around zero. We have two variables that may be correlated negatively or positively, but to find R^2 we square them, and therefore they all become positive. Every time we add a variable, R^2 will increase; it will never decrease. If we have enough variables, we will find that R^2 is equal to 1.00 – we will have explained all of the variance in our sample, but this will of course tell us nothing about the population. In the long run, values of R^2 in our sample will tend to be higher than values of R^2 in the population (this does not mean that R^2 is always higher in the sample than in the population). In order to correct for this, we use adjusted R^2, calculated using:

$$\text{Adj.}R^2 = 1 - (1 - R^2)\frac{N-1}{N-k-1}, \tag{5}$$

where N is the sample size, and k is the number of predictor variables in the analysis. Smaller values for N, and larger values for k, lead to greater downward adjustment of R^2. In samples taken from a population where the population value of R^2 is 0, the sample R^2 will always be greater than 0. Adjusted R^2 is centered on 0, and hence can become negative; but R^2 is a proportion of variance, and a variance can never be negative (it is the sum of squares) – a negative variance estimate therefore does not make sense and this must be an underestimate.

A useful source of further information is [2].

References

[1] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edition, Lawrence Erlbaum, Mahwah, NJ.

[2] Cohen, J., Cohen, P., West, S.G. & Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd Edition, Lawrence Erlbaum, Mahwah, NJ.

JEREMY MILES

Random Effects and Fixed Effects Fallacy

PHILIP T. SMITH

Volume 4, pp. 1657–1661

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Random Effects and Fixed Effects Fallacy

Introduction

In most psychological experiments, the factors investigated consist of *fixed effects*, that is, all possible levels of each factor are included in the experiment. Clear examples of fixed factors include *Sex* (if both *male* and *female* are included) and *Interference* (in an experiment that manipulated distraction during a task, and included such conditions as *no interference, verbal interference*, and *visual interference*). In contrast, a *random effect* is where the levels of a factor included in the experiment do not exhaust the possible levels of the factor, but consist of a random sample from a population of levels. In most psychological experiments, there is one random effect, *Subjects*: the experimenter does not claim to have tested all the subjects who might have undertaken the task, but hopes that the conclusions from any statistical test apply not only to the people tested but also to the population from which they have been drawn.

Analysis of factorial experiments where there are several fixed-effects factors and one random-effects factor (usually subjects) is the core of an introductory course on **analysis of variance**, and methods and results for all simple designs are well understood. The situation is less clear if there are two random-effects factors in the same experiment, and some researchers have argued that this is the case in experiments involving materials drawn from language. Two artificial datasets have been constructed, shown in Table 1, to illustrate the problem (ignore for the moment the variable *AoA* in Table 1(b); this will be discussed later).

An experimenter is interested in word frequency effects in a categorization task. He selects three high-frequency words, w_1 to w_3, and three low-frequency words, w_4 to w_6. Four subjects, s_1 to s_4, make decisions for all the words. Their decision times are recorded and shown as 'RT' in Table 1. Thus, this is a repeated measures design with the factor *Words* nested within the factor *Frequency*. This is a common design in psycholinguistic experiments, which has been chosen because it is used in many discussions of this topic (e.g., [1, 6, 7]). The actual examples are for illustrative purposes only: It is certainly not being suggested that it is appropriate to design experiments using such small numbers of subjects and stimuli.

Table 1 Two artificial data sets

(a) Small variance between words				(b) Large variance between words				
S	*W*	*Freq*	RT	*S*	*W*	*Freq*	*AoA*	RT
S_1	w_1	*hi*	10	s_1	w_1	*hi*	1	9
S_1	w_2	*hi*	11	s_1	w_2	*hi*	3	11
S_1	w_3	*hi*	12	s_1	w_3	*hi*	4	13
S_1	w_4	*lo*	13	s_1	w_4	*lo*	2	12
S_1	w_5	*lo*	14	s_1	w_5	*lo*	4	14
S_1	w_6	*lo*	15	s_1	w_6	*lo*	6	16
S_2	w_1	*hi*	10	s_2	w_1	*hi*	1	9
S_2	w_2	*hi*	12	s_2	w_2	*hi*	3	12
S_2	w_3	*hi*	12	s_2	w_3	*hi*	4	13
S_2	w_4	*lo*	14	s_2	w_4	*lo*	2	13
S_2	w_5	*lo*	14	s_2	w_5	*lo*	4	14
S_2	w_6	*lo*	15	s_2	w_6	*lo*	6	16
S_3	w_1	*hi*	11	s_3	w_1	*hi*	1	10
S_3	w_2	*hi*	11	s_3	w_2	*hi*	3	11
S_3	w_3	*hi*	12	s_3	w_3	*hi*	4	13
S_3	w_4	*lo*	13	s_3	w_4	*lo*	2	12
S_3	w_5	*lo*	13	s_3	w_5	*lo*	4	13
S_3	w_6	*lo*	15	s_3	w_6	*lo*	6	16
S_4	w_1	*hi*	10	s_4	w_1	*hi*	1	9
S_4	w_2	*hi*	10	s_4	w_2	*hi*	3	10
S_4	w_3	*hi*	11	s_4	w_3	*hi*	4	12
S_4	w_4	*lo*	13	s_4	w_4	*lo*	2	12
S_4	w_5	*lo*	15	s_4	w_5	*lo*	4	15
S_4	w_6	*lo*	15	s_4	w_6	*lo*	6	16

One possible analysis is to treat *Frequency* and *Words* as fixed-effect factors. Such an analysis uses the corresponding interactions with *Subjects* as the appropriate error terms. Analyzed in this way, *Frequency* turns out to be significant ($F(1, 3) = 80.53$, $P < 0.01$) for both datasets in Table 1. There is something disturbing about obtaining the same results for both datasets: it is true that the means for Frequency are the same in both sets (hi Frequency has a mean RT of 11.00, lo Frequency has a mean RT of 14.08, in both sets), but the effect looks more consistent in dataset (a), where all the hi Frequency words have lower means than all the lo Frequency words, than in dataset (b), where there is much more variation. In an immensely influential paper, Herb Clark [1] suggested that the analyses we have just described are invalid, because *Words* is being treated as a fixed effect: other words could have been selected that meet our selection criteria (in the present case, to be of hi or lo *Frequency*) so *Words* should be treated as a random effect. Treating *Words* as a fixed effect is, according to Clark, the *Language-as-Fixed-Effect Fallacy*.

Statistical Methods for Dealing with Two Random Effects in the Same Experiment

F_1 and F_2

Treating *Words* as a fixed effect, as we did in the previous paragraph, is equivalent to averaging across *Words*, and carrying out an ANOVA based purely on *Subjects* as a random effect. This is known as a *by-subjects* analysis and the F values derived from it usually carry the suffix '1'; so, the above analyses have shown that $F_1(1, 3) = 80.53$, $P < 0.01$. An alternative analysis would be, for each word, to average across subjects and carry out an ANOVA based purely on Words as a random effect. This is known as a *by-materials* analysis (the phrase *by-items* is sometimes used). The F values derived from this analysis usually carry the suffix '2'. In the present case, the dataset (a) by-materials analysis yields $F_2(1, 4) = 21.39$, $P = 0.01$, which is significant, whereas dataset (b) by-materials analysis yields $F_2(1, 4) = 4.33$, $P = 0.106$, clearly nonsignificant. This accords with our informal inspection of Table 1, which shows a more consistent frequency effect for dataset (a) than for dataset (b).

It is sometimes said that F_1 assesses the extent to which the experimental results may generalize to new samples of subjects, and that F_2 assesses the extent to which the results will generalize to new samples of words. These statements are not quite accurate: neither F_1 nor F_2 are pure assessments of the presence of an effect. The standard procedure in an ANOVA is to estimate the variance due to an effect, via its mean square (*MS*), and compare this mean square with other mean squares in the analysis to assess significance. Using formulas, to be found in many textbooks (e.g. [13, 14]), the analysis of the Table 1 data as a three-factor experiment where *Freq* and *W* are treated as fixed effects and *S* is treated as a random effect yields the following equations:

$$\mathrm{E}(MS_{\mathrm{Freq}}) = \sigma_e^2 + q\sigma^2_{\mathrm{Freq}\times S} + nq\sigma^2_{\mathrm{Freq}} \tag{1}$$

$$\mathrm{E}(MS_{W(\mathrm{Freq})}) = \sigma_e^2 + \sigma^2_{W(\mathrm{Freq})\times S} + n\sigma^2_{W(\mathrm{Freq})} \tag{2}$$

$$\mathrm{E}(MS_S) = \sigma_e^2 + pq\sigma_S^2 \tag{3}$$

$$\mathrm{E}(MS_{\mathrm{Freq}\times S}) = \sigma_e^2 + q\sigma^2_{\mathrm{Freq}\times S} \tag{4}$$

$$\mathrm{E}(MS_{W(\mathrm{Freq})\times S}) = \sigma_e^2 + \sigma^2_{W(\mathrm{Freq})\times S} \tag{5}$$

E means expected or theoretical value, σ_A^2 refers the variance attributable to A, e is random error, n is the number of *Subjects*, p the number of levels of *Frequency*, and q the number of *Words*. The researcher rarely needs to know the precise detail of these equations, but we include them to make an important point about the choice of error terms in hypothesis testing: (1), $\mathrm{E}(MS_{\mathrm{Freq}})$, differs from (4) only in having a term referring to the variance in the data attributable to the *Frequency* factor, so we can test for whether *Frequency* makes a nonzero contribution to the variance by comparing (1) with (4), more precisely by dividing the estimate of variance described in (1) (MS_{Freq}) by the estimate of variance described in (4) ($MS_{\mathrm{Freq}\times S}$). In other words, $MS_{\mathrm{Freq}\times S}$ is the appropriate error term for testing for the effect of *Frequency*.

If *Frequency* is treated as a fixed effect, but *Words* and *Subjects* are treated as random effects, the variance equations change, as shown below.

$$\begin{aligned}\mathrm{E}(MS_{\mathrm{Freq}}) = \sigma_e^2 + \sigma^2_{W(\mathrm{Freq})\times S} + q\sigma^2_{\mathrm{Freq}\times S} \\ + n\sigma^2_{W(\mathrm{Freq})} + nq\sigma^2_{\mathrm{Freq}}\end{aligned} \tag{6}$$

$$\mathrm{E}(MS_{W(\mathrm{Freq})}) = \sigma_e^2 + \sigma^2_{W(\mathrm{Freq})\times S} + n\sigma^2_{W(\mathrm{Freq})} \tag{7}$$

$$\mathrm{E}(MS_S) = \sigma_e^2 + \sigma^2_{W(\mathrm{Freq})\times S} + pq\sigma_S^2 \tag{8}$$

$$\mathrm{E}(MS_{\mathrm{Freq}\times S}) = \sigma_e^2 + \sigma^2_{W(\mathrm{Freq})\times S} + q\sigma^2_{\mathrm{Freq}\times S} \tag{9}$$

$$\mathrm{E}(MS_{W(\mathrm{Freq})\times S}) = \sigma_e^2 + \sigma^2_{W(\mathrm{Freq})\times S} \tag{10}$$

The most important change is the difference between (1) and (6). Unlike (1), (6) contains terms involving the factor *Words*. Equation (6) is telling us that some of the variance in calculating MS_{Freq} is due to the possible variation of the effect for different selections of words (in contrast, (1) assumes all relevant words that have been included in the experiment, so different selections are not possible). This means that F_1 (derived from dividing (6) by (9)) is contaminated by *Words*: a significant F_1 could arise from a fortuitous selection of words. Similarly, F_2 (derived by dividing (6) by (7) is contaminated by *Subjects*: a significant F_2 could arise from a fortuitous selection of subjects. By themselves, F_1 and F_2 are insufficient to solve the problem. (This is not to say that significant F_1 or F_2 are never worth reporting: for example, practical constraints when testing very young children or patients might mean the number of stimuli that can be used is too small to permit an F_2 test of reasonable power: here F_1 would be worth reporting, though the researcher needs to accept that generalization to other word sets has yet to be established.)

Quasi-F Ratios

There are ratios that can be derived from (6) to (10), which have the desired property that the numerator differs from the denominator by only a term involving σ^2_{Freq}. Two possibilities are

$$F' = \frac{(MS_{\text{Freq}} + MS_{W(\text{Freq})\times S})}{(MS_{W(\text{Freq})} + MS_{\text{Freq}\times S})} \qquad (11)$$

and

$$F'' = \frac{MS_{\text{Freq}}}{(MS_{W(\text{Freq})} + MS_{\text{Freq}\times S} - MS_{W(\text{Freq})\times S})} \qquad (12)$$

These Fs are called *Quasi-F ratios*, reflecting the fact that they are similar to standard F ratios, but, because they are not the simple ratio of two mean squares, their distribution is only approximated by the standard F distribution. Winer [13, pp. 377–378] and Winer et al. [14, pp. 374–377], give the basic formulas for degrees of freedom in (11) and (12), derived from Satterthwaite [10]. It is doubtful that the reader will ever need to calculate such expressions by hand, so these are not given here. SPSS (Version **12**) uses (12) to calculate quasi-F ratios: for example, if the data in Table 1 are entered into SPSS in exactly the form shown in the table, and if *Freq* is entered as a Fixed Factor and S and W are entered as Random Factors, and Type I SS are used, then SPSS suggests there is a significant effect of *Frequency* for dataset (a) ($F(1, 4.916) = 18.42$, $P < 0.01$), but not for dataset (b) ($F(1, 4.249) = 4.20$, $P = 0.106$). If S and W are truly random effects, then this is the correct statistical method. Many authorities, for example, [1, 6, 13], prefer (11) to (12) since (12) may on occasion lead to a negative number (*see* [5]).

Min F′

The method outlined in section 'Quasi-F Ratios' can be cumbersome: for any experiment with realistic numbers of subjects and items, data entry into SPSS or similar packages can be very time-consuming, and if there are missing data (quite common in reaction-time experiments), additional corrections need to be made. A short cut is to calculate min F', which, as its name suggests, is an estimate of F' that falls slightly below true F'. The formula is as follows:

$$\min F' = \frac{F_1 \cdot F_2}{(F_1 + F_2)} \qquad (13)$$

The degrees of freedom for the numerator of the F ratio remains unchanged $(p - 1)$, and the degrees of freedom for the denominator is given by

$$\text{df} = \frac{(F_1^2 + F_2^2)}{(F_1^2/\text{df}_2 + F_2^2/\text{df}_1)} \qquad (14)$$

where df_1 is the error degrees of freedom for F_1 and df_2 is the error degrees of freedom for F_2. For the data in Table 1, dataset (a) has min $F'(1, 5.86) = 16.90$, $P < 0.01$, and dataset (b) has min $F'(1, 4.42) = 4.11$, $P = 0.11$, all values being close to the true F' values shown in the previous section.

Best practice, then, is that when you conduct an ANOVA with two random-effects factors, use (11) or (12) if you can, but if you cannot, (13) and (14) provide an adequate approximation.

Critique

Clark's paper had an enormous impact. From 1975 onward, researchers publishing in leading psycholinguistic journals, such as *Journal of Verbal Learning and Verbal Behavior* (now known as *Journal of Memory and Language*) accepted that something needed to be done in addition to a *by-subjects* analysis, and at first min F' was the preferred solution. Nowadays, min F' is hardly ever reported, but it is very common to report F_1 and F_2, concluding that the overall result is significant if both F_1 and F_2 are significant. As Raaijmakers et al. [8] have correctly pointed out, this latter practice is wrong: simulations [4] have shown that using the simultaneous significance of F_1 and F_2 to reject the null hypothesis of no effect can lead to serious inflation of Type I errors. It has also been claimed [4] and [12] that min F' is too conservative, but this conservatism is quite small and the procedure quite robust to modest violations of the standard ANOVA assumptions [7, 9].

One reason for min F''s falling into disuse is its absence from textbooks psychology researchers are likely to read: Only one graduate level textbook has been found, by Allen Edwards [3], that gives a full treatment to the topic. Jackson & Brashers [6] give a very useful short overview, though their description of calculations using statistical packages is inevitably out of date. Another reason for not calculating min F' is that we have become so cosseted by statistics packages that do all our calculations for us, that we are not prepared to work out min F' and its fiddly degrees of freedom by hand. (If you belong to this camp, there is a website (`www.pallier.org/ressources/MinF/compminf.htm`) that will work out min F', its degrees of freedom and its significance for you.)

The main area of contention in the application of these statistical methods is whether materials should be treated as a *random* effect. This point was picked up by early critics of Clark [2, 12]: researchers do not select words at random, and indeed often go to considerable lengths to select words with appropriate properties ('It has often seemed to me that workers in this field counterbalance and constrain word lists to such an extreme that there may in fact *be* no other lists possible within the current English language.' [2, p.262]). Counterbalancing (for example, arranging that half the subjects receive word set A in condition C_1 and word set B in condition C_2, and the other half of the subjects receive word set B in condition C_1 and word set A in condition C_2) would enable a by-subjects analysis to be carried out uncontaminated by effects of materials [8]. Counterbalancing, however, is not possible when the effect of interest involves intrinsic differences between words, as in the examples in Table 1: different words must be used if we want to examine word frequency effects.

Constraining word lists, that is selecting sets of words that are matched on variables we know to influence the task they are to be used in, is often a sensible procedure: it makes little sense to select words that differ widely in their frequency of occurrence in the language when we know that frequency often has a substantial influence on performance. The trouble with such procedures is that matching can never be perfect because there are too many variables that influence performance. One danger of using a constrained word set, which appears to give good results in an experiment, is that the experimenter, and others who wish to replicate or extend his or her work, are tempted to use the same set of words in subsequent experiments. Such a procedure may be capitalizing on some as yet undetected idiosyncratic feature of the word set, and new sets of words should be used wherever possible. A further drawback is that results from constrained word sets can be generalized only to other word sets that have been constrained in a similar manner.

An alternative to using highly constrained lists is to include influential variables in the statistical model used to analyze the data (e.g., [11]). For example, another variable known to influence performance with words is *Age of Acquisition* (*AoA*). A researcher dissatisfied with the large variance displayed by different words in Table 1(b), and believing *AoA* was not adequately controlled, might add *AoA* as a covariate to the analysis, still treating *Subjects* and *Words* as random effects. This now transforms the previously nonsignificant quasi-F ratio for *Frequency* to a significant one ($F(1, 3.470) = 18.80$, $P < 0.05$).

A final remark is that many of the comments about treating *Words* as a random effect apply to treating *Subjects* as a random effect. In psycholinguistic experiments, we frequently reject subjects of low IQ or whose first language is not English, and, when we are testing older populations, we generally deal with a self-selected sample of above average individuals. In an area such as morphology, which is often not taught formally in schools, there may be considerable individual differences in the way morphemically complex words are represented. All of these examples suggest that **a**ttempts to model an individual *subject's* knowledge and abilities, for example, via covariates in **analyses of covariance**, could be just as important as modeling the distinct properties of individual *words*.

References

[1] Clark, H.H. (1973). The language-as-fixed-effect fallacy: a critique of language statistics in psychological research, *Journal of Verbal Learning and Verbal Behavior* **12**, 335–359.

[2] Clark, H.H., Cohen, J., Keith Smith, J.E. & Keppel, G. (1976). Discussion of Wike and Church's comments, *Journal of Verbal Learning and Verbal Behavior* **15**, 257–266.

[3] Edwards, A.L. (1985). *Experimental Design in Psychological Research*, 2nd Edition, Harper & Row, New York.

[4] Forster, K.I. & Dickinson, R.G. (1976). More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F_1, F_2, F', and min F', *Journal of Verbal Learning and Verbal Behavior* **15**, 135–142.

[5] Gaylor, D.W. & Hopper, F.N. (1969). Estimating the degrees of freedom for linear combinations of means squares by Satterthwaite's formula, *Technometrics* **11**, 691–706.

[6] Jackson, S. & Brashers, D.E. (1994). *Random Factors in ANOVA*, Sage Publications, Thousand Oaks.

[7] Maxwell, S.F. & Bray, J.H. (1986). Robustness of the quasi F statistic to violations of sphericity, *Psychological Bulletin* **99**, 416–421.

[8] Raaijmakers, J.G.W., Schrijnemakers, J.M.C. & Gremmen, F. (1999). How to deal with "The Language-as-Fixed-Effect Fallacy": common misconceptions and alternative solutions, *Journal of Memory and Language* **41**, 416–426.

[9] Santa, J.L., Miller, J.J. & Shaw, M.L. (1979). Using quasi F to prevent alpha inflation due to stimulus variation, *Psychological Bulletin* **86**, 37–46.

[10] Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components, *Biometrics Bulletin* **2**, 110–114.

[11] Smith, P.T. (1988). How to conduct experiments with morphologically complex words, *Linguistics* **26**, 699–714.

[12] Wike, E.L. & Church, J.D. (1976). Comments on Clark's "The Language-as-Fixed-Effect Fallacy", *Journal of Verbal Learning and Verbal Behavior* **15**, 249–255.

[13] Winer, B.J. (1971). *Statistical Principles in Experimental Design*, 2nd Edition, McGraw-Hill Kogakusha, Tokyo.

[14] Winer, B.J., Brown, D.R. & Michels, K.M. (1991). *Statistical Principles in Experimental Design*, 3rd Edition, McGraw-Hill, New York.

PHILIP T. SMITH

Random Effects in Multivariate Linear Models: Prediction

NICHOLAS T. LONGFORD

Volume 4, pp. 1661–1665

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Random Effects in Multivariate Linear Models: Prediction

Random effects are a standard device for representing the differences among observational or experimental units that are or can be perceived as having been drawn from a well-defined population. Such units may be subjects (individuals), their organizations (businesses, classrooms, families, administrative units, or teams), or settings within individuals (time periods, academic subjects, sets of tasks, and the like). The units are associated with random effects because they are incidental to the principal goal of the analysis – to make inferences about an *a priori* specified population of individuals, geographical areas, conditions, or other factors (contexts).

In modeling, random effects have the advantage of parsimonious representation, that a large number of quantities are summarized by a few parameters that describe their distribution. When the random effects have a (univariate) normal distribution, it is described completely by a single variance; the mean is usually absorbed in the regression part of the model. The fixed-effects counterparts of these models are **analysis of covariance** (ANCOVA) models, in which each effect is represented by one or a set of parameters.

In the resampling perspective, fixed effects are used for factors that are not altered in hypothetical replications. Typically, factors with few levels (categories), such as experimental conditions or treatments, which are the focus of the inference, are regarded as fixed. In contrast, a different set of random effects is realized in each hypothetical replication; the replications share only the distribution of the effects. A logical inconsistency arises when the analyzed sample is an enumeration. For example, when the districts of a country are associated with random effects, a replication would yield a different set of district-level effects. Yet, a more natural replication, considered in sampling theory in particular, keeps the effects fixed for each district – the same set of districts would be realized. This conflict is resolved by a reference to a superpopulation, arguing that inferences are desired for a domain *like* the one analyzed, and in each replication a different domain is realized with a different division into districts.

A constructive way of addressing the issue of fixed *versus* random effects is by admitting that incorrect models may be useful for inference. That is, the effects are fixed, but it is advantageous to treat them in inference as random. Apart from a compact description of the collection of units, by their (estimated) distribution or its parameters, random effects enable estimation of unit-specific quantities that is more efficient than in the **maximum likelihood** or **least squares** for standard fixed effects ANCOVA. The efficiency is achieved by *borrowing strength* across the units [9]. Its theoretical antecedent is the work on shrinkage estimation [5] and its application to small-area statistics [3].

Borrowing strength can be motivated by the following general example. If the units are similar, then the pooled (domain-related) estimator of the quantity of interest θ may be more efficient for the corresponding unit-related quantity θ_j, because the squared bias $(\theta_j - \theta)^2$ is much smaller than the sampling variance $\mathrm{var}(\hat{\theta}_j)$ of the unbiased estimator of θ_j. Instead of selecting the domain estimator $\hat{\theta}$ or the unbiased large-variance estimator $\hat{\theta}_j$, these two estimators are combined,

$$\tilde{\theta}_j = (1 - b_j)\hat{\theta}_j + b_j\hat{\theta}, \qquad (1)$$

with a constant b_j, or its estimator $\hat{b}_j$, for which the combination has some optimal properties, such as minimum mean squared error (MSE). The combination (*composition*) $\tilde{\theta}_j$ can be interpreted as exploiting the similarity of the units. The gains are quite dramatic when the units are similar and $\mathrm{var}(\hat{\theta}_j) \gg \mathrm{var}(\hat{\theta})$. That occurs when there are many (aggregate-level) units j and most of them are represented in the dataset by only a few observations each.

Inference about the individual units is usually secondary to studying the population as a whole. Nevertheless, interest in units on their own may arise as a result of an inspection of the data or their analysis that aimed originally at some population features. Model diagnostics are a notable example of this.

Estimation of random effects is usually referred to as *prediction*, to avoid the terminological conflict of ‘estimating random variables’, a contradiction in terms if taken literally. The task of prediction is to define a function of the data that is, in a well-defined sense, as close to the target as possible. With random effects, the target does not appear to be stationary.

In fact, *the realizations* of random variables, or quantities that are fixed across replications but regarded as random in the model, are estimated. Alternatively, the prediction can be described as estimating the quantity of interest *given* that it is fixed in the replications; the corresponding quantities for the other units are assumed to vary across replications. The properties of such an estimator (predictor) should be assessed *conditionally* on the realized value of the target.

Random-effects models involving normality and linearity are greatly preferred because of their analytical tractability, easier interpretation, and conceptual proximity to ordinary regression (*see* **Multiple Linear Regression**) and ANCOVA. We discuss first the prediction of random effects with the model

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \delta_j + \boldsymbol{\varepsilon}_j, \tag{2}$$

where $\mathbf{y}_j$ is the $n_j \times 1$ vector of (univariate) outcomes for (aggregate or level-2) unit $j = 1, \ldots, N_2$, $\mathbf{X}_j$ is the regression design matrix for unit j, $\boldsymbol{\beta}$ the vector of regression parameters, δ_j the random effect for unit j, and $\boldsymbol{\varepsilon}_j$ the vector of its elementary-level (residual) terms, or deviations (*see* **Variance Components**). The random terms δ_j and $\boldsymbol{\varepsilon}_j$ are mutually independent, with respective centered normal distributions $\mathcal{N}(0, \sigma_2^2)$ and $\mathcal{N}(0, \sigma^2\mathbf{I}_{n_j})$; $\mathbf{I}$ is the identity matrix of the size given in the subscript.

The model in (2) can be interpreted as a set of related regressions for the level-2 units. They have all the regression coefficients in common, except for the intercept $\beta_0 + \delta_j$. The regressions are parallel. The obvious generalization allows any regression coefficients to vary, in analogy with introducing group-by-covariate interactions in ANCOVA. Thus, a subset of the covariates in $\mathbf{X}$ is *associated with variation*. The corresponding submatrix of $\mathbf{X}_j$ is denoted by $\mathbf{Z}_j$ (its dimensions are $n_j \times r$), and the model is

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\delta}_j + \boldsymbol{\varepsilon}_j, \tag{3}$$

where $\boldsymbol{\delta}_j \sim \mathcal{N}(\mathbf{0}_r, \boldsymbol{\Sigma})$, independently ($\mathbf{0}_r$ is the $r \times 1$ column vector of zeros), [4] and [7]. In agreement with the ANCOVA conventions, $\mathbf{Z}_j$ usually contains the intercept column $\mathbf{1}_{n_j}$. Variables that are constant within groups j can be included in $\mathbf{Z}$, but the interpretation in terms of varying regressions does not apply to them because the within-group regressions are not identified for them.

The random effects $\boldsymbol{\delta}_j$ are estimated from their conditional expectations given the outcomes $\mathbf{y}$. The matrices $\mathbf{X}_j$ and $\mathbf{Z}_j$ are assumed to be known, or are conditioned on, even when they depend on the sampling or the data-generation process. Assuming that the parameters $\boldsymbol{\beta}$, σ^2 and those involved in $\boldsymbol{\Sigma}$ are known, the conditional distribution of $\boldsymbol{\delta}_j$ is normal,

$$(\boldsymbol{\delta}_j|\mathbf{y}, \boldsymbol{\theta}) \sim \mathcal{N}\left\{\frac{1}{\sigma^2}\boldsymbol{\Sigma}\mathbf{G}_j^{-1}\mathbf{Z}_j^\top\mathbf{e}_j,\ \boldsymbol{\Sigma}\mathbf{G}_j^{-1}\right\}, \tag{4}$$

where $\mathbf{G}_j = \mathbf{I}_r + \sigma^{-2}\mathbf{Z}_j^\top\mathbf{Z}_j\boldsymbol{\Sigma}$ and $\mathbf{e}_j = \mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}$. The vector $\boldsymbol{\delta}_j$ is predicted by its (naively) estimated conditional expectation. The univariate version of this estimator, for the model in (2), corresponding to $\mathbf{Z}_j = \mathbf{1}_{n_j}$, is

$$\tilde{\delta}_j = \frac{n_j\hat{\omega}}{1 + n_j\hat{\omega}}\overline{e}_j, \tag{5}$$

where $\hat{\omega}$ is an estimate of the variance ratio $\omega = \sigma_2^2/\sigma^2$ (σ_2^2 is the univariate version of $\boldsymbol{\Sigma}$) and $\overline{e}_j = (\mathbf{y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}})^\top\mathbf{1}_{n_j}/n_j$ is the average residual in unit j. Full or restricted maximum likelihood estimation (MLE) can be applied for $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2$ and $\hat{\boldsymbol{\Sigma}}$ or $\hat{\sigma}_2^2$. In general, restricted MLE is preferred because the estimators of σ^2 and σ_2^2 are unbiased. As the absence of bias is not maintained by nonlinear transformations, this preference has a poor foundation for predicting $\boldsymbol{\delta}_j$. Thus, even the restricted MLE of ω, the ratio of two unbiased estimators, $\hat{\omega} = \hat{\sigma}_2^2/\hat{\sigma}^2$, is biased. The bias of $\hat{\omega}$ can be corrected, but it does not lead to an unbiased estimator $\tilde{\delta}_j$, because $1/(1 + n_j\omega)$ is estimated with bias.

These arguments should not be interpreted as claiming superiority of full MLE over restricted MLE, merely that no bias is not the right goal to aim for. Absence of bias is not a suitable criterion for estimation in general; minimum MSE, combining bias and sampling variance, is more appropriate. Prediction of random effects is an outstanding example of successful (efficient) biased estimation. Bias, its presence and magnitude, depend on the resampling (replication) perspective adopted. Reference [10] presents a viewpoint in which the estimators we consider are unbiased. In fact, the terminology 'best linear unbiased predictor' (BLUP) is commonly used, and is appropriate when different units are realized in replications. Indeed, $\mathrm{E}(\tilde{\delta}_j - \delta_j|\omega) = 0$ when the expectation is taken both over sampling within units j and over the population of units j, because $\mathrm{E}(\tilde{\delta}_j) = \mathrm{E}(\delta_j) = 0$. In contrast,

$$\mathrm{E}(\tilde{\delta}_j|\delta_j;\omega) = \frac{n_j\omega}{1+n_j\omega}\delta_j, \qquad (6)$$

so $\tilde{\delta}_j$ is *conditionally* biased. The conditional properties of $\tilde{\delta}_j$ are usually more relevant. The return, sometimes quite generous, for the bias is reduced sampling variance.

When δ_j are regarded as fixed effects, their least-squares estimator (and also MLE) is $\hat{\delta}_j = \overline{e}_j$. As $\tilde{\delta}_j = q_j\hat{\delta}_j$, where $q_j = n_j\hat{\omega}/(1+n_j\hat{\omega}) < 1$, $\tilde{\delta}_j$ can be interpreted as a *shrinkage* estimator and q_j, or more appropriately $1 - q_j = 1/(1+n_j\hat{\omega})$, as a shrinkage coefficient. The coefficient q_j is an increasing function of both the sample size n_j and $\hat{\omega}$; more shrinkage takes place (q_j is smaller) for units with smaller sample sizes and when $\hat{\omega}$ is smaller. That is, for units with small samples, greater weight is assigned to the overall domain (its average residual $\overline{e} = (n_1\overline{e}_1 + \cdots + n_{N_2}\overline{e}_{N_2})/(n_1 + \cdots + n_{N_2})$ vanishes either completely or approximately) – the resulting bias is preferred to the substantial sampling variance of $\overline{e}_j$. Small ω indicates that the units are very similar, so the average residuals $\overline{e}_j$ differ from zero mainly as a result of sampling variation; shrinkage then enables estimation more efficient than by $\overline{e}_j$. The same principles apply to multivariate random effects, although the discussion is not as simple and the motivation less obvious.

Shrinkage estimation is a form of empirical Bayes estimation (*see* **Bayesian Statistics**). In Bayes estimation, a prior distribution is imposed on the model parameters, in our case, the random effects δ_j. In empirical Bayes estimation, the prior distribution is derived (estimated) from the same data to which it is subsequently applied; see [8] for examples in educational measurement. Thus, the prior distribution of δ_j is $\mathcal{N}(0, \sigma_2^2)$, and the posterior, with σ_2^2 and other model parameters replaced by their estimates, is $\mathcal{N}\{\tilde{\delta}_j,\ \hat{\sigma}_2^2/(1+n_j\hat{\omega})\}$.

Somewhat loosely, $\hat{\sigma}_2^2/(1+n_j\hat{\omega})$ is quoted as the sampling variance of $\tilde{\delta}_j$, and its square root as the standard error. This is incorrect on several counts. First, these are *estimators* of the sampling variance or standard error. Next, they estimate the sampling variance for a particular replication scheme (with δ_j as a random effect) assuming that the model parameters $\boldsymbol{\beta}$, σ^2, and σ_2^2 are known. For large-scale data, the uncertainty about $\boldsymbol{\beta}$ and σ^2 can be ignored, because their estimation is based on many degrees of freedom. However, σ_2^2 is estimated with at most N_2 degrees of freedom, one for each level-2 unit, and N_2 is much smaller than the elementary-level sample size $n = n_1 + \cdots + n_{N_2}$. Two factors complicate the analytical treatment of this problem; $\tilde{\delta}_j$ is a nonlinear function of ω and the precision of $\hat{\omega}$ depends on the (unknown) ω. The next element of 'incorrectness' of using $\hat{\sigma}_2^2/(1+n_j\hat{\omega})$ is that it refers to an 'average' unit j with the sample size n_j. Conditioning on the sample size is a common practice, even when the sampling design does not guarantee a fixed sample size n_j for the unit. But the sample size can be regarded as auxiliary information, so the conditioning on it is justified. However, $\tilde{\delta}_j$ is biased, so we should be concerned with its MSE:

$$\begin{aligned}\mathrm{MSE}(\tilde{\delta}_j;\delta_j) &= \mathrm{E}\{(\tilde{\delta}_j - \delta_j)^2|\delta_j\} \\ &= \frac{(n_j\omega)^2}{(1+n_j\omega)^2}\frac{\sigma^2}{n_j} + \frac{\delta_j^2}{(1+n_j\omega)^2},\end{aligned}$$

assuming that $\boldsymbol{\beta}$, σ^2, and ω are known and the sample-average residual $\overline{e}$ vanishes. Rather inconveniently, the MSE depends on the target δ_j itself. The conditional variance is obtained by replacing δ_j^2 with its expectation σ_2^2.

Thus, $\hat{\sigma}_2^2/(1+n_j\hat{\omega})$ is an estimator of the *expected* MSE (eMSE), where the expectation is taken over the distribution of the random effects $\delta_{j'}$, $j' = 1, \ldots, N_2$. It underestimates $\mathrm{eMSE}(\tilde{\delta}_j;\delta_j)$ because some elements of uncertainty are ignored. As averaging is applied, it is not a particularly good estimator of $\mathrm{MSE}(\tilde{\delta}_j;\delta_j)$. It is sometimes referred to as the comparative standard error [4]. The MSE can be estimated more efficiently by bootstrap [2] or by framing the problem in terms of incomplete information [1], and representing the uncertainty by plausible values of the unknown parameters, using the principles of **multiple imputation**, [11] and [12]. Approximations by various expansions are not very effective because they depend on the variances that have to be estimated.

The normal distribution setting is unusual by its analytical tractability, facilitated by the property that the normality and homoscedasticity are maintained by conditioning. These advantages are foregone with **generalized mixed linear models**. They are an extension of **generalized linear models** that parallels the extension of linear regression to random coefficient models:

$$g\{\mathrm{E}(\mathbf{y}_j|\boldsymbol{\delta}_j)\} = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\delta}_j, \tag{7}$$

where g is a monotone function, called the *link function*, and the assumptions about all the other terms are the same as for the random coefficient model that corresponds to normality and identity link (*see* **Generalized Linear Mixed Models**). The conditional distribution of $\mathbf{y}_j$ given $\boldsymbol{\delta}_j$ has to be specified; extensive theory is developed for the case when this distribution belongs to the exponential family (*see* **Generalized Linear Models (GLM)**). The realization of $\boldsymbol{\delta}_j$ is estimated, in analogy with BLUP in the normality case, by estimating its conditional expectation given the data and parameter estimates. In general, the integral in this expectation is not tractable, and we have to resort to numerical approximations. These are computationally manageable for one- or two-dimensional $\boldsymbol{\delta}_j$, especially if the number of units j in the domain, or for which estimation of $\boldsymbol{\delta}_j$ is desired, is not excessive. Some approximations avoid the integration altogether, but are not very precise, especially when the between-unit variance σ_2^2 (or $\boldsymbol{\Sigma}$) is substantial. The key to such methods is an analytical approximation to the (marginal) likelihood, based on Laplace transformation or quasilikelihood. These methods have been developed to their apparently logical conclusion in the h-likelihood [6]. Fitting models by h-likelihood involves no integration, the random effects can be predicted without any extensive computing, and more recent work by the authors is concerned with joint modeling of location and variation structures and detailed diagnostics.

In principle, any model can be extended to its random-effects version by assuming that a separate model applies to each (aggregate) unit, and specifying how the parameters vary across the units. No variation (identical within-unit models) is a special case in such a model formulation. Modeling is then concerned with the associations in an average or typical unit, and with variation within and across the units. Unit-level random effects represent the deviation of the model for a given unit from the average unit. Units can differ in all aspects imaginable, including their level of variation, so random effects need not be associated only with regression or location, but can be considered also for variation and any other model features.

References

[1] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood for incomplete data via the EM algorithm, *Journal of the Royal Statistical Society* **39**, 1–38.

[2] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.

[3] Fay, R.E. & Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data, *Journal of the American Statistical Association* **74**, 269–277.

[4] Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd Edition, Edward Arnold, London.

[5] James, W. & Stein, C. (1961). Estimation with quadratic loss, *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, pp. 361–379.

[6] Lee, Y. & Nelder, J.A. (1996). Hierarchical generalized linear models, *Journal of the Royal Statistical Society* **58**, 619–678.

[7] Longford, N.T. (1993). *Random Coefficient Models*, Oxford University Press, Oxford.

[8] Longford, N.T. (1995). *Models for Uncertainty in Educational Testing*, Springer-Verlag, New York.

[9] Morris, C.N. (1983). Parametric empirical Bayes inference: theory and applications, *Journal of the American Statistical Association* **78**, 55–65.

[10] Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects, *Statistical Science* **6**, 15–51.

[11] Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley and Sons, New York.

[12] Rubin, D.B. (1996). Multiple imputation after 18+ years, *Journal of the American Statistical Association* **91**, 473–489.

Nicholas T. Longford

Random Forests

Adele Cutler

Volume 4, pp. 1665–1667

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Random Forests

Random forests were introduced by Leo Breiman in 2001 [1], and can be thought of as **bagging classification and regression trees (CART)**, for which each node is split using a random subset of the variables, and not pruning. More explicitly, we select a bootstrap sample (*see* **Bootstrap Inference**) from the data and fit a binary decision tree to the bootstrap sample. To fit the tree, we split nodes by randomly choosing a small number of variables and finding the best split on these variables only. For example, in a classification problem for which we have, say, 100 input variables, we might choose 10 variables at random, independently, each time a split is to be made. For every distinct split on these 10 variables, we compute some measure of node purity, such as the *gini index* [2], and we select the split that optimizes this measure. Cases on each side of the split form new nodes in the tree, and the splitting procedure is repeated until all the nodes are pure. We typically grow the tree until it is large, with no pruning, and then combine the trees as with bagging (averaging for regression and voting for classification).

To illustrate, we use the R [8] function *randomForest* to fit a classifier to the data in Figure 1. The classification boundary and the data are given in Figure 1(a). In Figures 1(b), 1(c), and 1(d), the shading intensity indicates the weighted vote for class 1. As more trees are included, the nonlinear boundary is estimated more accurately.

Studies (e.g., [4]) show that random forests are about as accurate as *support vector machines* [6] and **boosting** [3], but unlike these competitors, random forests are interpretable using several quantities that we can compute from the forest.

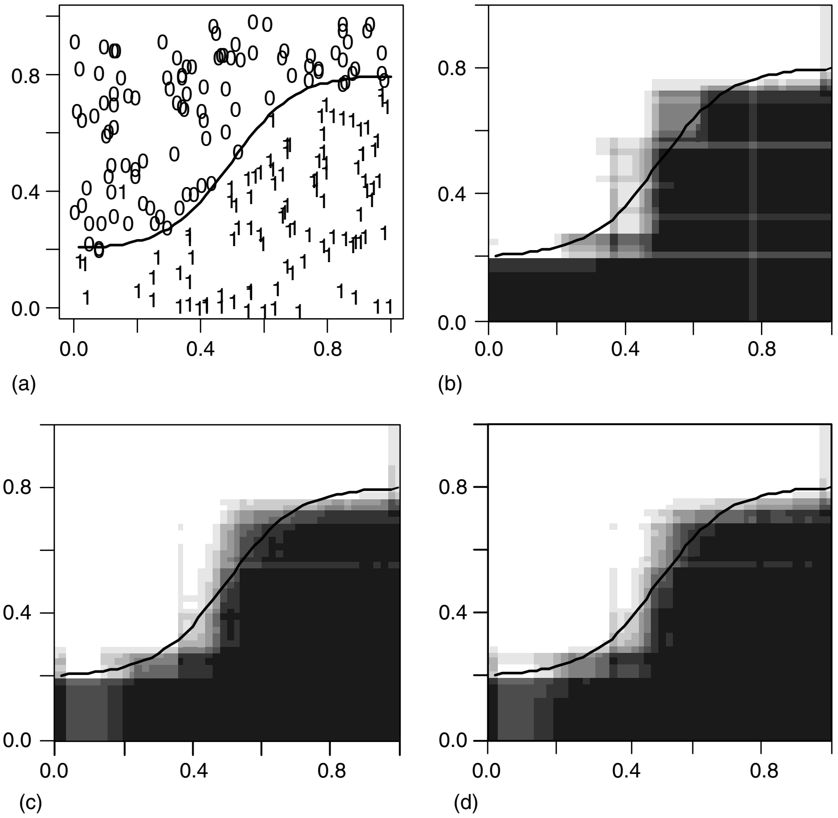


Figure 1 (a) Data and underlying function; (b) random forests, 10 trees; (c) random forests, 100 trees; and (d) random forests, 400 trees

The first such quantity is *variable importance*. We compute variable importance by considering the cases that are left out of a bootstrap sample ('out-of-bag'). If we are interested in the importance of variable 3, for example, we randomly permute variable 3 in the out-of-bag data. Then, using the tree that we obtained from the bootstrap sample, we subtract the prediction accuracy for the permuted out-of-bag data from that for the original out-of-bag data. If variable 3 is important, the permuted out-of-bag data will have lower prediction accuracy than the original out-of-bag data, so the difference will be positive. This measure of variable importance for variable 3 is averaged over all the bootstrap samples, and the procedure is repeated for each of the other input variables.

A second important quantity for interpreting random forests is the proximity matrix (*see* **Proximity Measures**). The proximity between any two cases is computed by looking at how often they end up in the same terminal node. These quantities, suitably standardized, can be used in a proximity-based clustering (*see* **Hierarchical Clustering**) or **multidimensional scaling** procedure to give insight about the data structure. For example, we might pick out subgroups of cases that almost always stay together in the trees, or **outliers** that are almost always alone in a terminal node.

Random forests can be used in a clustering context by thinking of the observed data as class 1, creating a synthetic second class, and using the random forests' classifier. The synthetic second class is created by randomly permuting the values of each input variable. The proximities from random forests can be used in a proximity-based clustering procedure.

More details on random forests can be obtained from `http://stat-www.berkeley.edu/users/breiman/RandomForests`, along with freely available software.

References

[1] Breiman, L. (2001). Random Forests, *Machine Learning* **45**(1), 5–32.

[2] Breiman, L., Friedman, J.H., Olshen, R. & Stone, C. (1983). *Classification and Regression Trees*, Wadsworth.

[3] Freund, Y. & Schapire, R. (1996). Experiments with a new boosting algorithm, in *Machine Learning: Proceedings of the Thirteenth International* Morgan Kauffman, San Francisco, CA. *Conference*, 148–156.

[4] Meyer, D., Leisch, F., Hornik, K. (2002). Adaptive information systems and modeling in economics and management science, Report No 78., Benchmarking support vector machines. `http://www.wu-wien.ac.at/am/`

[5] R Development Core Team, (2004). *R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing*, Vienna. `www.R-project.org`.

[6] Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*, Springer.

ADELE CUTLER

Random Walks

Donald Laming

Volume 4, pp. 1667–1669

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Random Walks

Suppose there is an accident on a motorway that reduces traffic past that point to a cautious one-vehicle-at-a-time on the hard shoulder. A mile or so in advance, the traffic is channelled into two lanes and, as you reach that two-lane restriction, you find yourself level with a Rolls-Royce. Thereafter, sometimes the Rolls-Royce edges forward a length, sometimes it is your turn, and Figure 1 shows how the relative position of the two cars develops. To your chagrin, the Rolls-Royce has crept ahead; will you ever catch up?

A random walk is the cumulative sum of a series of independent and identically distributed random variables, $\sum_1^n X_i$, and Figure 1 is a simple example (as also is the forward progress of either car). As a vehicle somewhere ahead edges past the scene of the accident, you or the Rolls-Royce (but not both) can move forward one car length – one step in the random walk. Assuming that the two lanes feed equally and at random past the accident, then the relative positions of the two cars is analogous to the difference in the numbers of heads and tails in a sequence of tosses of a fair coin. If the sequence continues for long enough, it is certain that the numbers of heads and tails will, at some point, be equal, but the mean wait is infinite. That is to say, you will most probably pass the point of restriction before you draw level with the Rolls-Royce.

Each step in Figure 1 could, of course, be itself the sum of a number of independent and identically distributed random variables. Suppose I let a drop of black ink fall into a glass of water. The ink slowly diffuses throughout the water, driven by Brownian motion. Suppose Figure 1 represents the drift of a notional particle of black ink on the left–right axis. Each step can then be split into an arbitrary number of substeps. If the substeps are independent and identically distributed, then the random walk is actually a random process, unfolding in continuous time. But such a decomposition (into an arbitrary number of independent and identically distributed substeps) is possible only if the distribution of each step is *infinitely divisible*. Amongst well-known probability distributions, the normal, the Poisson, and the gamma (or chi-squared) (*see* **Catalogue of Probability Density Functions**) distributions are infinitely divisible. In addition, a compound Poisson distribution (in which each Poisson event is itself a random variable) is infinitely divisible with respect to its Poisson parameter, so the class of infinitely divisible distributions is very broad. But amongst these possibilities only the normal (or Wiener) process is continuous with respect to the spatial dimension; all the other random processes contain jumps.

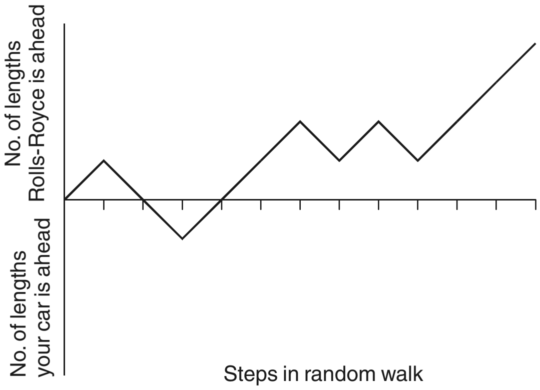


Figure 1 Relative position of two cars on a motorway

Interest in random walks began in the 18th century with gamblers wanting to know the chances of their being ruined. Suppose the game is 'absolutely fair' (*see* **Martingales**), so that the probabilities of winning and losing are equal. The paths in Figure 2 trace out different gamblers' cumulative wins and losses. If a gambler should ever lose his or her entire fortune, he or she will have nothing left to gamble with, and this is represented by his or her time line (path) in Figure 2 descending to the axis at 0. This poses the following question: What is the probability that the random walk will ever fall below a certain specified value (the gambler's entire fortune)? In more detail, how does the random walk behave if we delete all those gamblers who are ruined from the point of their bankruptcy onwards (broken lines in Figure 2)? For the binomial walk of Figure 1, this is a simple problem. Clearly, any walk that strays below the horizontal boundary must be struck out from that point onwards. But we must also delete those continuations of such a random walk that happen to rise upwards from the boundary as well as those that continue below. This may be achieved by introducing a mirror-image source (dotted lines in Figure 2) below the boundary. The fortunes of a gambler who has escaped ruin may be represented by the difference

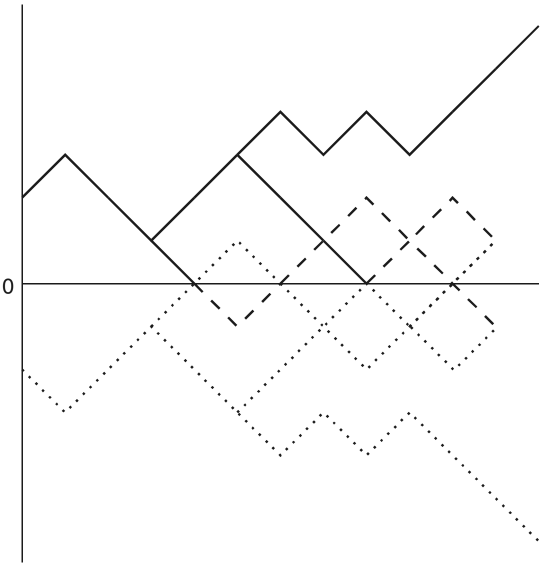


Figure 2 Calculation of a random walk with one absorbing barrier by deletion of an image process

between these two random processes, the original and the mirror image, above the horizontal boundary [2].

The mirror-image technique works equally for random processes in continuous time; and a simple modification to the argument adjusts it to the case where the random walk drifts up or down (the gambler is playing with skilled card sharps and loses more often than he wins) [1, p. 50]. If the basic random process is a (normal) Wiener process, the time taken to reach the boundary (to be ruined) is given by the *Wald distribution*

$$f(t) = \left[\frac{a}{\sqrt{(2\pi\sigma^2 t^3)}}\right] \exp\left\{\frac{-(a-\mu t)^2}{2\sigma^2 t}\right\}, \quad (1)$$

where a is the distance to the boundary (the gambler's fortune) and μ and σ^2 the rates at which the mean and variance of the random process increase per unit time. Schwartz [5] has used this distribution, convolved with an exponential to provide the characteristic long tail, as a model for simple reaction times.

Random walks have also been proposed as models for two-choice reaction times [6]. There are now two boundaries placed on either side of a starting point (Figure 3), one corresponding to each response. The response depends on which boundary is reached first; the reaction time is, of course, the time taken to reach that boundary. So this model bids to determine both choice of response and latency from a common process.

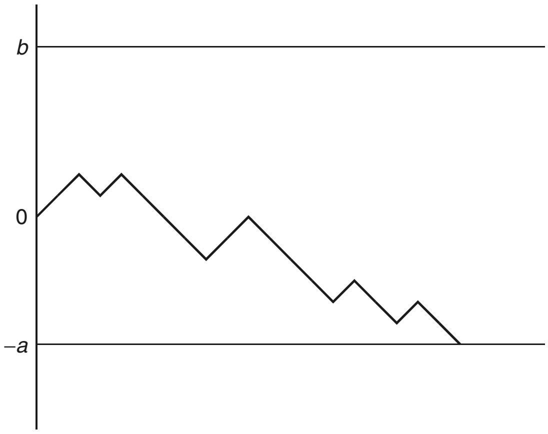


Figure 3 Random walk with two absorbing barriers

The distribution of reaction time now depends on the location of two absorbing boundaries as well as the statistical properties of the processes representing the two alternative stimuli. Unfortunately, the simple argument using mirror-image sources is not practicable now; it generates an infinite series of sources stretching out beyond both boundaries. But the response probabilities and the moments of the reaction time distributions can be readily obtained from Wald's identity [7, p. 160]. Let $\varphi(\omega)$ be the characteristic function of the random process per unit time. Let Z be the terminal value of the process on one or the other boundary. Then

$$E\left\{\frac{\exp(Z\omega)}{[\varphi(\omega)]^t}\right\} = 1. \quad (2)$$

Even so, the development of this model is not simple except in two special cases.

There are two distinct random processes involved, representing the two alternative stimuli, A and B. The special cases are distinguished by the relationship between these two processes.

1. Suppose that $f_B(x) = e^x f_A(x)$. The variable x is then the probability ratio between the two alternatives, and the reaction time model may be interpreted as a sequential probability ratio test between the two stimuli [3].

2. Suppose instead that $f_B(x) = f_A(-x)$. The two processes are now mirror images of each other [4].

The nature of reaction times is such that a random walk has an intuitive appeal as a possible model; this is especially so with two-choice reaction times in which both probabilities of error and reaction times derive from a common source. But the relationship of experimental data to model predictions has not provided great grounds for confidence; it has typically been disappointing.

References

[1] Bartlett, M.S. (1955). *An Introduction to Stochastic Processes*, Cambridge University Press, Cambridge.

[2] Chandrasekhar, S. (1943). Stochastic problems in physics and astronomy, *Reviews of Modern Physics* **15**, 1.

[3] Laming, D. (1968). *Information Theory of Choice-Reaction Times*, Academic Press, London.

[4] Link, S.W. (1975). The relative judgment theory of two choice response time, *Journal of Mathematical Psychology* **12**, 114–135.

[5] Schwartz, W. (2001). The ex-Wald distribution as a descriptive model of response times, *Behavior Research Methods, Instruments & Computers* **33**, 457–469.

[6] Stone, M. (1960). Models for choice-reaction time, *Psychometrika* **25**, 251–260.

[7] Wald, A. (1947). *Sequential Analysis*, Wiley, New York.

DONALD LAMING

Randomization

Robert J. Boik

Volume 4, pp. 1669–1674

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Randomization

Introduction

Randomization is the intentional use of a random process either in the design phase (prerandomization) or in the analysis phase (postrandomization) of an investigation. Prerandomization includes random selection, random assignment, and randomized response methods. Postrandomization includes randomized decision rules and randomization-based inference methods such as permutation tests and **bootstrap methods**. The focus of this article is on random selection, random assignment, and their relationship to randomization-based inference. Definitions of simple random assignment and selection along with a brief discussion of the origins of randomization are given in this section. Justifications for and criticisms of randomization are given in the next section.

Simple random selection refers to any process that selects a sample of size n without replacement from a population of size $N > n$, such that each of the $N!/[n!(N-n)!]$ possible samples is equally likely to be selected. Simple random assignment refers to any process that assigns one of t treatments to each of N subjects, such that each of the $N!/(n_1!n_2!, \ldots n_t!)$ possible assignments in which treatment j is assigned to n_j subjects is equally likely. A method for performing random assignment, as well as a warning about a faulty assignment method, is described in [26]. For convenience, the term randomization will be used in this article to refer either to random selection or to random assignment. Details on **randomized response** methods can be found in [67]. Randomized decision rules are described in [11, §1.5] and [40, §9.1].

The prevailing use of random assignment in experimentation owes much to **R. A. Fisher**, who in 1926 [17], apparently was the first to use the term randomization [7]. Random assignment, however, had been used much earlier, particularly in behavioral science research. Richet [49] used random assignment in an 1884 telepathy experiment in which subjects guessed the suit of a card. The Society for Psychical Research in London was receptive to using random assignment and by 1912 it was being used in parapsychological research at American universities. Random assignment also was used as early as 1885 in psychophysics experiments [42], but the procedure was not as readily accepted here as it was in parapsychological research. Fisher made random assignment a formal component of experimental design, and he introduced the method of permutation tests (*see* **Permutation Based Inference**) [18]. Pitman [44–46] provided a theoretical framework for permutation tests and extended them to tests on correlations and to **analysis of variance**.

Random sampling also has origins within social science research. Kiaer [34] proposed in 1897 that a representative (purposive) sample rather than a census be used to gather data about an existing population. The proposal met substantial opposition, in part because the method lacked a theoretical framework. Bowley, in 1906 [3], showed how the **central limit theorem** could be used to assess the accuracy of population estimates based on simple random samples. Work on the theory of stratified random sampling (*see* **Stratification**) had begun by 1923 [64], but as late as 1926 random sampling and purposive sampling were still treated on equal grounds [4]. It was **Neyman** [41] who provided the theoretical framework for random sampling and set the course for future research. In this landmark paper, Neyman introduced the randomization-based sampling distribution, described the theory of stratified random sampling with optimal allocation, and developed the theory of **confidence intervals**. Additional discussions on the history of randomization can be found in [16, 25, 29, 43, 47, 48, 57, 59], and [61].

Why Randomize?

It might seem unnecessary even to ask the question posed by the title of this section. For most behavioral scientists, the issue was resolved in the sophomore-level experimental psychology course. It is apparent, however, that not all statisticians take this course. At least two articles with the title 'Why Randomize' are widely cited [23, 32]. A third article asked 'Must We Randomize Our Experiment'? and answered – 'sometimes' [2]. A fourth asked 'Experimental Randomization: Who Needs It?' and answered – 'nobody' [27]. Additional articles that have addressed the question posed in this section include [5, 6, 9, 24, 36, 38, 55, 60–63, 65], and [66]. Arguments for randomization as well as selected criticisms are summarized next.

To Ensure that Linear Estimators are Unbiased and Consistent. A statistic is unbiased for estimating a parameter if the mean of its **sampling distribution** is equal to the value of the parameter, regardless of which value the parameter might have. A statistic is consistent for a parameter if the statistic converges in probability to the value of the parameter as sample size increases.

A sampling distribution for a statistic can be generated by means of postrandomization, provided that prerandomization was employed for data collection. Sampling distributions generated in this manner are called *randomization distributions*. In an observational study, random sampling is sufficient to ensure that sample means are unbiased and consistent for population means. Random sampling also ensures that the empirical cumulative distribution function is unbiased and consistent for the population cumulative distribution function and this, in turn, is the basis of the bootstrap. Likewise, random assignment together with unit-treatment additivity ensures that differences between means of treatment groups are unbiased and consistent estimators of the true treatment differences, even if important explanatory variables have been omitted from the model.

Unbiasedness is generally thought of as a desirable property, but an estimator may be quite useful without being unbiased. First, biased estimators are sometimes superior to unbiased estimators with respect to mean squared error (variance plus squared bias). Second, unbiasedness can be criticized as an artificial advantage because it is based on averaging over treatment assignments or subject selections that could have been but were not observed [5]. Averaging over data that were not observed violates the likelihood principle, which states that inferences should be based solely on the likelihood function given the observed data. A brief introduction to the issues regarding the likelihood principle can be found in [50] (*see* **Maximum Likelihood Estimation**).

To Justify Randomization-based Inference. One of the major contributions of Neyman [41] was to introduce the randomization distribution for survey sampling. Randomization distributions provide a basis for assessing the accuracy of an estimator (e.g., standard error) as well as a framework for constructing confidence intervals.

Randomization distributions based on designed experiments are particularly useful for testing sharp null hypotheses. For example, suppose that treatment and control conditions are randomly assigned to subjects and that administration of the treatment would have an additive effect, say δ, for each subject. The permutation test, based on the randomization distribution of treatment versus control means, provides an exact test of the hypothesis $\delta = 0$. Furthermore, a confidence interval for δ can be obtained as the set of all values δ_0 for which H_0: $\delta = \delta_0$ is not rejected using the permutation test. In general, randomization plus subject-treatment additivity eliminates the need to know the exact process that generated the data.

It has been suggested that permutation tests are meaningful even when treatments are not randomly assigned [9, 12–15, 19, 39, 54]. The resulting P values might have descriptive value but without random assignment they do not have inferential value. In particular, they cannot be used to make inferences about causation [24].

Randomization-based inference has been criticized because it violates the conditionality principle [1, 10, 27]. This principle states that inference should be made conditional on the values of ancillary statistics; that is, statistics whose distributions do not depend on the parameter of interest. The outcome of randomization is ancillary. Accordingly, to obey the conditionality principle, inference must be made conditional on the observed treatment assignment or sample selection. Postrandomizations do not yield additional information. This criticism of randomization-based inference loses much of its force if the model that generated the data is unknown. Having a valid inference procedure in the absence of distributional knowledge is appealing and appears to outweigh the cost of violating the conditionality principle.

To Justify Normal-theory Tests. Kempthorne [30, 31] showed that if treatments are randomly assigned to subjects and unit-treatment additivity holds, then the conventional F Test is justified even in the absence of normality. The randomization distribution of the test statistic under H_0 is closely approximated by the central F distribution. Accordingly, the conventional F Test can be viewed as an approximation to the randomization test. In addition, this result implies that the choice of the linear model is not *ad hoc*, but follows from randomization together with unit-treatment additivity.

To Protect Against Subjective Biases of the Investigator. Randomization ensures that treatment assignment is not affected by conscious or subconscious biases of the experimenter. This justification has been criticized on the grounds that an investigator who cannot be trusted without randomization does not become trustworthy by using randomization [27]. The issue, however, is not only about trust. Even the most trustworthy experimenter could have an unintentional influence on the outcome of the study. The existence of unintentional experimenter effects is well documented [53] and there is little reason to believe that purposive selection or purposive assignment would be immune from such effects.

To Elucidate Causation. It has been argued that the scientific (as opposed to statistical) purpose of randomization is to elucidate causation [33, 62]. Accurate inferences about causality are difficult to make because experimental subjects are heterogeneous and this variability can lead to bias. Random assignment guards against pretreatment differences between groups on recorded variables (overt bias) as well as on unobserved variables (hidden bias).

In a designed experiment, the causal effect of one treatment relative to a second treatment for a specific subject can be defined as the difference between the responses under the two treatments. Most often in practice, however, only one treatment can be administered to a specific subject. In the counterfactual approach (*see* **Counterfactual Reasoning**), the causal effect is taken to be the difference between potential responses that would be observed under the two treatments, assuming subject-treatment additivity and no **carryover effects** [28, 52, 55, 56]. These treatment effects cannot be observed, but random assignment of treatments is sufficient to ensure that differences between the sample means of the treatment groups are unbiased and consistent for the true causal effects. The counterfactual approach has been criticized on the grounds that assumptions such as subject-treatment additivity cannot be empirically verified [8, 20, 63].

'Randomization is rather like insurance [22].' It protects one against biases, but it does not guarantee that treatment groups will be free of pretreatment differences. It guarantees only that over the long run, average pretreatment differences are zero. Nonetheless, even after a bad random assignment, it is unlikely that treatment contrasts will be completely confounded with pretreatment differences. Accordingly, if treatment groups are found to differ on important variables after random assignment, then covariate adjustment still can be used (*see* **Analysis of Covariance**). This role of reducing the probability of confounding is not limited to frequentist analyses; it also is relevant in Bayesian analyses [37].

Under certain conditions, causality can be inferred without random assignment. In particular, if experimental units are homogeneous, then random assignment is unnecessary [55]. Also, random assignment is unnecessary (but may still be useful) under covariate sufficiency. Covariate sufficiency is said to exist if all covariates that affect the response are observed [63]. Under covariate sufficiency, hidden bias is nonexistent and adjustment for differences among the observed covariates is sufficient to remove overt bias, even if treatments are not randomly assigned. Causal inferences from structural equation models fitted to observational data as in [58] implicitly require covariate sufficiency. Without this condition, inferences are limited to ruling out causal patterns that are inconsistent with the observed data. Causal inference in observational studies without covariance sufficiency is substantially more difficult and is sensitive to model misspecification [68, 69].

Furthermore, random assignment is unnecessary for making causal inferences from experiments whenever treatments are assigned solely on the basis of observed covariates, even if the exact assignment mechanism is unknown [21]. The conditional probability of treatment assignment given the observed covariates is known as the **propensity score** [52]. If treatments are assigned solely on the basis of observed covariates, then adjustment for differences among the propensity scores is sufficient to remove bias. One way to ensure that treatment assignment is solely a function of observed covariates is to randomly assign treatments to subjects, possibly after blocking on one or more covariates.

To Enhance Robustness of Inferences. Proponents of optimal experimental design and sampling recommend that treatments be purposively assigned and that subjects be purposively selected using rational judgments rather than random processes. The advantage of purposive assignment and selection is that they can yield estimators that are more efficient than those based on randomization [35, 51]. If the presumed model is not correct, however, then the resulting inferences may be faulty. Randomization guards against making incorrect inferences due to model misspecification.

For example, consider the problem of constructing a regression function for a response, Y, given a single explanatory variable, X. If the regression function is known to be linear, then the variance of the least squares slope estimator is minimized by selecting observations for which half of the Xs are at the maximum and half of the Xs are at the minimum value (*see* **Optimal Design for Categorical Variables**). If the true regression function is not linear, however, then the resulting inference will be incorrect and the investigator will be unable to perform diagnostic checks on the model. In contrast, if (X, Y) pairs are randomly selected then standard regression diagnostic plots can be used to detect model misspecification and to guide selection of a more appropriate model.

References

[1] Basu, D. (1980). Randomization analysis of experimental data: the fisher randomization test, (with discussion), *Journal of the American Statistical Association* **75**, 575–595.

[2] Box, G.E.P. (1990). Must we randomize our experiment? *Journal of Quality Engineering* **2**, 497–502.

[3] Bowley, A.L. (1906). Address to the economic science and statistics section of the British Association for the Advancement of Science, 1906, *Journal of the Royal Statistical Society* **69**, 540–558.

[4] Bowley, A.L. (1926). Measurement of the precision attained in sampling, *Bulletin of the International Statistics Institute* **22** (Supplement to Liv. 1), 6–62.

[5] Bunke, H. & Bunke, O. (1978). Randomization. Pro and contra, *Mathematische Operationsforschung und Statistik, Series Statistics* **9**, 607–623.

[6] Cox, D.R. (1986). Some general aspects of the theory of statistics, *International Statistical Review* **54**, 117–126.

[7] David, H.A. (1995). First (?) occurrence of common terms in mathematical statistics, *The American Statistician* **49**, 121–133.

[8] Dawid, A.P. (2000). Causal inference without counterfactuals (with discussion), *Journal of the American Statistical Association* **95**, 407–448.

[9] Easterling, R.G. (1975). Randomization and statistical inference, *Communications in Statistics* **4**, 723–735.

[10] Efron, B. (1978). Controversies in the foundations of statistics, *The American Mathematical Monthly* **85**, 231–246.

[11] Ferguson, T.S. (1967). *Mathematical Statistics A Decision Theoretic Approach*, Academic Press, New York.

[12] Finch, P.D. (1979). Description and analogy in the practice of statistics, *Biometrika* **67**, 195–208.

[13] Finch, P.D. (1981). A descriptive interpretation of standard error, *Australian Journal of Statistics* **23**, 296–299.

[14] Finch, P.D. (1982). Comments on the role of randomization in model-based inference, *Australian Journal of Statistics* **24**, 146–147.

[15] Finch, P. (1986). Randomization – I, in *Encyclopedia of Statistical Sciences*, Vol. 7, S. Kotz & N.L. Johnson, eds, John Wiley & Sons, New York, pp. 516–519.

[16] Fienberg, S.E. & Tanur, J.M. (1987). Experimental and sampling structures: parallels diverging and meeting, *International Statistical Review* **55**, 75–96.

[17] Fisher, R.A. (1926). The arrangement of field experiments, *Journal of the Ministry of Agriculture of Great Britain* **33**, 700–725.

[18] Fisher, R.A. (1935). *The Design of Experiments*, Oliver & Boyd, Edinburgh.

[19] Freedman, D.A. & Lane, D. (1983). Significance testing in a nonstochastic setting, *Journal of Business and Economic Statistics* **1**, 292–298.

[20] Gadbury, G.L. (2001). Randomization inference and bias of standard errors, *The American Statistician* **55**, 310–313.

[21] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2003). *Bayesian Data Analysis*, 2nd Edition. Chapman & Hall/CRC, Boca Raton.

[22] Gibson, B. (2000). How many Rs are there in statistics? *Teaching Statistics* **22**, 22–25.

[23] Greenberg, B.G. (1951). Why randomize? *Biometrics* **7**, 309–322.

[24] Greenland, S. (1990). Randomization, statistics, and casual inference, *Epidemiology* **1**, 421–429.

[25] Hacking, I. (1988). Telepathy: origins of randomization in experimental design, *Isis* **79**, 427–451.

[26] Hader, R.J. (1973). An improper method of randomization in experimental design, *The American Statistician* **27**, 82–84.

[27] Harville, D.A. (1975). Experimental randomization: who needs it? *The American Statistician* **29**, 27–31.

[28] Holland, P.W. (1986). Statistics and causal inference (with discussion), *Journal of the American Statistical Association* **81**, 945–970.

[29] Holschuh, N. (1980). Randomization and design: I, in *R. A. Fisher: An Appreciation*, S.E. Fienberg & D.V. Hinkley, eds, Springer Verlag, New York, pp. 35–45.

[30] Kempthorne O. (1952). *The Design and Analysis of Experiments*, New York: John Wiley & Sons.

[31] Kempthorne O. (1955). The randomization theory of experimental inference, *Journal of the American Statistical Association* **50**, 946–967.

[32] Kempthorne, O. (1977). Why randomize? *Journal of Statistical Planning and Inference* **1**, 1–25.

[33] Kempthorne, O. (1986). Randomization – II, in *Encyclopedia of Statistical Sciences*, Vol. 7, S. Kotz & N.L. Johnson, eds, John Wiley & Sons, New York, pp. 519–525.

[34] Kiaer, A.N. (1976). *The Representative Method of Statistical Surveys*, Central Bureau of Statistics of Norway, Oslo. English translation of 1897 edition which was

issued as No. 4 of The Norwegian Academy of Science and Letters, The Historical, Philosophical Section.

[35] Kiefer, J. (1959). Optimal experimental designs, *Journal of the Royal Statistical Society, Series B* **21**, 272–319.

[36] Kish, L. (1989). Randomization for statistical designs, *Statistica* **49**, 299–303.

[37] Lindley, D.V. & Novic, M. (1981). The role of exchangeability in inference, *Annals of Statistics* **9**, 45–58.

[38] Mayo, O. (1987). Comments on randomization and the Design of Experiments by P. Urbach, *Philosophy of Science* **54**, 592–596.

[39] Mier, P. (1986). Damned liars and expert witnesses, *Journal of the American Statistical Association* **81**, 269–276.

[40] Mood, A.M., Graybill, F.A. & Boes, D.C. (1974). *Introduction to the Theory of Statistics*, 3rd Edition, McGraw-Hill, New York.

[41] Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection (with discussion), *Journal of the Royal Statistical Society* **97**, 558–625.

[42] Peirce, C.S. & Jastrow, J. (1885). On small differences of sensation, *Memoirs of the National Academy of Sciences* **3**, 75–83.

[43] Picard, R. (1980). Randomization and design: II, in *R. A. Fisher: An Appreciation*, S.E. Fienberg & D.V. Hinkley, eds, Springer Verlag, New York, pp. 46–58.

[44] Pitman, E.G. (1937a). Significance tests which may be applied to samples from any populations, *Journal of the Royal Statistical Society Supplement* **4**, 119–130.

[45] Pitman, E.G. (1937b). Significance tests which may be applied to samples from any populations: II. The correlation coefficient test, *Journal of the Royal Statistical Society Supplement* **4**, 225–232.

[46] Pitman, E.G. (1938). Significance tests which may be applied to samples from any populations: III. The analysis of variance test, *Biometrika* **29**, 322–335.

[47] Preece, D.A. (1990). R. A. Fisher and experimental design: a review, *Biometrics* **46**, 925–935.

[48] Rao, J.N.K. & Bellhouse, D.R. (1990). History and development of the theoretical foundations of survey based estimation and analysis, *Survey Methodology* **16**, 3–29.

[49] Richet, C. (1884). La suggestion mentale et le calcul des probabilités, *Revue Philosophique de la France et de Étranger* **18**, 609–674.

[50] Robins, J. & Wasserman, L. (2002). Conditioning, likelihood, and coherence: a review of some foundational concepts, in *Statistics in the 21st Century*, A.E. Raftery, M.A. Tanner & M.T. Wells, eds, Chapman & Hall/CRC, Boca Raton, pp. 431–443.

[51] Royall, R.M. (1970). On finite population sampling theory under certain linear regression models, *Biometrika* **57**, 377–387.

[52] Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41–55.

[53] Rosenthal, R. (1966). *Experimenter Effects in Behavioral Research*, Appleton-Century-Crofts, New York.

[54] Rouanet, H., Bernard, J.-M. & LeCoutre, B. (1986). Nonprobabilistic statistical inference: a set theoretic approach, *The American Statistician* **40**, 60–65.

[55] Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology* **66**, 688–701.

[56] Rubin, D.B. (1990). Formal modes of statistical inference for causal effects, *Journal of Statistical Planning and Inference* **25**, 279–292.

[57] Senn, S. (1994). Fisher's game with the devil, *Statistics in Medicine* **13**, 217–230.

[58] Shipley, B. (2000). *Cause and Correlation in Biology*, Cambridge University Press, Cambridge.

[59] Smith, T.M.F. (1976). The foundation of survey sampling: a review, *Journal of the Royal Statistical Society, Series A* **139**, 183–204.

[60] Smith, T.M.F. (1983). On the validity of inferences from non-random samples, *Journal of the Royal Statistical Society, Series A* **146**, 394–403.

[61] Smith, T.M.F. (1997). Social surveys and social science, *The Canadian Journal of Statistics* **25**, 23–44.

[62] Sprott, D.A. & Farewell, V.T. (1993). Randomization in experimental science, *Statistical Papers* **34**, 89–94.

[63] Stone, R. (1993). The assumptions on which causal inferences rest, *Journal of the Royal Statistical Society, Series B* **55**, 455–466.

[64] Tchuprov, A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations, *Metron* **2**, 646–680.

[65] Thornett, M.L. (1982). The role of randomization in model-based inference, *Australian Journal of Statistics* **24**, 137–145.

[66] Urbach, P. (1985). Randomization and the design of experiments, *Philosophy of Science* **52**, 256–273.

[67] Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association* **60**, 63–69.

[68] Winship, C. & Mare, R.D. (1992). Models for sample selection bias, *Annual Review of Sociology* **18**, 327–350.

[69] Winship, C. & Morgan, S.L. (1999). The estimation of causal effects from observational data, *Annual Review of Sociology* **25**, 659–706.

ROBERT J. BOIK

Randomization Based Tests

John Ludbrook

Volume 4, pp. 1674–1681

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Randomization Based Tests

Introduction

My brief is to provide an overview of the randomization model of statistical inference, and of the statistical tests that are appropriate to that model. I may have exceeded that brief. In the first instance, I have felt obliged to present and discuss one of its major rivals: the population model of inference. In the second instance, I have not confined my description to tests on continuous or rank-ordered data under the randomization model. It has seemed to me that the so-called 'exact' tests on categorical data should also be considered under this model for they, too, involve permutation (*see* **Exact Methods for Categorical Data**).

I shall try to differentiate between the population and randomization models of statistical inference. I shall not consider the Bayesian model, partly (at least) because it is not popular, and partly because I have to confess that I do not fully understand its application to real-life experimentation (*see* **Bayesian Statistics**).

The Population Model of Statistical Inference

This is sometimes known as the classical model. It was first articulated in stringent theoretical terms by **Neyman** and **Pearson** with respect to continuous data [21, 22] (*see* **Neyman–Pearson Inference**). It presupposes, or has as its assumptions, the following: (a) the samples are drawn randomly from defined populations, and (b) the frequency distributions of the populations are mathematically definable. An alternative definition embraces the notion that if a large (infinite) number of samples (with replacement) were to be taken from the population(s) actually sampled in the experiment, then the P value attached to the null hypothesis corresponds to the frequency with which the values of the test statistic (for instance, the difference between means) are equal to or exceed those in the actual experiment. It should be noted that Neyman and Pearson also introduced the notions of Type 1 error (α, or false rejection of the null hypothesis) and Type 2 error (β, or false acceptance of the null hypothesis) [21, 22]. By extension, this led to the notion of **power** to reject the null hypothesis $(1-\beta)$. Though Neyman and Pearsons's original proposals were concerned with tests of significance (P values), Neyman later introduced the notion of **confidence intervals** (CIs) [20].

It is important to note that statistical inferences under this model, whether they are made by hypothesis-testing (P values) or by estimation (CIs), refer to the parent populations from which the random samples were drawn.

The mathematically definable populations to which the inferences refer are usually normally distributed or are derivatives of the normal (Gaussian) distribution (for instance, t, F, χ^2).

A late entrant to the population model of inference is the technique of **bootstrapping**, invented by Bradley Efron in the late 1970s [5]. Bootstrapping is done, using a fast computer, by random resampling of the samples (because the populations are inaccessible), with replacement. It allows inferences to be made (P values, SEs, CIs) that refer to randomly sampled populations, but with the difference from classical statistical theory that no assumptions need be made about the frequency distribution of the populations.

The historical relationship between the enunciation of the population model of inference and the first descriptions of statistical tests that are valid under this model is rather curious. This is because the tests were described before the model was. 'Student' (**W.S. Gosset**) described what came to be known as the t distribution in 1908 [28], later converted into a practical test of significance by Fisher [6]. **R.A. Fisher** gave, in 1923, a detailed account of his use of **analysis of variance (ANOVA)** to evaluate the results of a complex experiment involving 12 varieties of potato, 6 different manures, and 3 replicates in a **randomized block design** [9]. The analysis included the Variety × Manure interaction. All this, performed with pencil and paper! But it was not until 1928 that Neyman and Pearson expounded the population model of inference [21, 22].

So, what is wrong with the population model of inference? As experimental biologists (not least, behavioral scientists) should know – but rarely admit – we never take random samples (*see* **Randomization**). At best, we take nonrandom samples of the experimental units (humans, animals, or whatever) that are available – 'samples of convenience' [16]. The availability may come about because units have presented themselves to a clinic, have responded to a call for volunteers, or can be purchased from animal breeders. We then randomize the experimental units to 'treatment groups', for instance, no treatment (control), placebo, various drugs, or various environmental manipulations. In these circumstances, it is impossible to argue that genuine populations have been randomly sampled. Enter, the randomization model of inference and tests under that model.

The Randomization Model of Statistical Inference

This was enunciated explicitly only about 50 years ago [12], even though statistical tests under this model had been described and performed since the early 1930s. This is undoubtedly because, until computers were available, it might take days, weeks, or even months to analyze the results of a single experiment. Much more recently, this and other models have been critically appraised by Rubin [26]. The main features of the model are that (a) experimental groups are not acquired by random sampling, but by taking a nonrandom sample and allocating its members to two or more 'treatments' by a process of randomization; (b) tests under this model depend on a process of permutation (randomization); (c) the tests do not rely on mathematically defined frequency-distributions; (d) inferences under this model do not refer to populations, but only to the particular experiment; (e) any wider application of these statistical inferences depends on scientific (verbal), not statistical, argument. Good accounts of this model for nonmathematical statisticians are given in monographs [4, 11, 16, 18] (*see* **Permutation Based Inference**).

Arguments in favor of this model have been provided by R.A. Fisher [7]: '....conclusions [from t or F Tests] have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method [randomization].' And by Kempthorne [12], who concluded: 'When one considers the whole problem of experimental inference, that is, of tests of significance, estimation of treatment differences and estimation of the errors of estimated differences, there seems little point in the present state of knowledge in using [a] method of inference other than randomization analysis.'

The randomization model of inference and the statistical tests conducted under that model have attracted little attention from theoretical statisticians. Why? My guess is that this is because, to theoretical statisticians, the randomization model is boring. There are some exceptions [2, 25] but, as best I can judge, these authors write of inferences being referable to populations. I argue that because the experimental designs involve randomization, not random sampling, they are wrong.

What statistical tests are appropriate to randomization? For continuous data, the easiest to understand are those designed to test for differences between or among means [15]. In 1935, R.A. Fisher analyzed, by permutation, data from an experiment on matched pairs of plants performed by Charles Darwin [8]. Fisher's goal was to show that breaches of the assumptions for Student's t Test did not affect the outcome. In 1936, Fisher analyzed craniometric data from two independent groups by permutation [7]. In 1933, Eden and Yates reported a much more ambitious analysis of a complex experiment by permutation, to show that analysis of variance (ANOVA) was not greatly affected by breaches of assumptions [3]. As I shall show shortly, the calculations involved in these studies can only be described as heroic. Perhaps, because of this, Fisher repudiated permutation tests in favor of t Tests and ANOVA. He said that their utility '... consists in their being able to supply confirmation. whether rightly or, more often, wrongly, when it is suspected that the simpler tests have been apparently injured by departure from normality' [8]. It is strange that Fisher, the inventor, practitioner, and proselytizer of randomization in experimental design [8], seems not to have made the connection between randomization in design and the use of permutation (randomization) tests to analyze the results. All Fisher's papers can be downloaded free-of-charge from the website `www.library.adelaide.edu.au/digitised/fisher/`.

Over the period 1937 to 1952, the notion of transforming continuous data into ranks was developed [10, 13, 19, 29] (*see* **Rank Based Inference**). The goal of the authors was to simplify and expedite analysis of variance, and to cater for data that do not fulfill the assumption of normality in the parent populations. All four sets of authors relied on calculating approximate (asymptotic) P values by way of the χ^2 or z distributions. It is true that **Wilcoxon** [29] and Kruskal and Wallis [13] did produce small tables of exact P values resulting from the permutation of ranks. But it is curious that only Kruskal and Wallis [13] refer to Fisher's idea of exact tests based on the permutation of means [7, 8].

There is a misapprehension, commonplace among investigators but also among some statisticians [26], that the rank-order tests are concerned with differences between medians. This is simply untrue. These tests are concerned with differences in group mean ranks ($\bar{R}_1 - \bar{R}_2$), and it can easily be demonstrated that, because of the method of ranking, this is not the same as differences between medians [1]. It should be said that though Wilcoxon was the first to describe exact rank-order tests, it was Milton Friedman (later a Nobel prize-winner in economics) who first introduced the notion of converting continuous into rank-ordered data in 1937 [10], but he did not embrace the notion of permutation.

Now, the analysis of categorical data by permutation – this, too, was introduced by R.A. Fisher [8]. He conjured up a hypothetical experiment. It was that a colleague reckoned she could distinguish a cup of tea to which the milk had been added first from a cup in which the milk had been added after the tea (Table 3). This well-known 'thought' experiment was the basis for what is now known as the Fisher exact test. Subsequently, this exact method for the analysis of categorical data in **2 x 2 tables** of frequency has been extended to $r \times c$ tables of frequencies, both unordered and ordered.

As a piece of history, when **Frank Yates** described his correction for continuity for the χ^2 test on small 2×2 tables, he validated his correction by reference to Fisher's exact test [30].

Statistical Tests Under the Randomization Model of Inference

Differences Between or Among Means

Two Independent Groups of Size n_1 and n_2. The procedure followed is to exchange the members of the groups in all possible permutations, maintaining the original group sizes (Tables 1 and 2). That is, all possible ways in which the original randomization could have fallen out are listed. Then:

$$P = \frac{\begin{array}{c}\text{No. of permutations in which the difference}\\ \text{between group means is equal to}\\ \text{or more extreme than that observed}\end{array}}{\text{All possible permutations}}. \tag{1}$$

Table 1 Illustration of the process of permutation in the case of two independent, randomized, groups

Permutation	Group 1 $n = 2$	Group 2 $n = 3$
A	a, b	c, d, e
B	a, c	b, d, e
C	a, d	b, c, e
D	a, e	b, c, d
E	b, c	a, d, e
F	b, d	a, c, e
G	b, e	a, c, d
H	c, d	a, b, d
I	c, e	a, b, d
J	d, e	a, b, c

The number of possible permutations (see Formula 2) is $(2 + 3)!/(2)!(3)! = 10$, whether the entries are continuous or ranked data.

Table 2 Numerical datasets corresponding to Table 1

Dataset	Group 1	Group 2	$\bar{x}_2 - \bar{x}_1$	$\bar{R}_2 - \bar{R}_1$
A	1, 3	4, 7, 9	4.67	2.50
B	1, 4	3, 7, 9	3.83	1.67
C	1, 7	3, 4, 9	1.33	0.83
D	1, 9	3, 4, 7	−0.33	0.00
E	3, 4	1, 7, 9	2.17	0.83
F	3, 7	1, 4, 9	−0.33	0.00
G	3, 9	1, 4, 7	−2.00	−0.83
H	4, 7	1, 3, 9	−1.17	−0.83
I	4, 9	1, 3, 7	−2.83	−1.67
J	7, 9	1, 3, 4	−5.33	−2.50

1, 3, 4, 7 and 9 were substituted for a, b, c, d and e in Table 1, and the differences between means ($\bar{x}_2 - \bar{x}_1$) calculated. The ranks 1, 2, 3, 4 and 5 were substituted for a, b, c, d and e in Table 1, and the differences between mean ranks ($\bar{R}_2 - \bar{R}_1$) calculated.

Exact permutation test on difference between means [7]: dataset A, one-sided $P = 0.100$, two-sided $P = 0.200$; dataset B, one-sided $P = 0.200$, two-sided $P = 0.300$; dataset J, one-sided $P = 0.100$, two-sided $P = 0.100$.

Exact permutation test on difference between mean ranks [18, 28]: dataset A, one-sided $P = 0.100$, two-sided $P = 0.200$; dataset B, one-sided $P = 0.200$, two-sided $P = 0.400$; dataset J, one-sided $P = 0.100$, two-sided $P = 0.200$.

For a one-sided P, only differences that are equal to or greater than that observed (that is, with + sign) are used. For a two-sided P, differences, regardless of the sign, that are greater than that observed are used.

The number of all possible permutations increases steeply with group size. The formula for this, for group sizes of n_1 and n_2, and corresponding group means $\bar{x}_1$ and $\bar{x}_2$, is:

$$\frac{(n_1 + n_2)!}{(n_1)!(n_2)!} \tag{2}$$

This innocuous formula disguises the magnitude of the computational problem. Thus, for $n_1 = n_2 = 5$, the number of all possible permutations is 252. For $n_1 = n_2 = 10$, it is 184 756. And for $n_1 = n_2 = 15$, it is 155 117 520. A solution to the sometimes massive computational problem is to take Monte Carlo random samples (*see* **Monte Carlo Simulation**) of, say, 10 000 from the many millions of possible permutations. Interestingly, this is what Eden and Yates did in solving the massive problem of their complex analysis of variance – by the ingenious use of cards [3].

Matched Pairs. Given n matched pairs, the number of all possible permutations is

$$2^n. \tag{3}$$

Thus, if $n = 5$, the number of all possible permutations is 32; for $n = 10$, 1024; and for $n = 15$, 32768. The last is what R.A. Fisher computed with pencil and paper in 1935 [8].

k Independent Groups. This corresponds to one-way ANOVA. It is a simple extension of two independent groups, and the number of all possible permutations is given by an extension of formula (2). Usually, instead of using the difference between the means, one uses a simplified F statistic [4, 11, 18].

k Matched Groups. This corresponds to two- or multi-way ANOVA (*see* **Factorial Designs**). There is no great problem if only the main effects are to be extracted. But it is usually the interactions that are the focus of interest (*see* **Interaction Effects**). There is no consensus on how best to go about extracting these. Edgington [4], Manly [18], and Good [11] have suggested how to go about this.

In the case of a two-way, factorial, design first advocated by Fisher [9], there seems to be no great problem. Good [11] describes clearly how, first, the main (fixed) effects should be factored out, leaving the two-way interaction for analysis by permutation.

But what about a three-way factorial design? Not uncommon in biomedical experimentation. But this involves no fewer than three two-way interactions, and one three-way interaction. It is the last that might test the null hypothesis of prime interest. Can this interaction be extracted by permutation? Good [11] shows, by rather complex theoretical argument, how this could be done.

Then, what about **repeated-measures designs**? Not an uncommon design in biomedical experimentation. If the order of repeated measurements is randomized, Lunneborg shows how this can be handled by permutation [16, 17]. But, if the order of the repeated measurements is not randomized (for example, time cannot be randomized, nor can ascending dose- or stimulus-response designs), surely, analysis of the results cannot be done under the randomization model of inference?

My pragmatic view is that the more complex the experimental design, the less practicable is a randomization approach. Or, to put it another way, the more complex the experimental design, the closer tests under the classical and randomization models of inference approach each other.

Confidence Intervals (CIs) Under the Randomization Model. It seems to me that CIs are irrelevant to the randomization model. This is because they refer to populations that have been randomly sampled [20], and this is emphatically not the case under the randomization model.

Minimal Group Size and Power in Randomization Tests. Conventional ways of thinking about these have to be abandoned, because they depend on the population model of inference. My practical solution is to calculate the maximum number of possible permutations (formulae 2, 3). It must be at least 20 in order to be able to achieve $P \leq 0.05$.

Differences Between or Among Group Mean Ranks

As indicated above, the computational problem of evaluating these differences by permutation is much less than that presented by differences between/among means. It is, therefore, rarely necessary to resort to Monte Carlo random sampling.

An enormous number of rank-order tests has been described: the **Wilcoxon-Mann-Whitney test** for two independent groups [19, 29], the Wilcoxon matched pairs, signed-rank, test [29], the **Kruskal-Wallis test** (*see* **Kruskal–Wallis Test**) on k independent groups [13], and the **Friedman test** on k related groups [4], are well known. But there is a host of other, eponymous, rank-order tests [14]. A word of caution is necessary. Almost all general statistics programs offer these tests, though executed asymptotically by normal or χ^2 approximations rather than by permutation. The problem with the asymptotic versions is that the algorithms used are often not described. How ties are handled is of critical importance. This matter has been addressed recently [1]. An example of the exact Wilcoxon-Mann-Whitney test is given in Table 2.

Exact Tests on Categorical Data

The simplest case is a 2×2 table of frequencies (Table 3). As Fisher described in simple terms [8], and Siegel and Castellan in modern notation [27], the point probability of H_0 is described by:

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!} \quad (4)$$

However, the probability of H_0 refers to the probability of occurrence of the observed values, plus the probabilities of *more extreme values in the same direction*. And, if a two-sided P is sought, *the same or more extreme values in either direction* must be taken into account. Thus, (4) must be applied to all these tables, and the P values summed. In the example of Table 3, the two-sided P is twice the one-sided value. This is only because of the equality and symmetry of the marginal totals.

There is a multitude of **exact tests** on categorical data for 2×2 tables of frequencies, and for unordered and ordered $r \times c$ tables. But even in the simplest case of 2×2 tables, there are problems about the null hypotheses being tested. It seems to me that the Fisher exact test, and the various formulations of the χ^2 test, are concerned with very vague null hypotheses (H_0). I prefer tests in which H_0 is much more specific: for instance, that OR = 1, $p_1 - p_2 = 0$, or $p_1/p_2 = 1$ (Table 4).

There is a further problem with the Fisher exact test. It is a conditional test, in which the column and marginal totals in a 2×2 table are fixed in advance. This was so in Fisher's thought experiment (Table 3), but it is almost inconceivable that this could be achieved in a real-life experiment. In the case that H_0 is that OR = 1, there is no difficulty if

Table 3 Results of R.A. Fisher's thought experiment on the ability of a lady to distinguish whether milk or tea was added first to the cup [8]

		Actual design		Row
		Milk first	Tea first	Total
Lady's decisions	Milk first	3	1	4
	Tea first	1	3	4
	Column total	4	4	8

The lady was informed in advance that she would be presented with 8 cups of tea, in 4 of which the tea had been added first, in 4 the milk. The exact probability for the null hypothesis that the lady was unable to distinguish the order is obtained by the sum of the point probabilities (Formula 4) for rearrangements of the Table in which the observed, or more extreme, values would occur. Thus, two-sided $P = 0.22857 + 0.0142857 + 0.22857 + 0.0142857 = 0.48573$.

Table 4 Properties of exact tests on 2×2 tables of frequencies

Test	Null hypothesis	Conditioning
Fisher's	Vague	Both marginal totals fixed
Odds ratio	OR = 1	Column marginal totals fixed
Exact χ^2	Columns and rows independent	Unconditional
Difference between proportions	$p_1 - p_2 = 0$	Unconditional
Ratio between proportions	$p_1/p_2 = 1$	Unconditional

Conditioning means 'fixed in advance by the design of the experiment'. In real-life experiments, both marginal totals can never be fixed in advance. Column totals are usually fixed in advance if they represent treatment groups.

one regards the column totals as group sizes. Tests on proportions, for instance, that $p_1 - p_2 = 0$, or $p_1/p_2 = 1$, are unconditional. But there are complex theoretical and computational difficulties with these.

Acknowledgements

I am most grateful to colleagues who have read and commented on my manuscript, especially my Section Editor, Cliff Lunneborg.

Appendix: Software Packages for Permutation Tests

I list only those programs that are commercially available. I give first place to those that operate under the Microsoft Windows system. There is an excellent recent and comparative review of several of these [23, 24]. The remaining programs operate under DOS. I shall mention only one of the latter, which I have used. The remaining DOS programs require the user to undertake programming, and I do not list these. All the commercial programs have their own datafile systems. I have found dfPowerDBMS/Copy v. 8 (Dataflux Corporation, Cary NC) invaluable in converting any spreadsheet into the appropriate format for almost any statistics program.

Microsoft Windows

StatXact v.6 with Cytel Studio (Cytel Software Corporation, Cambridge MA). StatXact is menu-driven. For differences between/among group means, it caters for two independent groups, matched pairs, and the equivalent of one-way ANOVA. It does not have a routine for two-way ANOVA. For differences between/among group mean ranks, it caters for the Wilcoxon-Mann-Whitney test, the Wilcoxon signed-rank test for matched pairs, the Kruskal-Wallis test for k independent group mean ranks, and the Friedman test for k matched groups. For categorical data, it provides a great number of tests. These include the Fisher exact test, exact χ^2, a test on OR = 1, tests on differences and ratios between proportions; and for larger and more complex tables of frequencies, the Cochran–Armitage test on ordered categorical data.

LogXact (Cytel Software Corporation, Cambridge MA). This deals exclusively with exact logistic regression analysis for small samples. However, it does not cater for stepwise logistic regression analysis.

SAS v. 8.2 (SAS Institute Inc, Cary NC). SAS has introduced modules for exact tests, developed by the Cytel Software Corporation. These include PROC FREQ, PROC MULTTEST, and PROC NPAR1WAY, PROC UNIVARIATE, and PROC RANK. NPAR1WAY provides permutation tests on the means of two or more independent groups, but not on more than two related groups. PROCRANK caters for a variety of tests on mean ranks, and PROCFREQ a large number of exact tests on tables of frequencies.

Testimate v. 6 (Institute for Data Analysis and Study Planning, Gauting/Munich). This caters for a variety of exact rank-order tests and tests on tables of frequency, but no tests on means.

SPSS (SPSS Inc, Chicago IL). This very popular statistics package has an Exact Tests add-on (leased from StatXact), with routines for a wide range of exact tests on categorical data, some on rank-ordered data, but none on differences between/among means.

DOS Programs

RT v. 2.1 (West Inc., Cheyenne WY). This is Bryan Manly's program, based on his book [18]. One important attribute is that it provides for two-way ANOVA carried out by permutation, in which the interaction can be extracted. However, it has not been developed since 1991, though Bryan Manly hopes that someone will take on the task of translating it onto a Windows platform (personal communication).

References

[1] Bergmann, R., Ludbrook, J. & Spooren, W.P.M.J. (2000). Different outcomes of the Wilcoxon-Mann-Whitney test from different statistics packages, *The American Statistician* **54**, 72–77.

[2] Boik, R.J. (1987). The Fisher-Pitman permutation test: a non-robust alternative to the normal theory F test when variances are heterogeneous, *British Journal of Mathematical and Statistical Psychology* **40**, 26–42.

[3] Eden, T. & Yates, F. (1933). On the validity of Fisher's z test when applied to an actual example of nonnormal data, *Journal of Agricultural Science* **23**, 6–16.

[4] Edgington, E.S. (1995). *Randomization Tests*, 3rd Edition, Marcel Dekker, New York.

[5] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, London.

[6] Fisher, R.A. (1925). Applications of "Student's" distribution, *Metron* **5**, 90–104.

[7] Fisher, R.A. (1936). The coefficient of racial likeness and the future of craniometry, *Journal of the Royal Anthropological Institute* **66**, 57–63.

[8] Fisher, R.A. (1971). *The Design of Experiments*, in *Statistical Methods, Experimental Design and Scientific Inference*, 8th Edition, (1st Edition 1935), (1990) J.H. Bennett, ed., Oxford University Press, Oxford.

[9] Fisher, R.A. & Mackenzie, W.A. (1923). Studies in crop variation. II: the manurial response of different potato varieties, *Journal of Agricultural Science* **13**, 311–320.

[10] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the American Statistical Association* **32**, 675–701.

[11] Good, P.I. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, 2nd Edition, Springer-Verlag, New York.

[12] Kempthorne, O. (1955). The randomization theory of experimental inference, *Journal of the American Statistical Association* **50**, 946–967.

[13] Kruskal, W.H. & Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association* **47**, 583–621.

[14] Lehmann, E.L. (1998). *Nonparametrics: Statistical Methods Based on Ranks*, 1st Edition, revised, Prentice-Hall, Upper Saddle River.

[15] Ludbrook, J. & Dudley, H.A.F. (1998). Why permutation tests are superior to t- and F-tests in biomedical research, *The American Statistician* **52**, 127–132.

[16] Lunneborg, C.E. (2000). *Data Analysis by Resampling: Concepts and Applications*, Duxbury Press, Pacific Lodge.

[17] Lunneborg, C.E. (2002). Randomized treatment sequence designs: the randomization test as a nonparametric replacement of ANOVA and MANOVA, *Multiple Linear Regression Viewpoints* **28**, 1–9.

[18] Manly, B.F.J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd Edition, Chapman & Hall, London.

[19] Mann, H.B. & Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics* **18**, 50–60.

[20] Neyman, J. (1941). Fiducial argument and the theory of confidence intervals, *Biometrika* **32**, 128–150.

[21] Neyman, J. & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part I, *Biometrika* **20A**, 175–240.

[22] Neyman, J. & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part II, *Biometrika* **20A**, 263–294.

[23] Oster, R.A. (2002a). An examination of statistical software packages for categorical data analysis using exact methods, *The American Statistician* **56**, 235–246.

[24] Oster, R.A. (2002b). An examination of statistical software packages for categorical data analysis using exact methods – Part II, *The American Statistician* **57**, 201–213.

[25] Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any population, *Journal of the Royal Statistical Society B* **4**, 119–130.

[26] Rubin, D.B. (1991). Practical applications of modes of statistical inference for causal effects and the critical role of the assignment mechanism, *Biometrics* **47**, 1213–1234.

[27] Siegel, S. & Castellan, N.J. (1988). *Nonparametric Statistics for the Behavioral Sciences*, 2nd Edition, McGraw-Hill, New York.

[28] "Student". (1908). The probable error of a mean, *Biometrika* **6**, 1–25.

[29] Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics* **1**, 80–83.

[30] Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test, *Supplement Journal of the Royal Statistical Society* **1**, 217–235.

JOHN LUDBROOK

Randomized Block Design: Nonparametric Analyses

Li Liu and Vance W. Berger

Volume 4, pp. 1681–1686

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Randomized Block Design: Nonparametric Analyses

In a **randomized block design**, there are, in addition to the experimental factor or factors of interest, one or more nuisance factors (*see* **Nuisance Variables**) influencing the measured responses. We are not interested in evaluating the contributions of these nuisance factors. For example, gender may be relevant in studying smoking cessation, but, in a comparative evaluation of a particular set of smoking cessation techniques, we would not be concerned with assessing the influence of gender per se. However, we would be interested in controlling the gender influence. One can control such nuisance factors by a useful technique called *blocking* (*see* **Block Random Assignment**), which can reduce or eliminate the contribution these nuisance factors make to the experimental error. The basic idea is to create homogeneous blocks or **strata** in which the levels of the nuisance factors are held constant, while the levels of the experimental factors are allowed to vary. Each estimate of the experimental factor effect within a block is more efficient than estimates across all the samples. When one pools these more efficient estimates across blocks, one should obtain a more efficient estimate than would have been available without blocking [3, 9, 13].

One way to analyze the randomized block design is to use standard parametric **analysis of variance (ANOVA)** methods. However, these methods require the assumption that the experimental errors are normally distributed. If the errors are not normally distributed, or if there are outliers in the data, then parametric analyses may not be valid [2]. In this article, we will present a few distribution-free tests, which do not require the normality assumption. These nonparametric analyses include the Wilcoxon signed rank test (*see* **Signed Ranks Test**), **Friedman's test**, the aligned rank test, and Durbin's test.

The Sign Test for the Randomized Complete Block Design with Two Treatments

Consider the boys' shoe-wear data in Table 1, which is from [3]. Two sole materials, A and B, were randomly assigned to the left sole or right sole for each boy. Each boy is a block, and has one A-soled shoe and one B-soled shoe. The goal was to determine whether or not there was any difference in wearing quality between the two materials, A and B.

Table 1 Boys' shoe-wear. Example: the amount of wear measured for two different materials A and B

Boy	Material A	Material B
1	13.2 (L)[a]	14.0 (R)[b]
2	8.2 (L)	8.8 (R)
3	10.9 (R)	11.2 (L)
4	14.3 (L)	14.2 (R)
5	10.7 (R)	11.8 (L)
6	6.6 (L)	6.4 (R)
7	9.5 (L)	9.8 (R)
8	10.8 (L)	11.3 (R)
9	8.8 (R)	9.3 (L)
10	13.3 (L)	13.6 (R)

[a]left sole; [b]right sole.

The **sign test** (*see* **Binomial Distribution: Estimating and Testing Parameters**) is a nonparametric test to compare the two treatments in such designs. It uses the signs of the paired differences, (B – A), to construct the test statistic [12]. To perform the sign test, we count the number of positive paired differences, P_+. Under the null hypothesis of no treatment difference, P_+ has a binomial distribution with parameters $n = 10$ and $p = 0.5$, where n is the number of blocks with nonzero differences. If the sample size is large, then one might use the normal approximation. That is,

$$Z = \frac{P_+ - n/2}{\sqrt{n/4}} \tag{1}$$

is approximately distributed as the standard normal distribution. For the small sample data in Table 1, we have $P_+ = 8$, and the exact binomial P value is 0.109.

The Wilcoxon Signed Rank Test for the Randomized Complete Block Design with Two Treatments

The sign test uses only the signs of the paired differences, but ignores their magnitudes. A more powerful test, called the *Wilcoxon signed rank test*, uses both the signs and the magnitudes of the differences. This signed rank test can also be used to analyze the paired comparison designs. The following steps show how to construct the Wilcoxon signed rank test [12]. First, compute the paired differences and drop all the pairs whose paired differences are zero. Second, rank the absolute values of the paired differences across the remaining blocks (pairs) (*see* **Rank Based Inference**). Third, assign to the resulting ranks the sign of the differences whose absolute value yielded that rank. If there is a tie among the ranks, then use the mid ranks. Fourth, compute the sum of the ranks with positive signs T_+ and the sum of the ranks with negative signs T_-. $T = \min(T_+, T_-)$ is the test statistic. Reject the null hypothesis of no difference if T is small.

If the sample size is small, then the exact distribution of the test statistic should be used. Tables of critical values for T appear in many textbooks, and can be computed in most statistical software packages. If the sample size n (the number of pairs with nonzero differences) is large, then the quantity

$$Z = \frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \tag{2}$$

is approximately distributed as the standard normal. For the boys' shoe-wear data in Table 1, the signed ranks are listed in Table 2. We find that $T = 3$, $Z = -2.5$, and the approximate P value, based on the normal approximation, is 0.0125. Since the sample size is not especially large, we also compute the P value based on the exact distribution, which is 0.0078.

The small P value suggests that the two sole materials were different. Compared to the sign test P value of 0.109, the Wilcoxon signed rank test P value is much smaller. This is not surprising, since the Wilcoxon signed rank test considers both the signs and the ranks of the differences, and hence it is more powerful (*see* **Power**).

Table 2 Signed ranks for Boys' shoe-wear data

Boy	Material A	Material B	Difference (B − A)	Signed
1	13.2 (L)[a]	14.0 (R)[b]	0.8	9
2	8.2 (L)	8.8 (R)	0.6	8
3	10.9 (R)	11.2 (L)	0.3	4
4	14.3 (L)	14.2 (R)	−0.1	−1
5	10.7 (R)	11.8 (L)	1.1	10
6	6.6 (L)	6.4 (R)	−0.2	−2
7	9.5 (L)	9.8 (R)	0.3	4
8	10.8 (L)	11.3 (R)	0.5	6.5
9	8.8 (R)	9.3 (L)	0.5	6.5
10	13.3 (L)	13.6 (R)	0.3	4

[a]left sole; [b]right sole.

Friedman's Test for the Randomized Complete Block Design

Friedman's test [5] is a nonparametric method for analyzing randomized complete block designs, and it is based on the ranks of the observations within each block. Consider the example in Table 3, which is an experiment designed to study the effect of four drugs [8]. In this experiment, there are five litters of mice, and four mice from each litter were chosen and randomly assigned to four drugs (A, B, C, and D). The lymphocyte counts (in thousands per cubic millimeter) were measured for each mouse. Here, each litter was a block, and all four drugs were assigned once and only once per block. Since all four treatments (drugs) are compared in each block (litter), this is a randomized complete block design.

To compute the Friedman test statistic, we first rank the observations from different treatments within each block. Assign mid ranks in case of ties. Let b denotes the number of blocks, t denotes the number of treatments, y_{ij} $(i = 1, \ldots, b, j = 1, \ldots, t)$ denotes the observation in the ith block and jth treatment, and R_{ij} denotes the rank of y_{ij} within each block. The Friedman test statistic is based on the sum of squared differences of the average rank for each treatment from the overall average rank. That is,

$$F_R = \sum_{j=1}^{t} \frac{(\bar{R}_j - \bar{R})^2}{\sigma^2_{\bar{R}_j}} \tag{3}$$

where $\bar{R}_j$ is the average rank for treatment j, $\bar{R}$ is the overall average rank, and the denominator is the

Table 3 Lymphocyte counts

	Drugs			
Litter	A	B	C	D
1	7.1	6.7	7.1	6.7
2	6.1	5.1	5.8	5.4
3	6.9	5.9	6.2	5.7
4	5.6	5.1	5.0	5.2
5	6.4	5.8	6.2	5.3

variance of the first term in the numerator. Note that this variance does not depend on the treatment group, but retains the subscript because it is the variance of a quantity indexed by this subscript. Under the null hypothesis, the Friedman test statistic has a chi-square distribution with $(t-1)$ degrees of freedom. We reject the null hypothesis that there is no difference between treatments if F_R is large.

If there are no ties within blocks, then we have

$$\sigma^2_{\bar{R}_j} = \frac{t(t+1)}{12b}, \tag{4}$$

and Friedman's test statistic reduces to

$$F_R = \frac{12}{bt(t+1)} \sum_{j=1}^{t} R^2_{.j} - 3b(t+1). \tag{5}$$

If there are ties within blocks, then we have

$$\sigma^2_{\bar{R}_j} = \frac{t(t+1)}{12b} \times D, \tag{6}$$

where

$$D = 1 - \frac{\sum_{i=1}^{b} T_i}{bt(t^2-1)}, \; T_i = \sum_k (S_k^3 - S_k), \tag{7}$$

k ranges over the number of ties in the ith block, and S_k is the number of observations at the kth tied value.

To compute Friedman's test statistic for the data in Table 3, we first compute the ranks within each block. These are listed in Table 4.

We have

$$\begin{aligned}\sum_{j=1}^{t} (\bar{R}_j - \bar{R})^2 &= (19.5/5 - 2.5)^2 \\ &\quad + (8.5/5 - 2.5)^2 + (13.5/5 - 2.5)^2 \\ &\quad + (8.5/5 - 2.5)^2 = 3.28,\end{aligned}$$

$$\begin{aligned} D &= 1 - \frac{\sum_{i=1}^{b} T_i}{bt(t^2-1)} = 1 - \frac{12}{5 \times 4 \times (4^2 - 1)} \\ &= 0.96, \\ \sigma^2_{\bar{R}_j} &= \frac{t(t+1)}{12b} \times D = \frac{4 \times (4+1)}{12 \times 5} \times 0.96 = 0.32. \end{aligned} \tag{8}$$

Table 4 The ranks of the Lymphocyte count data

	Drugs			
Litter	A	B	C	D
1	3.5	1.5	3.5	1.5
2	4	1	3	2
3	4	2	3	1
4	4	2	1	3
5	4	2	3	1
Rank sum	19.5	8.5	13.5	8.5

So $F_R = 3.28/0.32 = 10.25$. Asymptotically, F_R has as its null distribution that of the chi-squared random variable with $(t-1)$ degrees of freedom. The P value for our obtained test statistic of 10.25, based on the chi-squared distribution with 3 df, is 0.0166, and indicates that the drug effects are different. The exact distribution (*see* **Exact Methods for Categorical Data**) of the Friedman test statistic can be used for small randomized complete block designs. Odeh et al. [10] provided the critical values of the Friedman test for up to six blocks and six treatments.

Aligned Rank Test for Randomized Complete Block Design

Friedman's test is based on the rankings within blocks, and it has relatively low power when the number of blocks or treatments is small. An alternative is to align the rankings. That is, we subtract from the observation in each block some estimate of the location of the block to make the blocks more comparable. The location-subtracted observations are called *aligned observations*. We rank all the aligned observations from all the blocks instead of ranking them only within each block, and we use the aligned ranks to compute the test statistic. This is called the *aligned rank test* [6]. The aligned rank test statistic is the same as Friedman's test statistic except that the aligned rank test computes the ranks differently.

For the example in Table 3, the aligned rank test statistic is 10.53. We again evaluate this against the chi-squared distribution with $(t-1) = 3$ df and obtain a P value of 0.0146, slightly smaller than that associated with Friedman's test. For this data, the difference between the aligned rank test and Friedman's test is not very pronounced, but, in some cases, the more powerful aligned rank test can have a much smaller P value than the Friedman's test.

Durbin's Test for Balanced Incomplete Blocks Design

In a **balanced incomplete block design**, the block size k is smaller than the number of treatments t because it is impractical or impossible to form homogeneous blocks of subjects as large as t. As a result, not all of the treatments can be compared within a block, and this explains the term incomplete. The design is balanced, which means that each pair of treatments is compared in the same number of blocks as every other pair of treatments. And each block contains the same number of subjects and each treatment occurs the same number of times. In such a design, the appropriate subset of k treatments are randomized to the subjects within a particular block.

Table 5 is an example of a balanced incomplete block design [1]. In this study, the measurements are (percentage elongation -300) of specimens of rubber stressed at 400 psi. The blocks are 10 bales of rubber, and two specimens were taken from each bale. Each specimen was assigned to one of the five tests (treatments). We are interested in finding out whether there is any difference among the five tests.

Notice that each pair of treatments occurs together in exactly one bale. Durbin's test [4] can be used to test the treatment difference in a balanced incomplete block design. We first rank the observations within each block, and assign mid ranks for ties. Durbin's test statistic is

$$D_R = \frac{12(t-1)}{rt(k-1)(k+1)} \sum_{j=1}^{t} R_{.j}^2 - \frac{3r(t-1)(k+1)}{k-1}, \tag{9}$$

where $t = 5$ is the number of treatments, $k = 2$ is the number of subjects per block ($k < t$), $r = 4$ is the number of times each treatment occurs, and $R_{.j}$ is the sum of the ranks assigned to the jth treatment. Under the null hypothesis that there are no treatment differences, D_R is distributed approximately as chi-squared, with $(t-1)$ degrees of freedom. For the example in Table 5, we have

$$\begin{aligned} D_R = {} & \frac{12 \times (5-1)}{4 \times 5 \times (2-1) \times (2+1)} \\ & \times (7^2 + 6^2 + 4^2 + 8^2 + 5^2) \\ & - \frac{3 \times 4 \times (5-1) \times (2+1)}{2-1} = 8 \end{aligned} \tag{10}$$

with four degrees of freedom. The P value of 0.0916 indicates that the treatments are somewhat different. Durbin's test reduces to Friedman's test if the number of treatments is the same as the number of units per block.

Table 5 Measurements (percentage elongation −300) of rubber stressed at 400 psi

	Treatment				
Bale	1	2	3	4	5
1	35	16			
2	20		10		
3	13			26	
4	25				21
5		16	5		
6		21		24	
7		27			16
8			20	37	
9			15		20
10				31	17

Cochran–Mantel–Haenszel Row Mean Score Statistic

The Sign test, Friedman's test, the aligned rank test, and Durbin's test are all special cases of the Cochran–Mantel–Haenszel (CMH) (*see* **Mantel–Haenszel Methods**) row mean score statistic with the ranks as the scores (*see* [11]). Suppose that we have a set of $qs \times r$ contingency tables. Let n_{hij} be the number of subjects in the hth stratum in the ith group and jth response categories, $p_{hi+} = n_{hi+}/n_h$ and $p_{h+j} = n_{h+j}/n_h$. Let

$$\begin{aligned} \mathbf{n}_h' &= (n_{h11}, n_{h12}, \ldots, n_{h1r}, \ldots, n_{hs1}, \ldots, n_{hsr}), \\ \mathbf{P}_{h*+}' &= (p_{h1+}, \ldots, p_{hs+}), \\ \mathbf{P}_{h+*}' &= (p_{h+1}, \ldots, p_{h+r}). \end{aligned} \tag{11}$$

Then

$$\begin{aligned} E(n_{hij}|H_0) &= m_{hij} = n_h p_{hi+} p_{h+j}, \\ \mathrm{E}(\mathbf{n}_h|H_0) &= \mathbf{m}_h = n_h[\mathbf{P}_{h+*} \otimes \mathbf{P}_{h*+}], \end{aligned}$$

$$Var(\mathbf{n}_h|H_0) = \mathbf{V}_h = \frac{n_h^2}{(n_h - 1)}((\mathbf{D}_{Ph+*} - \mathbf{P}_{h+*}\mathbf{P}_{h+*}^{'}) \otimes (\mathbf{D}_{Ph*+} - \mathbf{P}_{h*+}\mathbf{P}_{h*+}^{'})) \quad (12)$$

where $\otimes$ denotes the left-hand Kronecker product, $\mathbf{D}_{Ph+*}$ and $\mathbf{D}_{Ph*+}$ are diagonal matrices with elements of the vectors $\mathbf{P}_{h+*}$ and $\mathbf{P}_{h*+}$ as the main diagonals. The general CMH statistic [7] is

$$\mathrm{Q}_{CMH} = \mathbf{G}^{'}\mathbf{V}_G^{-1}\mathbf{G}, \quad (13)$$

where

$$\mathbf{G} = \sum_h \mathbf{A}_h(\mathbf{n}_h - \mathbf{m}_h),$$

$$\mathbf{V}_G = \sum_h \mathbf{A}_h\mathbf{V}_h\mathbf{A}_h^{'}. \quad (14)$$

where $\mathbf{A}_h$ is a matrix, and different choices of $\mathbf{A}_h$ provide different statistics, such as the correlation statistic, the general association statistic, and so on. To perform the test based on the mean score statistic, we have $\mathbf{A}_h = \mathbf{a}_h^{'} \otimes [\mathbf{I}_{(s-1)}, \mathbf{0}_{(s-1)}]$, where $\mathbf{a}_h = (a_{h1}, \ldots, a_{hr})$ are the scores for the jth response level in the hth stratum [11]. If we use the ranks as the scores, and there is one subject per row and one subject per column in the contingency table of each stratum, then the CMH mean score statistic is identical to Friedman's test. The sign test, aligned rank test, and Durbin's test can also be computed using the CMH mean score statistic with the ranks as the scores.

References

[1] Bennett, C.A. & Franklin, N.L. (1954). *Statistical Analysis in Chemistry and the Chemical Industry*, Wiley, New York.

[2] Berger, V.W. (2000). Pros and cons of permutation tests, *Statistics in Medicine* **19**, 1319–1328.

[3] Box, G., Hunter, W.G. & Hunter, J.S. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, John Wiley & Sons, New York.

[4] Durbin, J. (1951). Incomplete blocks in ranking experiments, *British Journal of Mathematical and Statistical Psychology* **4**, 85–90.

[5] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of American Statistical Association* **32**, 675–701.

[6] Hodges, J.L. & Lehmann, E.L. (1962). Rank methods for combination of independent experiments in analysis of variance, *Annals of Mathematical Statistics* **33**, 482–497.

[7] Landis, J.R., Heyman, E.R. & Koch, C.G. (1978). Average partial association in three-way contingency tables: a review and discussion of alternative tests, *International Statistical Review* **46**, 237–254.

[8] Mead, R., Curnow, R.N. & Hasted, A.M. (1993). *Statistical Methods in Agriculture and Experimental Biology*, Chapman & Hall, London.

[9] NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/.

[10] Odeh, R.E., Owen, D.B., Birnbaum, Z.W. & Fisher, L.D. (1977). *Pocket Book of Statistical Tables*, Marcel Dekker, New York.

[11] Stokes, M.E., Davis, C.S., & Koch, G., (1995). *Categorical Data Analysis Using the SAS System*, SAS Institute, Cary.

[12] Wayne, W.D. (1978). *Applied Nonparametric Statistics*, Houghton-Mifflin, New York.

[13] Wu, C.F.J. & Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*, John Wiley & Son, New York.

(*See also* **Distribution-free Inference, an Overview**)

Li Liu and Vance W. Berger

Randomized Block Designs

Scott E. Maxwell

Volume 4, pp. 1686–1687

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Randomized Block Designs

Sir Ronald A. Fisher's insight that random assignment to treatment groups probabilistically balances all possible confounders and thus allows for a causal inference is one of the most important statistical contributions of the twentieth century (*see* **Randomization**). This contribution undoubtedly explains the popularity of randomized designs in many disciplines, including the behavioral sciences. However, even though **nuisance variables** no longer need to be regarded as possible confounders with random assignment, these variables nevertheless continue to contribute to variance within groups in completely randomized designs. The magnitude of these effects is often substantial in the behavioral sciences, because for many phenomena of interest there are sizable systematic individual differences between individuals. The presence of such individual differences within groups serves to lower **power** and precision because variance due to such differences is attributed to error in completely randomized designs. In contrast, this variance is separately accounted for in the randomized block design (RBD) and thus does not inflate the variance of the error term, thus typically yielding greater power and precision in the RBD.

The basic idea of an RBD is first to form homogeneous blocks of individuals. Then individuals are randomly assigned to treatment groups within each block. Kirk [2] describes four types of possible dependencies: (a) repeated measures, (b) subject matching, (c) identical twins or littermates, and (d) pairs, triplets, and so forth, matched by mutual selection such as spouses, roommates, or business partners. The repeated measures (also known as *within-subjects*) application of the RBD (*see* **Repeated Measures Analysis of Variance**) is especially popular in psychology because it is typically much more efficient than a between-subjects design.

It is incumbent on the researcher to identify a blocking variable that is likely to correlate with the dependent variable of interest in the study (*see* **Matching**). For example, consider a study with 60 research participants comparing three methods of treating depression, where the Beck Depression Inventory (BDI) will serve as the dependent variable at the end of the study. An ideal blocking variable in this situation would be each individual's BDI score at the beginning of the study prior to treatment. If these scores are available to the researcher, he/she can form 20 blocks of 3 individuals each: the first block would consist of the 3 individuals with the 3 highest BDI scores, the next block would consist of the 3 individuals with the 3 next highest scores, and so forth. Then the researcher would randomly assign 1 person within each block to each of the 3 treatment groups. The random assignment component of the RBD resembles that of the completely randomized design, but the random assignment is said to be restricted in the RBD because it occurs within levels of the blocking variable, in this case, baseline score on the BDI.

The RBD offers two important benefits relative to completely randomized designs. First, the reduction in error variance brought about by blocking typically produces a more powerful test of the treatment effect as well as more precise estimates of effects. In this respect, blocking is similar to **analysis of covariance** [3, 4]. Second, a potential problem in a **completely randomized design** is **covariate** imbalance [1], which occurs when groups differ substantially from one another on a nuisance variable despite random assignment. The RBD minimizes the risk of such imbalance for the blocking variable. It is also important to acknowledge that blocking a variable that turns out not to be related to the dependent variable comes at a cost, because degrees of freedom are lost in the RBD, thus requiring a larger critical value and thereby risking a loss instead of a gain in power and precision.

References

[1] Altman, D.G. (1998). Adjustment for covariate imbalance, in *Encyclopedia of Biostatistics*, P. Armitage & T. Colton, eds, Wiley, Chichester, pp. 1000–1005.

[2] Kirk, R.E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, 3rd Edition, Brooks/Cole, Pacific Grove.

[3] Maxwell, S.E. & Delaney, H.D. (2004). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, Lawrence Erlbaum Associates, Mahwah.

[4] Maxwell, S.E., Delaney, H.D. & Dill, C.A. (1984). Another look at ANCOVA versus blocking, *Psychological Bulletin* **95**, 136–147.

(*See also* **Block Random Assignment**)

SCOTT E. MAXWELL

Randomized Response Technique

Brian S. Everitt

Volume 4, pp. 1687–1687

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Randomized Response Technique

The randomized response technique is an approach that aims to get accurate answers to a sensitive question that respondents might be reluctant to answer truthfully, for example, 'have you ever had an abortion?' The randomized response technique protects the respondent's anonymity by offering both the question of interest and an innocuous question that has a known probability (α) of yielding a 'yes' response, for example,

1. Flip a coin. Have you ever had an abortion?
2. Flip a coin. Did you get a head?

A random device is then used by the respondent to determine which question to answer. The outcome of the randomizing device is seen only by the respondent, not by the interviewer. Consequently when the interviewer records a 'yes' response, it will not be known whether this was a yes to the first or second question [2]. If the probability of the random device posing question one (p) is known, it is possible to estimate the proportion of yes responses to questions one (π), from the overall proportion of yes responses ($P = n_1/n$), where n is the total number of yes responses in the sample size n.

$$\hat{\Pi} = \frac{P - (1-p)\alpha}{p} \tag{1}$$

So, for example, if $P = 0.60$, $(360/600)p = 0.80$ and $\alpha = 0.5$, then $\hat{\Pi} = 0.125$. The estimated variance of $\hat{\Pi}$ is

$$\mathrm{Var}(\hat{\Pi}) = \frac{\hat{\Pi}(1-\hat{\Pi})}{n} + \frac{\begin{array}{c}(1-p)^2\alpha(1-\alpha) + p(1-p)\\ {[p(1-\alpha) + \alpha(1-p)]}\end{array}}{np^2} \tag{2}$$

For the example here, this gives $\mathrm{Var}(\hat{\Pi}) = 0.0004$.

Further examples of the application of the technique are given in [1].

References

[1] Chaudhuri, A. & Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*, Marcel Dekker, New York.

[2] Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association* **60**, 63–69.

BRIAN S. EVERITT

Range

David Clark-Carter

Volume 4, pp. 1687–1688

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Range

If the smallest score in a set of data is 10 and the largest is 300 then the range is $300 - 10 = 290$. The range has the advantage over just quoting the maximum and minimum values in that it is independent of the part of the scale where those scores occur. Thus, if the largest value were 400 and the smallest 110, then the range would still be 290.

A disadvantage of the range, compared to many other measures of spread, is that it is based solely on the numbers at the two ends of a set. Thus, if a single value at one extremity of a set of data is widely separated from the rest of the set, then this will have a large effect on the range. For example, the range of the set 5, 7, 9, 11, 13 is 8, whereas if the largest value were 200 instead of 13, then the range would be 195. (Notice that the same range would be produced when an extreme score occurred at the other end of the set: 5, 193, 195, 197, 200.) Hence, the range is better quoted alongside other measures of spread that are less affected by extreme scores, such as the **interquartile range**, rather than on its own.

DAVID CLARK-CARTER

Rank Based Inference

T.P. Hettmansperger, J.W. McKean and S.J. Sheather

Volume 4, pp. 1688–1691

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Rank Based Inference

In the behavioral sciences, a researcher often does not wish to assume that the measurements in an experiment came from one or more normal populations. When the population distributions are unknown (but have the same shape), an alternative approach to hypothesis testing can be based on the ranks of the data. For example, data on reaction times are typically skewed and **outliers** or unusual results can exist in situations where results from a new experimental treatment are being compared with those from a well established procedure. When comparing treatment and control groups, the two samples are combined and ranked. Then, rather than use the traditional two-sample t statistic that compares the sample averages, we use the Mann–Whitney–Wilcoxon (MWW) (*see* **Wilcoxon–Mann–Whitney Test**) statistic that compares the average ranks in the two samples. One advantage is that the null distribution of the MWW statistic does not depend on the common but unspecified shape of the underlying populations. Another advantage is that the MWW statistic is more robust against outliers and gross errors than the t statistic. Finally, the MWW test is more efficient (have greater **power**) than the t Test when the tails of the populations are only slightly heavier than normal and almost as efficient (95.5%) when the populations are normal. Consider also that the data may come directly as ranks. A group of judges may be asked to rank several objects and the researcher may wish to test for differences among the objects. The Friedman rank statistic (*see* **Friedman's Test**) is appropriate in this case.

Most of the simple experimental designs (one-sample, two-sample, one-way layout, two-way layout (with one observation per cell), correlation, and regression) have rank tests that serve as alternatives to the traditional tests that are based on least squares methods (sample means). For the designs listed above, alternatives include the Wilcoxon **signed rank test**, the MWW test, the **Kruskal–Wallis test**, the **Friedman test**, Spearman's rank correlation (*see* **Spearman's Rho**), and rank tests for regression coefficients (*see* **Nonparametric Regression**), respectively. There were just under 1500 citations to the above tests in social science papers as determined by the Social Science Citation Index (1956–2004). The Kruskal–Wallis (KW) statistic is used for the one-way layout and can be thought of as an extension of the MWW test for two samples. As in the case of the MWW, the data is combined and ranked. Then the average ranks for the treatments are compared to the overall rank average which is $(N + 1)/2$ under the null hypothesis of equal populations, where N is the combined sample size. When the average ranks for the treatments diverge sufficiently from this overall average rank, the null hypothesis of equal treatment populations is rejected. This test corresponds directly to the one-way analysis of variance F Test.

The nonparametric test statistics in the one- and two-sample designs have corresponding estimates associated with them. In the case of the Wilcoxon signed rank statistic, the corresponding estimate of the center of the symmetric distribution (including the mean and median) is the median of the pairwise averages of the data values. This estimate, called the *one-sample* **Hodges–Lehmann estimate**, combines the robustness of the median with the efficiency of averaging data from a symmetric population. In the two-sample case, the Hodges–Lehmann estimate of the difference in the locations of the two populations is the median of the pairwise differences. In addition, the test statistics can be inverted to provide **confidence intervals** for the location of the symmetric population or for the difference of locations for two populations. The statistical package, Minitab, provides these estimates and confidence intervals along with the tests. Reference [3] is an excellent source for further reading on rank-based methods. Next we discuss estimation and testing in the linear model.

In a general linear model, the least squares approach entails minimizing a sum of squared residuals to produce estimates of the regression coefficients (*see* **Multiple Linear Regression**). The corresponding F Tests are based on the reduction in sum of squares when passing from the reduced to the full model, where the reduced model reflects the null hypothesis. Rank methods in the linear model follow this same strategy. We replace the least squares criterion by a criterion that is based on the ranks of the residuals. Then we proceed in the same way as a least squares analysis. Good robustness and efficiency properties of the MWW carry over directly to the estimates and tests in the linear model. For a detailed account of this approach, see [3] for an applied perspective and [2] for the underlying theory with examples.

Computations are always an important issue. Many statistical packages contain rank tests for the simple designs mentioned above but not estimates. Minitab includes estimates and confidence intervals along with rank tests for the simple designs. Minitab also has an undocumented rank regression command rreg that follows the same syntax as the regular regression command. The website `http://www.stat.wmich.edu/slab/RGLM/` developed by McKean provides a broad range of tests, estimates, standard errors, and data plots for the general linear model. See reference [1] for additional discussion and examples based on the website. Below, we will illustrate the use of the website.

First we sketch the rank-based approach to inference in the linear model, and then we will outline how this approach works in a simple **analysis of covariance**. For $i = 1, \ldots, n$, let $e_i = y_i - \beta_0 - \mathbf{x}_i^T \boldsymbol{\beta}$ be the ith residual or error term, where $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$ are the known regression constants and $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$ are the unknown regression coefficients. The error distribution is assumed to be continuous and have median zero with no further assumptions. We denote by $F(e)$ and $f(e)$ the distribution and density functions of the errors, respectively. We call this the general linear model, since regression, **analysis of variance**, and analysis of covariance can be analyzed with this model. In matrix notation, we have $\mathbf{y} = \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ where $\mathbf{y}$ is an $n \times 1$ vector of responses, $\mathbf{1}$ is an $n \times 1$ vector of all ones, $\mathbf{X}$ is the $n \times p$ full rank design matrix, $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients, and $\mathbf{e}$ is the $n \times 1$ vector of errors.

In least squares, we minimize the criterion function Σe_i^2 to find the estimates of the regression coefficients. This least squares criterion function is equivalent to $\Sigma\Sigma_{i<j}(e_i - e_j)^2$ for inference on $(\beta_1, \ldots, \beta_p)$. For rank-based inference, we replace this criterion function by $\Sigma\Sigma_{i<j}|e_i - e_j|$. But $\Sigma\Sigma_{i<j}|e_i - e_j|$ is proportional to $D(\text{full}) = \surd 12 \Sigma(\text{Rank}(e_i) - (N+1)/2)e_i$. Hence, rather than a quadratic function of the residuals, we have a linear function of the residuals with the weights determined by the ranks of the residuals. The rank estimate of $\boldsymbol{\beta}$ is found by minimizing $D(\text{full})$, which is a measure of the dispersion of the full model residuals. For testing a null hypothesis, we write $D(\text{reduced})$ for the dispersion of the residuals with $\boldsymbol{\beta}$ constrained to the null hypothesis. Then $RD = \widehat{D}(\text{reduced}) - \widehat{D}(\text{full})$ is the reduction in dispersion as a result of fitting the reduced (null) model, where the hat indicates that D has been evaluated at the respective reduced and full model estimates of $\boldsymbol{\beta}$.

Throughout the inference, we need a scaling factor similar to σ^2, the variance of the error distribution, in least squares. This factor is $\tau = (\surd 12 \int f^2(e)de)^{-1}$, where f is the density function of the error distribution. In the case of normal errors, τ is equal to $\surd(\pi/3)\sigma$. This scaling parameter can be estimated using a density estimate based on the full model residuals.

This leads to the following basic results for inference: the rank-based estimate $\widehat{\boldsymbol{\beta}}$ is approximately normally distributed with mean $\boldsymbol{\beta}$ and covariance matrix $\tau^2 \mathbf{X}^T \mathbf{X}$, where $\mathbf{X}$ is the matrix of regression constants. The estimate of β_0 is the median of the residuals, $y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}$, and it is approximately normally distributed with mean β_0 and variance $(n4f^2(0))^{-1}$ when the design matrix has been centered. Further, for testing $H_0 : M\boldsymbol{\beta} = \mathbf{0}$, where M is a $q \times p$ matrix of full row rank, $F_R = 2RD/q\widehat{\tau}$ is approximately distributed as $F(q, n - p - 1)$. We now illustrate this approach on a simple analysis of covariance.

The website `http://www.stat.wmich.edu/slab/RGLM/` was used for computations for this example, and it can be used for a wide variety of models and designs. In this experiment, three advertising media (radio, newspaper, and television) were compared. The experimental units were 15 fast food restaurants located in comparable but different cities, five for each of the media. The response variable y was profits in thousands of dollars. The restaurants were roughly of the same size but had differing levels of food wastage. The percentage of food wastage x was used as a covariate. For example, there was 1.0% in the first restaurant under radio. The data is given in Table 1.

The website provides, along with the significance testing and estimation (with standard errors), data plots, **residual plots**, and standardized residual plots including **histograms**, **boxplots**, and Q–Q plots (*see* **Probability Plots**). We report here the results of testing for a covariate effect and for a media effect. We find $F_R = 2RD/q\widehat{\tau} = 92.6$ with a P value $=$ 0.0001 for the null hypothesis that all media are the same. For the null hypothesis that the coefficient of the covariate x is zero, the P value is effectively

Table 1 Profits in thousands of dollars(y) and Percent food wastage(x)

radio		newspaper		television	
x	y	x	y	x	y
1.0	30	2.1	24	3.4	17
1.4	18	2.6	20	3.9	11
1.9	13	3.1	7	4.3	3
2.5	6	3.6	4	4.7	−6
2.7	3	4.1	−5	5.2	−10

zero. Removing the covariate effect and considering data plots shows that the profits decline from radio to television. Least squares results in a very similar analysis. However, if $y = 6$ in the fourth row of radio is entered incorrectly as $y = 60$, the analysis is essentially the same for the rank tests but the least squares test is no longer significant for either of the null hypotheses, illustrating the relative robustness of the rank tests. In the original data, the coefficient of determination based on the rank approach is 0.90 while it is 0.97 for least squares. This coefficient is often used in data analysis to assess the quality of the fitted model. In the following paragraphs, we discuss robust coefficients of determination.

We now consider the correlation model in which $\mathbf{x}$ is a p-dimensional random vector with distribution function $M(\mathbf{x})$ and density function $m(\mathbf{x})$. We also let $H(\mathbf{x}, y)$ and $h(\mathbf{x}, y)$ denote the joint distribution and density functions of $(\mathbf{x}, y)$, respectively. Then $h(\mathbf{x}, y) = f(y - \beta_0 - \mathbf{x}^T\boldsymbol{\beta})m(\mathbf{x})$. We are interested in a robust measure of the relationship between y and $\mathbf{x}$; that is, a robust measure of association between y and $\mathbf{x}$. As with the traditional measure of determination, our robust measure will be zero if and only if y and $\mathbf{x}$ are independent. Hence, independence becomes the null hypothesis and this translates into $\beta = \mathbf{0}$ so that $h(\mathbf{x}, y) = f(y - \beta_0)m(\mathbf{x})$.

We consider the traditional measure first. Assume without loss of generality that $E(\mathbf{x}) = \mathbf{0}$ and $E(e) = 0$. Let $\mathrm{Var}(e) = \sigma_e^2$ and let $\Sigma = E(\mathbf{x}\mathbf{x}^T)$ denote the variance–covariance matrix of $\mathbf{x}$. Then the traditional coefficient of determination is given by

$$\overline{R}^2 = \frac{\boldsymbol{\beta}^T\Sigma\boldsymbol{\beta}}{\sigma_e^2 + \boldsymbol{\beta}^T\Sigma\boldsymbol{\beta}}. \qquad (1)$$

Note that $\overline{R}^2$ is a measure of association between y and $\mathbf{x}$. It lies between 0 and 1, and it is 0 if and only if y and $\mathbf{x}$ are independent, since y and $\mathbf{x}$ are independent if and only if $\boldsymbol{\beta} = \mathbf{0}$.

Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ be a random sample from the above correlation model. In order to obtain a consistent estimate of $\overline{R}^2$, treat $\mathbf{x}_i$ as fixed and estimate the parameters by least squares in the full model and then again in the reduced model in which $\boldsymbol{\beta} = \mathbf{0}$. The reduction in sum of squares is $SSR = SST - SSE$, and a consistent estimate of $\overline{R}^2$ is the familiar $R^2 = SSR/SST$. The traditional F statistic is $F = (SSR/p)/MSE$ where $MSE = SSE/(n - p - 1)$. Finally, recall that R^2 can be reexpressed as

$$R^2 = \frac{SSR}{SSR + (n - p - 1)SSE} = \frac{\frac{p}{n-p-1}F}{1 + \frac{p}{n-p-1}F}. \qquad (2)$$

We first introduce the robust estimate and then discuss the measure that it estimates (*see* **Robust Testing Procedures**). The rank test described above for testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is $F_R = 2RD/p\widehat{\tau}$, and RD is the reduction in dispersion when passing from the reduced to the full model. It is analogous to the reduction in sum of squares in the traditional approach based on least squares. We simply replace F by F_R in the expression above to get

$$R_{\text{rank}} = \frac{\frac{p}{n-p-1}F_R}{1 + \frac{p}{n-p-1}F_R} = \frac{RD}{RD + (n - p - 1)(\widehat{\tau}/2)}.$$

The statistic R_{rank} is robust because it is a one-to-one function of the robust test statistic F_R. Further, it lies between 0 and 1, having the values 1 for a perfect fit and 0 for a complete lack of fit. These properties make R_{rank} an attractive coefficient of determination for regression problems. Note that R_{rank} is a ratio of scales, while R^2 is a ratio of variances. In general, R_{rank} estimates a different parameter than R^2. Thus caution is needed in comparing the two statistics. However, like R^2, R_{rank} can be used for comparison of hierarchical models.

The statistic R_{rank} is a consistent estimate for $\overline{R}_{\text{rank}} = \overline{RD}/(\overline{RD} + \tau/2)$, where

$$\overline{RD} = \int \sqrt{12}(G(y) - \tfrac{1}{2})yg(y)dy - \int \sqrt{12}(F(y) - \tfrac{1}{2})yf(y)dy, \qquad (3)$$

and $G(y)$ and $g(y)$ are the marginal distribution and density functions of y, respectively. This measure is between 0 and 1 and is 0 if and only if y and $\mathbf{x}$ are independent. This is the robust coefficient of determination. In general, $\overline{R}_{\text{rank}}$ and $\overline{R}^2$ differ, they are one-to-one functions of each other when $(\mathbf{x}, y)$ has a multivariate normal distribution. Define

$$\overline{R}^*_{\text{rank}} = 1 - \left[\frac{1 - \overline{R}_{\text{rank}}}{1 - \overline{R}_{\text{rank}}(1 - \pi/6)}\right]^2. \qquad (4)$$

Then, under multivariate normality, $\overline{R}^*_{\text{rank}} = \overline{R}^2$. The corresponding statistic is R^*_{rank} which is defined in terms of R_{rank}. Now, under multivariate normality (*see* **Catalogue of Probability Density Functions**), R^*_{rank} and R^2 estimate the same quantity.

References

[1] Abebe, A., Crimin, K., McKean, J.W., Haas, J. & Vidmar, T. (2001). Rank-based procedures for linear models: applications to pharmaceutical science data, *Drug Information Journal* **35**, 947–971.

[2] Hettmansperger, T.P. & McKean, J.W. (1998). *Robust Nonparametric Statistical Methods*, Arnold Publishing, London.

[3] Hollander, M. & Wolfe, D.A. (1999). *Nonparametric Statistical Methods*, John Wiley, New York.

T.P. Hettmansperger, J.W. McKean and S.J. Sheather

Rasch Modeling

Gerhard H. Fischer

Volume 4, pp. 1691–1698

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Rasch Modeling

History of the Rasch Model

In 1952, the Danish mathematician Georg Rasch (1901–1980), then consultant for a project of the Ministry of Social Affairs, introduced a multiplicative Poisson model for the analysis of reading errors of school children. He considered the number of errors made by testee S_v in text I_i as a realization of a Poisson variable with parameter λ_{vi} (*see* **Catalogue of Probability Density Functions**) measuring the testee's 'proneness' to errors when reading that particular text. He then split λ_{vi} into a factor pertaining to the testee, S_v's reading ability θ_v, and a factor pertaining to the text, the difficulty δ_i of text I_i. To him as a mathematician it was immediate, in virtue of a well-known theorem about Poisson variables, to draw the following conclusion: if a testee had read two texts I_i and I_j, the probability of observing k_{vi} and k_{vj} errors in these two texts, conditional on the total sum of $k_{vi} + k_{vj} = k_{v.}$ errors, had to follow a Binomial distribution characterized by $k_{v.}$ and parameter $\pi = 1/(1 + \delta_j/\delta_i)$ (*see* **Catalogue of Probability Density Functions**).

This opened Rasch's eyes for a novel approach to measurement in behavioral science: since parameter π no longer depended on the testee parameter θ_v, data from all testees who had read the same two texts could be pooled to obtain a (**maximum likelihood**) estimate (*see* **Maximum Likelihood Estimation**) of π, from which in turn an estimate of the quotient of the text difficulties, δ_i/δ_j, was obtainable. This enabled a measurement of the *relative* difficulties of the texts – in a specific sense – independently of the sample of testees, and this procedure could of course be extended to the simultaneous comparison of more than two texts. He considered such comparisons of text difficulties as *objective* because the result was generalizable over particular children tested, whatever their individual reading abilities might have been. Later Rasch [19] denoted functions that made such a comparison feasible *comparators*, and comparisons carried out in that manner *specifically objective*.

Instrumental for this was the splitting of parameter λ into a product of a testee's ability and the text's difficulty. When Rasch in 1953 analyzed intelligence test data for the Danish army, he decided to carry the same principle over to the area of test analysis. First, he sought for a suitable *item response function* (IRF) $P(+|S_v, I_i) = f(\xi_v, \epsilon_i)$, where '+' denoted a correct response, ξ_v the testee's ability, and ϵ_i the easiness of test item I_i; as a particularly simple algebraic function mapping the positive reals on the semiopen interval $[0, 1)$, he chose $f = x/(1 + x)$. Then he conceived x to be a product of the testee's ability ξ_v and the item's easiness ϵ_i, namely, $x = \xi_v\epsilon_i$, with $\xi_v \geq 0$ and $\epsilon_i \geq 0$. This model is now generally denoted the 'Rasch Model' (RM), but is usually reparameterized by taking the logarithms rather than Rasch's original item and person parameters: $\theta_v = \ln(\xi_v)$ as testee *ability* and $\beta_i = -\ln(\epsilon_i)$ as item *difficulty*.

Definition and Some Basic Properties of the RM

The RM for dichotomous responses (denoted '+' and '−') is defined as

$$P(X_{vi} = 1|\theta_v, \beta_i) = \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)}, \qquad (1)$$

where X_{vi} is the response variable with realization $x_{vi} = 1$ if testee S_v gives response '+' to I_i, and $x_{vi} = 0$ if S_v's response to I_i is '−'; $-\infty < \theta_v < \infty$ is the latent ability of S_v, and $-\infty < \beta_i < \infty$ the difficulty of item I_i. All item responses are assumed to be 'locally independent', that is, the probability of any response pattern in a test of length k is the product of k probabilities (1) or complements thereof. Parameters θ_v and β_i are defined only up to an arbitrary additive normalization constant c; the latter is usually specified by either setting one item parameter to zero (e.g., $\beta_1 = 0$), or setting $\sum_i \beta_i = 0$.

The form of (1) implies that all IRFs are parallel curves. This means that, if the RM is to fit a set of data, all items must have (approximately) *equal discrimination*.

The probability of a complete $(n \times k)$ item score matrix $\boldsymbol{X}$ as a function of the unknown parameters (i.e., the so-called *likelihood function*) can be shown to depend only on the marginal sums of $\boldsymbol{X}$, namely, on the *raw scores* $r_v = \sum_i x_{vi}$ and the item marginals $x_{.i} = \sum_v x_{vi}$. This implies that the r_{vi} and the $x_{.i}$ contain all the relevant information in the data with respect to the parameters (i.e., are jointly *sufficient statistics-see* **Estimation**), whereas the individual response patterns yield no additional information about the parameters. This is a remarkable asset of the

RM: throughout a century of psychological testing, the number-correct (or raw) score in intelligence and other attainment tests has been employed as a summary of a testee's test achievement; if the RM holds for a particular test, this fact yields a rigorous justification for the use of the raw scores. Therefore, ascertaining whether the RM fits a given set of test data is an enterprise of considerable practical importance.

Another noteworthy property of the RM is that the conditional probability (or likelihood) of an item score matrix $\boldsymbol{X}$, given all raw scores r_v, is a function of the item parameters only. (This parallels Rasch's earlier observation about the conditional distribution of the number of reading errors in two texts, given the testee's total number of reading errors.) The conditional probability can thus serve as a comparator for the item parameters, enabling specifically objective comparisons of the item difficulties. (Symmetrically, specifically objective comparisons between persons are also possible, but carrying them out directly is impractical under most realistic conditions. The way to compare persons is to first estimate all item parameters and then to estimate the person parameters by considering the item parameters as given constants.)

These two favorable properties of the RM – sufficiency of the raw scores and of the item marginals, specifically objective comparisons between items and between testees – raises the question of whether other models exist that share these properties with the RM. Within a framework of locally independent items with continuous, strictly monotone IRFs with lower limits zero (i.e., no guessing) and upper limits one, the answer is 'no'. It can be shown that, within this framework, a *family* of RMs follows from either of the following assumptions: (a) sufficiency of the raw score for the person parameter; (b) existence of a nontrivial likelihood function that can serve as a specifically objective comparator for the items or the testees; (c) existence of a nontrivial sufficient statistic for the testee parameter that is independent of the item parameters; and (d) stochastically consistent ordering of the items. (On formal definitions and proofs, see Chapter 2 of [7].) The term 'family of RMs' refers to all models of the form (1) where, however, the parameters θ_v and β_i are replaced by $a\theta_v + c$ and $a\beta_i + c$, respectively. Therein, the constant c, which immediately cancels from (1), is the normalization constant mentioned above, and $a > 0$ is an unspecified *discrimination* parameter, namely, the maximal slope of the (parallel) IRFs. One may set $\alpha = 1$, of course, which immediately yields the RM, but it has to be kept in mind that this specification is arbitrary.

These results have important consequences regarding an age-old question in behavioral science: What are the measurement properties of psychological or educational tests? If, for a given set of data $\boldsymbol{X}$, the RM is found to fit and the parameters are adequately estimated, the scales of the parameters θ and β are unique only up to linear transformations $a\theta + c$ and $a\beta + c$, with arbitrary $a > 0$ and arbitrary normalization constant c. From this it is concluded that the scales are *interval scales* (*see* **Scales of Measurement**) with a common measurement unit.

There is one more *caveat*, though: The proofs leading to these conclusions rest on the assumption that there are a testee population and a *dense* universe of items such that both the person and the item parameters can vary continually. While the first assumption appears to be acceptable – for instance, growth of ability is usually conceived as continuous – the second assumption may be questioned. In the special case, however, that the item parameters are assumed to be k fixed discrete *rational* numbers, the conclusion about the measurement properties of the RM becomes somewhat weaker: abilities are measurable only on an *ordered metric scale* which has interval scale properties at certain lattice points (which are typically spaced narrowly), but elsewhere allows only an ordering of abilities. For practical purposes, however, the scale can still be considered an interval scale.

Parameter Estimation and Testing of Fit

Various approximate methods for parameter estimation have been proposed by Rasch and some of his students, but under the perspective that estimators should be *unique*, *statistically consistent*, and should have known *asymptotic properties*, only two approaches seem to prevail. The probably most attractive method is the *conditional maximum likelihood* (CML) method which maximizes the conditional probability of $\boldsymbol{X}$, given the raw scores r_v, in terms of the item parameters. As mentioned above, this likelihood function is independent of the person parameters and thus is a comparator function – in the spirit of Rasch – for establishing specifically objective comparisons of the items. In finite samples, the normalized CML estimates are unique if a certain directed graph C associated with matrix $\boldsymbol{X}$ is *strongly connected*, see [7]; checking this weak connectivity condition is easy in practice, using standard tools of graph theory. If a weak necessary and sufficient condition given by Pfanzagl [16] is met by the distribution of the person parameters in the respective testee population, the estimators are consistent and asymptotically normally distributed around the true parameters β_i for fixed test length k and $n \to \infty$. The standard errors of the estimates $\hat{\beta}_i$ can be determined from the **information matrix**. The estimator is not *efficient*, though, but the loss of statistical information entailed by conditioning on the raw scores is very slight, see [5]. Programs for CML estimation are, for example, LPCM-Win [8] and OPLM [21].

Another approach to the estimation of the β_i is the *marginal maximum likelihood* (MML) method. In the so-called *parametric* MML, a latent population distribution of the θ-parameters is specified, for instance, a normal distribution with unknown mean μ and standard deviation σ. To calculate the likelihood of the data under such an enhanced RM requires integration over θ, leading to the elimination of the person parameters from the likelihood function. The latter is then maximized with respect to both the item and distribution parameters. If the assumption about the latent population distribution happens to be true, then the parametric MML method is consistent and asymptotically efficient. If, on the other hand, the distributional assumption is not true, then the MML method can be strongly biased and even loses the property of consistency. A popular program for parametric MML estimation in the RM (and also for more general item response models like the two- and three-parameter logistic models) is BILOG [14].

Yet another method is *nonparametric* MML where the latent distribution is replaced by a step function with at most $(k+2)/2$ (if k is even) or $(k+1)/2$ (if k is odd) nodes on the θ-axis. The respective areas under the step function are then estimated along with the item parameters. De Leeuw and Verhelst [4] have shown that with the RM this method is asymptotically equivalent to CML. Pfanzagl [16] has proved that the RM is the only model where, under the assumption of an unknown latent distribution of θ, the item parameters are identifiable via nonparametric MML. (This is one more argument for the RM.)

From a practical point of view, the two most promising candidates among the estimation methods seem to be CML and parametric MML. (Nonparametric MML is asymptotically equivalent to CML and thus needs not be considered separately.) The choice should depend on the researcher's faith in a particular latent population distribution. Specifying the latent distribution, however, is always problematic: it has to be kept in mind that assuming, for instance, a normal distribution for a given population makes it practically impossible that the same holds for subpopulations like the male and female testees as well. Choosing CML implies a slight loss of precision of the estimates but, on the other hand, circumvents distributional assumptions that may entail a serious bias of the estimators.

Closely related to the question of item parameter estimation is the problem of testing of fit: only if the statistical consistency and asymptotic distribution (i.e., normality) of the estimators have been verified, it becomes possible to establish powerful test methods for assessing fit and/or testing other hypotheses on the parameters. Rasch [18] had mainly used heuristic graphical methods for the assessment of fit by checking whether the most characteristic property of the RM – independence of the item parameter estimates of the testees' abilities – could be empirically verified. Andersen [2] has shown that this null hypothesis can be tested by means of a *conditional likelihood ratio* (CLR) test: the sample is split in two or more subsamples which differ significantly with respect to their raw score distributions (e.g., in testees with above versus below average raw scores), and the item parameter estimates in these subgroups are compared with those of the total group by means of the respective conditional likelihoods. The same can be done by splitting the sample on the basis of an external criterion like age, gender, education, etc. Some authors, however, observed that these tests sometimes fail to reject the null hypothesis when there actually is lack of fit. Glas and Verhelst, see Chapter 5 of [7], therefore developed several powerful tests of fit, both global and item-wise. Ponocny [17] proposed nonparametric tests for hypotheses than can be chosen arbitrarily, based on a Monto-Carlo procedure. Klauer [11] and Fischer [6] developed exact single-case tests and a CML estimation method for the amount of change of θ_v between two time points, for instance, for the assessment of development or of a treatment effect in an individual.

Once the item parameters have been estimated and the fit of the RM has been established satisfactorily, the person parameters can be easily estimated by straightforward maximum likelihood. Many researchers see a disadvantage of ML estimates of θ in the fact that for raw score $r_v = 0$ the person parameter estimate diverges to $-\infty$, and for $r_v = k$, to ∞. Warm [22] has therefore suggested a weighted ML estimator that yields finite values even for these extreme scores. (On further details about person parameter estimation, see [10]).

Remarks on Application of the RM, and an Example

A very wide range of applications of the RM both to achievement tests and questionnaires with dichotomous item format is seen in the literature. These abundant applications are often motivated as follows. If a scale consisting of dichotomous items has been constructed with the intention to measure a single latent trait via the simple raw score, then the RM should hold for that scale. Therefore, the RM is applied to check these assumptions. Two necessary conditions, however, are often overlooked: first, undimensionality in the sense of the RM (*see* **Item Response Theory (IRT) Models for Dichotomous Data**) is a very strict requirement that can hold, for example, in a subscale of an intelligence test for items with homogeneous content, such like those of the Standard Progressive Matrices test or Gittler's [9] cubes test of spatial ability; but unidimensionality is extremely unlikely to occur in omnibus intelligence scales or, even more so, in questionnaires. Second, the IRFs are assumed (a) to tend to the lower limit zero and (b) the upper limit one. In order to satisfy these requirements, the response format must exclude guessing (or, at least, the guessing probabilities must be very low); this implies that a correct response occurs only if the testee has *solved* the item, so that this event is a reliable indicator of the respective ability. Technically, in questionnaires with response format 'yes' versus 'no' (or '+' versus '−') there is an extremely high guessing probability – namely, 0.50 – and hence there can be no certainty that the response '+' is an indicator of the trait of interest; the testee could as well have checked the response categories arbitrarily. Therefore, most applications of the dichotomous RM to questionnaire scales are of little scientific value. (On generalized Rasch models for polytomous ordered response items, see below.)

To illustrate the procedure of an application of the RM, the analysis of a data sample from $n = 1160$ testees who took Gittler's '3DW' test [9] of spatial ability is now sketched. In each of the $k = 17$ items, a cube X is presented to the testee which is known to display different patterns on each of its sides, however, only three of them can be seen. Simultaneously with X, six other cubes are also presented, one of which may be equal to X but is shown in a rotated position. The testee is asked to point out which of the latter cubes is the same as X. The testee is moreover instructed to choose one of the alternative response categories 'None of the cubes is equal to X' or 'Don't know' if she/he feels uncertain whether any of the cubes equals X.

First the item parameters β_i are estimated by means of the CML method for the total sample. Then the data are split in two subsamples with raw scores below versus above the average raw score (denoted 'Group L' and 'Group H' for short). By virtue of the property of specific objectivity, the item parameter estimates in these two subgroups, denoted $\hat{\beta}_i^L$ and $\hat{\beta}_i^H$, should be the same except for sampling errors. Therefore, the points with coordinates $(\hat{\beta}_i^L, \hat{\beta}_i^H)$ in Figure 1 should fall near the straight line through the origin with slope 1. As can be seen, this is the case; the ellipses around the $k = 17$ points can be interpreted as confidence regions for $\alpha = 0.05$ with semiaxes $1.96\sigma_i^L$ in the horizontal and $1.96\sigma_i^H$ in the vertical direction. None of the points falls significantly apart from the line with slope 1.

This graphical control of the model is complemented by Andersen's [2] asymptotic CLR test: the log-likelihood is $\ln L_T = -7500.95$ for the total sample, $\ln L_L = -3573.21$ for Group L, and $\ln L_H = -3916.91$ for Group H. Therefore, Andersen's test statistic is $\chi^2 = -2[\ln L_T - (\ln L_L + \ln L_H)] = -2$ $[-7500.95 - (-3573.21 - 3916.91)] = 21.65$ with $\mathrm{df} = 16$, which is nonsignificant for $\alpha = 0.05$ (the critical value being $\chi_\alpha^2 = 26.29$). The H_0 that the RM fits the data is therefore retained under this test. (Similarly, when the sample is split by either of the external criteria gender, age, or education level, the respective test statistics are also nonsignificant.) This supports the hypothesis that the RM fits the data sufficiently well.

Another graphical tool for the assessment of fit is shown in Figure 2 (for three arbitrarily selected items

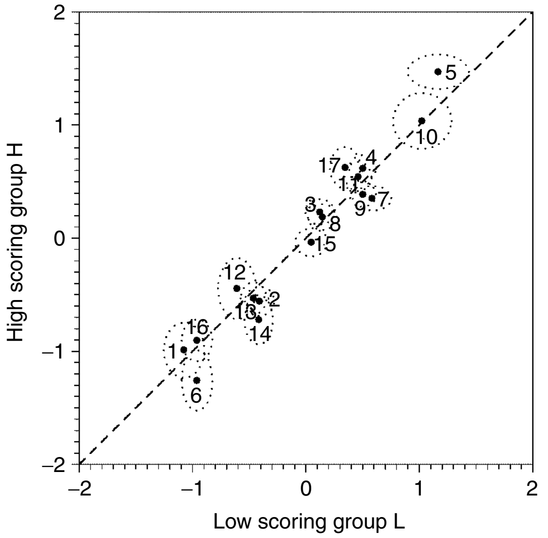


Figure 1 Graphic control of the RM for the 3DW items

of different difficulty levels, I_6 with $\hat{\beta}_6 = -1.06$, I_{10} with $\hat{\beta}_{10} = 1.02$, and I_{15} with $\hat{\beta}_{15} = -0.01$). It shows the IRFs (1) based on the parameter estimates as a function of θ, and 'empirical IRFs' based on the relative frequencies of correct responses in the different raw score subgroups. These relative frequencies were slightly smoothed using a so-called *normal kernel smoother* with a bandwidth of 3 adjacent frequencies at the margins and bandwidth 5 elsewhere. The confidence intervals for the estimates of the ordinates for $\alpha = 0.05$ are shown as dotted lines. (The graphs for the remaining items are similar.) This nicely supports the hypothesis that the RM fits the items.

If the researcher has satisfied him(her)self that the RM fits, further advantages can be taken of the particular properties of the model. For instance, under the assumption that the item parameters have been estimated sufficiently well, the estimates can be considered as known constants and *exact* conditional tests can be made of the H_0 that the person parameters of two individuals are equal, or that the person parameters of the same person tested at two time points or under two different testing conditions are equal. 'Exact' means that the tests are based on the exact conditional distribution of the two scores, given their sum, rather than on asymptotic theory. To make such tests, it is not required that the two persons (or the same person at two time points) take exactly the same items: it suffices to present two item samples from the (unidimensional) item pool for which the RM has been found to hold. These item samples may be identical, or overlapping, or disjoint. In the present case, given that the item pool comprises only $k = 17$ items, there would be little point in choosing different (i.e., smaller) subsamples of items. Suppose therefore that the complete test is given to two testees (or the same testee twice), and the H_0 is to be tested that the two person parameters are equal (i.e., the H_0 of 'no change'). The empirical researcher needs only to apply Table 1 which gives the significance levels for all possible score combinations r_1 and r_2 of the two testees (or of one testee at two time points) under the specified H_0, based on the so-called '*Mid-P-Method*' (cf. [6]). For example, in the following score

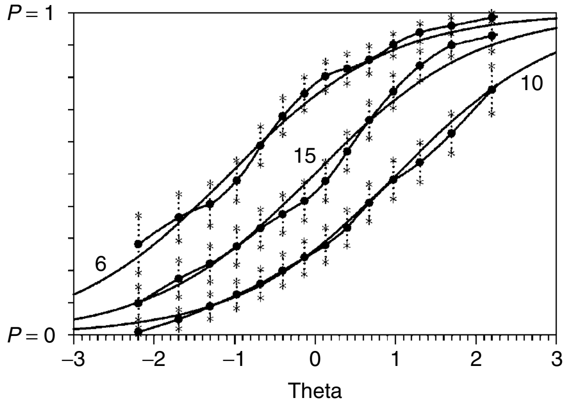


Figure 2 Theoretical and empirical IRFs for items 6, 10, and 15

Table 1 Significances of score combinations in the 3DW test (two-sided exact tests with $\alpha = 0.05$)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
0					s	s	S	S	T	T	T	T	T	T	T	T	T	T	0
1						.	s	s	S	S	T	T	T	T	T	T	T	T	1
2								.	s	S	S	S	T	T	T	T	T	T	2
3									.	s	s	S	S	T	T	T	T	T	3
4	s									.	s	s	S	S	T	T	T	T	4
5	s	.									.	s	s	S	S	T	T	T	5
6	S	s										.	s	s	S	S	T	T	6
7	S	s	.										.	s	s	S	T	T	7
8	T	S	s	.										.	s	S	S	T	8
9	T	S	S	s	.										.	s	S	T	9
10	T	T	S	s	s	.										.	s	S	10
11	T	T	S	S	s	s	.										s	S	11
12	T	T	T	S	S	s	s	.									.	s	12
13	T	T	T	T	S	S	s	s	.									s	13
14	T	T	T	T	T	S	S	s	s	.									14
15	T	T	T	T	T	T	S	S	S	s	.								15
16	T	T	T	T	T	T	T	T	S	S	s	s	.						16
17	T	T	T	T	T	T	T	T	T	T	S	S	s	s					17
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	

Note: Rows correspond to r_1, columns to r_2. Entries '.' denote significance level 0.10, 's' level 0.05, 'S' level 0.01, and 'T' level 0.001.

combinations (r_1, r_2), score r_2 would be the lowest score that is significantly higher than score r_1, under a two-sided alternative hypothesis with $\alpha = 0.05$: (0,4), (1,6), (2,8), (3,9), (4,10), (5,11), (6,12), (7,13), (8,14), (9,15), (10,16), (11,16), (12,17), (13,17).

The simplicity of the application of this table illustrates the advantages of having a test to which the RM can be fitted. Therefore, it is recommended to employ the RM whenever possible as a guideline for test development rather than just as a means for *post hoc* analysis of test data.

Some Extensions of the RM

Many extensions of the RM and of Rasch's approach to measurement have been proposed. One group of them comprises models for dichotomous item responses, the other polytomous response models.

A. Dichotomous extensions. The *linear logistic test model* (LLTM) assumes that the item difficulty parameters of an RM can be explained as weighted sums of certain *basic parameters* assigned, for instance, to rules or cognitive operations involved in the solution process. The CML estimation method, uniqueness conditions, asymptotic properties, and conditional likelihood ratio tests generalize directly form the RM to the LLTM. Studies trying to explain item parameters as weighted sums of parameters of cognitive operations or rules have often failed to be successful, however, because of lack of fit. The primary importance of the LLTM therefore seems to lie in studies with experimental or longitudinal designs (*see* **Clinical Trials and Intervention Studies** and **Longitudinal Data Analysis**) where the basic parameters are, for instance, effects of treatments given prior to – or experimental conditions prevailing at – the testing occasion. The LLTM can moreover be reformulated as a *multidimensional* longitudinal model, assigning different latent dimensions to different items without, however, requiring assumptions about the true dimensionality of the latent space. This *linear logistic model with relaxed assumptions* (LLRA) can serve to measure effects of treatments in educational, applied, or clinical psychology. (On these models, see Chapters 8 and 9 of [7].) Analogous to the LLTM is Linacre's [12] FACETS model. Another kind of dichotomous generalization of the RM is the *mixed* RM by Rost [20], which assumes the existence of $c > 1$ latent classes between which the item parameters of the RM are allowed to differ. Yet another direction of generalization aims at relaxing the assumption of equal discrimination of all items: the *one parameter logistic model* (OPLM) by Verhelst et al. [21] presumes a small set of discrete rational discrimination parameters α_i and assigns one of them, by means of a heuristic procedure, to each item. The difference between the OPLM and the *two-parameter logistic (or Birnbaum) model* lies in the nature of the discrimination parameters: in the latter, they are free parameters, while in the OPLM they are constants chosen by hypothesis. Sophisticated test methods are needed, of course, to verify or reject such hypotheses. Statistics to test the fit of the OPLM are given by Verhelst and Glas in Chapter 12 of [7].

B. Polytomous extensions. Andrich [3] and Masters [13] have introduced undimensional rating scale models for items with ordered response categories (like ‘strongly agree’, ‘rather agree’, ‘rather disagree’, ‘strongly disagree’), namely, the *rating scale model* (RSM) and the *partial credit model* (PCM). The former assumes equal response categories for all items, whereas the latter allows for a different number and different definition of the response categories per item. The RM is a special case of the RSM, and the RSM a special case of the PCM. Fischer and Ponocny (see Chapter 19 of [7]) have embedded a linear structure in the parameters of these models analogously to the LLTM mentioned above. CML estimation, conditional hypothesis tests, and multidimensional reparameterizations are generalizable to this framework, as in the LLTM. These models are particularly suited for longitudinal and treatment effect studies. The individual-centered exact conditional tests of change are also applicable in these polytomous models (cf. [6]). A still more general class of multidimensional IRT models with linear structures embedded in the parameters has been developed by Adams et al. [1, 23]; these authors rely on parametric MML methods. Müller [15], moreover, has proposed an extension of the RM allowing for continuous responses.

References

[1] Adams, R.J., Wilson, M. & Wang, W. (1997). The multidimensional random coefficients multinomial logit model, *Applied Psychological Measurement* **21**, 1–23.

[2] Andersen, E.B. (1973). A goodness of fit test for the Rasch model, *Psychometrika* **38**, 123–140.

[3] Andrich, D. (1978). A rating formulation for ordered response categories, *Psychometrika* **43**, 561–573.

[4] De Leeuw, J. & Verhelst, N.D. (1986). Maximum likelihood estimation in generalized Rasch models, *Journal of Educational Statistics* **11**, 183–196.

[5] Eggen, T.J.H.M. (2000). On the loss of information in conditional maximum likelihood estimation of item parameters, *Psychometrika* **65**, 337–362.

[6] Fischer, G.H. (2003). The precision of gain scores under an item response theory perspective: a comparison of asymptotic and exact conditional inference about change, *Applied Psychological Measurement* **27**, 3–26.

[7] Fischer, G.H. & Molenaar, I.W. eds (1995). *Rasch Models: Foundations, Recent Developments, and Applications*, Springer-Verlag, New York.

[8] Fischer, G.H. & Ponocny-Seliger, E. (1998). *Structural Rasch Modeling. Handbook of the Usage of LPCM-WIN 1.0*, ProGAMMA, Groningen.

[9] Gittler, G. (1991). *Dreidimensionaler Würfeltest (3DW)*, [The three-dimensional cubes test (3DW).] Beltz Test, Weinheim.

[10] Hoijtink, H. & Boomsma, A. (1996). Statistical inference based on latent ability estimates, *Psychometrika* **61**, 313–330.

[11] Klauer, K.C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model, *Psychometrika* **56**, 213–228.

[12] Linacre, J.M. (1996). *A User’s Guide to FACETS*, MESA Press, Chicago.

[13] Masters, G.N. (1982). A Rasch model for partial credit scoring, *Psychometrika* **47**, 149–174.

[14] Mislevy, R.J. & Bock, R.D. (1986). *PC-Bilog: Item Analysis and Test Scoring with Binary Logistic Models*, Scientific software, Mooresville.

[15] Müller, H. (1987). A Rasch model for continuous ratings, *Psychometrika* **52**, 165–181.

[16] Pfanzagl, J. (1994). On item parameter estimation in certain latent trait models, in *Contributions to Mathematical Psychology, Psychometrics, and Methodology*, G.H. Fischer & D. Laming, eds, Springer-Verlag, New York, pp. 249–263.

[17] Ponocny, I. (2001). Non-parametric goodness-of-fit tests for the Rasch model, *Psychometrika* **66**, 437–460.

[18] Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*, Pædagogiske Institut, Copenhagen. (Expanded edition 1980. The University of Chicago Press, Chicago.).

[19] Rasch, G. (1977). On specific objectivity. An attempt at formalizing the request for generality and validity of scientific statements, in *The Danish Yearbook of Philosophy*, M. Blegvad, ed., Munksgaard, Copenhagen, pp. 58–94.

[20] Rost, J. (1990). Rasch models in latent classes: an integration of two approaches to item analysis, *Applied Psychological Measurement* **3**, 397–409.

[21] Verhelst, N.D., Glas, C.A.W. & Verstralen, H.H.F.M. (1994). *OPLM: Computer Program and Manual*, CITO, Arnhem.

[22] Warm, T.A. (1989). Weighted likelihood estimation of ability in item response models, *Psychometrika* **54**, 427–450.

[23] Wu, M.L., Adams, R.J. & Wilson, M.R. (1998). *ACER ConQest*, Australian Council for Educational Research, Melbourne.

GERHARD H. FISCHER

Rasch Models for Ordered Response Categories

David Andrich

Volume 4, pp. 1698–1707

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Rasch Models for Ordered Response Categories

Introduction

This entry explains the latent response structure and process compatible with the **Rasch model** (RM) for ordered response categories in standard formats (*see* **Rasch Modeling**). It considers the rationale for the RM but is not concerned with details of parameter estimation and tests of fit. There are a number of software packages that implement the RM at an advanced level. Detailed studies of the theory and applications of the RM can be found in [10], [11], [16], and [17].

Standard formats involve one response in one of the categories deemed *a priori* to reflect levels of the latent trait common in quantifying attitude, performance, and status in the social sciences. They are used by analogy to measurement in the natural sciences. Table 1 shows typical formats for four ordered categories.

The Model and its Motivation

The RM was derived from the following requirement of invariant comparisons:

The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison.

Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison [14, p. 332].

Rasch was not the first to require such invariance, but he was the first to formalize it in the form of a probabilistic mathematical model. Following a sequence of derivations in [14], [1], and [3], the model was expressed in the form

$$P\{X_{ni} = x\} = \frac{1}{\gamma_{ni}} \exp\left(-\sum_{k=0}^{x} \tau_k + x(\beta_n - \delta_i)\right), \tag{1}$$

where (a) $X_{ni} = x$ is an integer random variable characterizing $m + 1$ successive categories, (b) β_n and δ_i are respectively locations on the same latent continuum of person n and item i, (c) τ_k, $k = 1, 2, 3, \ldots, m$ are m thresholds which divide the continuum into to $m + 1$ ordered categories and which, without loss of generality, have the constraint $\sum_{k=0}^{m} \tau_k = 0$, and (d) $\gamma_{ni} = \sum_{x=0}^{m} \exp(-\sum_{k=0}^{x} \tau_k + x(\beta_n - \delta_i))$ is a normalizing factor that ensures that the probabilities in (1) sum to 1. For convenience of expression, though it is not present, $\tau_0 \equiv 0$. The thresholds are points at which the probabilities of responses in one of the two adjacent categories are equal.

Figure 1 shows the probabilities of responses in each category, known as category characteristic curves (CCCs) for an item with three thresholds and four categories, together with the location of the thresholds on the latent trait.

In (1) the model implies equal thresholds across items, and such a hypothesis might be relevant in example 3 of Table 1. Though the notation used in (1) focuses on the item–person response process and does not subscript the thresholds with item i, the derivation of the model is valid for the case that different items have different thresholds, giving [20]

$$P\{X_{ni} = x\} = \frac{1}{\gamma_{ni}} \exp\left(-\sum_{k=1}^{x} \tau_{ki} + x(\beta_n - \delta_i)\right), \tag{2}$$

in which the thresholds τ_{ki}, $k = 1, 2, 3, \ldots m_i$, $\sum_{k=0}^{m_i} \tau_{ki} = 0$, are subscripted by i as well as $k.\tau_{0i} \equiv 0$. Such differences in thresholds among items might be required in examples of 1 and 2 of Table 1.

Models of (1) and (2) have become known as the *rating scale* model and *partial credit* models

Table 1 Standard response formats for the Rasch model

Example	Category 1		Category 2		Category 3		Category 4
1	Fail	<	Pass	<	Credit	<	Distinction
2	Never	<	Sometimes	<	Often	<	Always
3	Strongly disagree	<	Disagree	<	Agree	<	Strongly agree

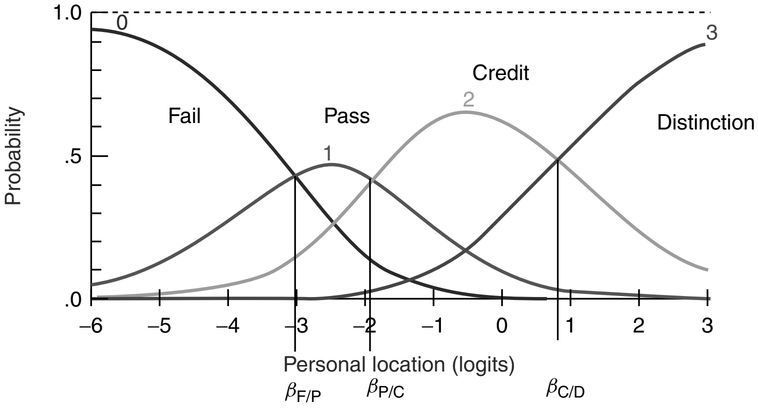


Figure 1 Category characteristic curves showing the probabilities of responses in each of four ordered categories

respectively. Nevertheless, they have an identical response structure and process for a single person responding to a single item.

Let $\delta_{ki} = \delta_i + \tau_{ki}$, $\delta_{0i} \equiv 0$. Then (2) simplifies to

$$P\{X_{ni} = x\} = \frac{1}{\gamma_{ni}} \exp \sum_{k=1}^{x} (\beta_n - \delta_{ki})$$
$$= \frac{1}{\gamma_{ni}} \exp \left(x\beta_n - \sum_{k=1}^{x} \delta_{ki} \right). \quad (3)$$

In this form, the thresholds δ_{ki} are immediately comparable across items; in the form of (2), the τ_{ki} are referenced to the location $\delta_i$of item i, which is the mean of the thresholds δ_{ki} for each item, that is, $\delta_i = \bar{\delta}_{ki}$. It is convenient to use (3) in some illustrations and applications.

Elimination of Person Parameters

The formalization of invariance of comparisons rests on the existence of sufficient statistics, which implies that conditional on that statistic, the resultant distribution is independent of the relevant parameter. The sufficient statistic for the person parameter $\beta_n$is simply the total score $r = \sum_{i=1}^{I} x_{ni}$. Then for a pair of items i and j, and using directly the responses x_{ni}, x_{nj} the conditional equation

$$P\{x_{ni}, x_{ni} | r_n = x_{ni} + x_{ni}\}$$
$$= \frac{1}{\Psi_{nij}} \exp \sum_{k=1}^{x_{ni}} (-\delta_{ki}) \exp \sum_{k=1}^{x_{nj}} (-\delta_{kj}), \quad (4)$$

where $\Psi_{nij} = \sum_{(x_{ni},\ x_{nj}|r_n)} \exp \sum_{k=1}^{x_{ni}} (-\delta_{ki}) \exp \sum_{k=1}^{x_{nj}} (-\delta_{kj})$ is the summation over all possible pairs of responses given a total score of r_n. Equation (4) is clearly independent of the person parameters β_n, $n = 1, 2, \ldots N$. It can be used to estimate the item threshold parameters independently of the person parameters. In the implementation of the estimation, (4) may be generalized by considering all possible pairs of items or conditioning on the total score across more than two items. Different software implements the estimation in different ways. The person parameters are generally estimated following the estimation of the item parameters using the item parameter estimates as known. Because there are generally many less items than persons, the procedure for estimating the person parameters by conditioning out the item parameters is not feasible. **Direct maximum likelihood estimates** of the person parameters are biased and methods for reducing the bias have been devised.

Controversy and the Rasch Model

The construction of a model on the basis of *a priori* requirements rather than on the basis of characterizing data involves a different paradigm from the traditional in the data model relationship. In the traditional paradigm, if the model does not fit the data, then consideration is routinely given to finding a different model which accounts for the data better. In the Rasch paradigm, the emphasis is on whether the data fit the chosen model, and if not, then consideration is routinely given to understanding what aspect of the data is failing to fit the model.

The RM has the required structure of fundamental measurement or additive conjoint measurement in a probabilistic framework (*see* **Measurement: Overview**) [9, 19]. This is the main reason that those who adhere to the RM and try to construct data that fit it, give for adhering to it and for trying to collect data that fit it. In addition, they argue that the constructing measuring instruments, is not simply a matter of characterizing data, but a deliberate attempt to construct measures which satisfy important properties, and that the RM provides an operational criterion for obtaining fundamental measurement. Thus, the data from a measuring instrument are seen to be deliberately constructed to be empirical valid and at the same time satisfy the requirements of measurements [5]. Some controversy, which has been discussed in [7], has arisen from this distinctive use of the RM.

Short biographies of Rasch can be found in [2], [6], and [18].

The Latent Structure of the Model

The RM for dichotomous responses, which specializes from (3) to

$$P\{X_{ni} = x\} = \frac{\exp x(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)}; \quad x \in \{0, 1\}, \quad (5)$$

is also the basis of the RM for more than two ordered categories. In this case there is only the one threshold, the location of the item δ_i.

The Derivation of the Latent Structure Assuming a Guttman Pattern

To derive the latent structure of the model, assume an instantaneous latent dichotomous response process $\{Y_{nki} = y\}$, $y \in \{0, 1\}$ at each threshold [3]. Let this latent response take the form

$$P\{Y_{nki} = y\} = \frac{1}{\eta_{ni}} \exp y(\beta_n - \delta_{ki}), \quad (6)$$

where η_{ki} is the normalizing factor $\eta_{ni} = 1 + \exp(\beta_n - \delta_{ki})$.

Although instantaneously assumed to be independent, there is only *one* response in *one* of $m + 1$ categories. Therefore, the responses must be *latent* (*see* **Latent Variable**). Furthermore, if the responses were independent, there would be 2^m possible response patterns. Therefore, the responses must also be *dependent* and a *constraint* must be placed on any process in which the latent responses at the thresholds are instantaneously considered independent. The Guttman structure provides this constraint in exactly the required way. Table 2 shows the responses according to the Guttman structure for three items.

The rationale for the Guttman patterns in Table 2 [12] is that for unidimensional responses across items, if a person succeeds on an item, then the person should succeed on all items that are easier than that item and that if a person failed on an item, then the person should fail on all items more difficult than that item. A key characteristic of the Guttman pattern is that the total score across items recovers the response pattern perfectly. Of course, with experimentally independent items, that is, where each item is responded to independently of every other item, it is possible that the Guttman structure will not be observed in data and that given a particular total score the Guttman pattern will not be observed.

The rationale for the Guttman structure, as with the ordering of items in terms of their difficulty, is that the thresholds with an item are required to be ordered, that is

$$\tau_1 < \tau_2 < \tau_3 \cdots < \tau_{m-1} < \tau_m. \quad (7)$$

This requirement of ordered thresholds is independent of the RM – it is required by the Guttman structure and ordering of the categories. However, it is completely *compatible* with the structure of the responses in the RM and implies that a person at the threshold of two higher categories is at a higher location than a person at the boundary of two lower categories. For example, it requires that a person who is at the threshold between a Credit and Distinction in the first example in Table 1, and reflected in Figure 1, has a higher ability than a person who is at the threshold between a Fail and a Pass. That is, if the categories are to be ordered, then the thresholds

Table 2 The Guttman structure with dichotomous items in difficulty order

Items	1	2	3	Total score
	0	0	0	0
	1	0	0	1
	1	1	0	2
	1	1	1	3

that define them should also be ordered. The derivation of the model with this requirement is shown in detail in [3].

To briefly outline this derivation, consider the case of only four ordered categories as in Table 1 and therefore three thresholds. Then if the responses at the thresholds were independent, the probability of any set of responses across the thresholds is given by

$$P\{y_{n1i}, y_{n2i}, y_{n3i}\} = \prod_{k=1}^{3} \frac{\exp y_{nki}(\beta_n - \delta_{ki})}{1 + \exp(\beta_n - \delta_{ki})}, \tag{8}$$

where the sum of probabilities of *all* patterns $\sum_{All} \prod_{k=1}^{3} \frac{\exp y_{nki}(\beta_n - \delta_{ki})}{1 + \exp(\beta_n - \delta_{ki})} = 1.$

The subset of Guttman G patterns has a probability of occurring

$$\Gamma = \sum_{G} \prod_{k=1}^{3} \frac{\exp y_{ni}(\beta_n - \delta_{xi})}{1 + \exp(\beta_n - \delta_{xi})}$$
$$= \frac{\sum_{G} \exp \sum_{k=1}^{3} y_{nki}(\beta_n - \delta_{ki})}{\prod_{k=1}^{3} (1 + \exp(\beta_n - \delta_{ki}))} < 1. \tag{9}$$

Then the probability of a particular Guttman response pattern, conditional on the response being one of the Guttman patterns, is given by

$$P\{y_{n1i}, y_{n2i}, y_{n3i}|G\}$$
$$= \frac{\exp \sum_{k=1}^{3} y_{ni}(\beta_n - \delta_{ki})}{\prod_{i=1}^{3} (1 + \exp(\beta_n - \delta_{ki}))} \Bigg/ \Gamma$$
$$= \frac{\exp \sum_{k=1}^{3} y_{ni}(\beta_n - \delta_{ki})}{\prod_{k=1}^{3} (1 + \exp(\beta_n - \delta_{ki}))}$$

$$\Bigg/ \frac{\sum_{G} \exp \sum_{k=1}^{3} y_{nki}(\beta_n - \delta_{ki})}{\prod_{k=1}^{3} (1 + \exp(\beta_n - \delta_{ki}))}$$
$$= \frac{\exp \sum_{k=1}^{3} y_{ni}(\beta_n - \delta_{ki})}{\sum_{G} \exp \sum_{k=1}^{3} y_{nki}(\beta_n - \delta_{ki})}. \tag{10}$$

For example, suppose the response pattern is $\{1, 1, 0\}$, in which case the total score is 2. Then

$$P\{1, 1, 0|G\}$$
$$= \frac{\exp[1(\beta_n - \delta_{1i}) + 1(\beta_n - \delta_{2i}) + 0(\beta_n - \delta_{3i})]}{\sum_{G} \exp \sum_{i=1}^{3} y_{nki}(\beta_n - \delta_{ki})}$$
$$= \frac{\exp(2\beta_n - \delta_{1i} - \delta_{2i})}{\sum_{G} \exp \sum_{i=1}^{3} y_{nki}(\beta_n - \delta_{ki})}. \tag{11}$$

Notice that the coefficient of the person location $\beta_n$in the numerator of (11) is 2, the total score of the number of successes at the thresholds. This scoring can be generalized and the total score can be used to define the Guttman response pattern. Thus, define the integer random variable $X_{ni} = x \in \{0, 1, 2, 3\}$as the total score for each of the Guttman patterns: $0 \equiv (0, 0, 0), 1 \equiv (1, 0, 0), 2 \equiv (1, 1, 0), 3 \equiv (1, 1, 1)$. Then (11) simplifies to

$$P\{X_{ni} = x\} = \frac{\exp(x\beta_n - \delta_{1i} - \delta_{2i} \ldots - \delta_{xi})}{\sum_{x=0}^{3} \exp \sum_{k=1}^{x} k(\beta_n - \delta_{ki})}, \tag{12}$$

which is the special case of (3) in the case of just three thresholds and four categories.

Effectively, the successive categories are scored with successive integers as in elementary analyses of ordered categories. However, no assumption of equal distances between thresholds is made; the thresholds are estimated from the data and may be unequally spaced. The successive integer scoring rests on the discrimination of the latent dichotomous responses

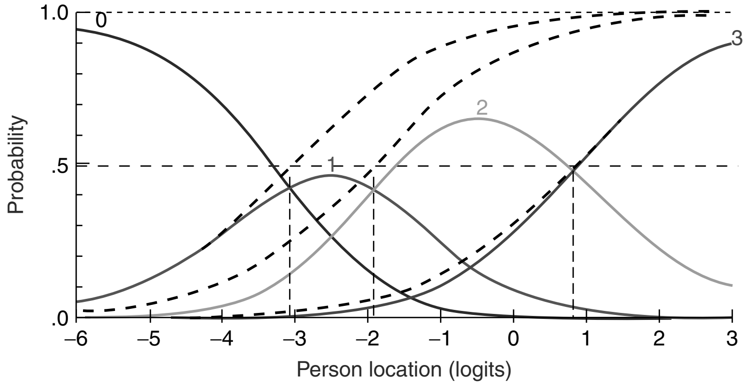


Figure 2 Probabilities of responses in each of four ordered categories showing the probabilities of the latent dichotomous responses at the thresholds

at the thresholds within an item being the same. Although the successive categories are scored with successive integers, it is essential to recognize that the response in any category implies a success at thresholds up to and including the lower threshold defining a category and failure on subsequent thresholds including the higher threshold defining a category. That is, the latent response structure of the model is a Guttman pattern of *successes* and *failures* at all the thresholds. Figure 2 shows Figure 1 augmented by the probabilities of the latent dichotomous responses at the thresholds according to (6).

The Derivation of the Latent Structure Resulting in the Guttman Pattern

This response structure is confirmed by considering the ratio of the probability of a response in any category, conditional on the response being in one of two adjacent categories. The probability of the response being in the higher of the two categories is readily shown to be

$$\frac{P\{X_{ni}=x\}}{P\{X_{ni}=x-1\}+P\{X_{ni}=x\}} = \frac{\exp(\beta_n-\delta_{xi})}{1+\exp(\beta_n-\delta_{xi})}, \tag{13}$$

which is just the dichotomous model at the threshold defined in (6).

This conditional latent response between two adjacent categories is dichotomous and again latent. It is an implication of the latent structure of the model because there is no sequence of observed conditional responses at the thresholds: there is just one response in one of the *m* categories.

In the above derivation, a Guttman pattern based on the ordering of the thresholds was imposed on an initially independent set of responses. Suppose now that the model is defined by the latent responses at the thresholds according to (13). Let

$$\frac{P\{X_{ni}=x\}}{P\{X_{ni}=x-1\}+P\{X_{ni}=x\}} = P_x, \tag{14}$$

and its complement

$$\frac{P\{X_{ni}=x-1\}}{P\{X_{ni}=x-1\}+P\{X_{ni}=x\}} = Q_x = 1-P_x. \tag{15}$$

Then it can be shown readily that

$$P\{X_{ni}=x\} = P_1P_2P_3\ldots.P_xQ_{x+1}Q_{x+2}\ldots Q_m/D, \tag{16}$$

where $D = Q_1Q_2Q_3\ldots Q_m + P_1Q_2Q_3\ldots Q_m + P_1 P_2Q_3\ldots.Q_m + \ldots P_1P_2P_3\ldots P_m$.

Clearly, the particular response $X_{ni}=x$ implies once again successes at the first $x$thresholds and failures at all the remaining thresholds. That is, the response structure results in successes at the thresholds consistent with the Guttman pattern. This in turn implies an ordering of the thresholds. Thus, both derivations lead to the same structure at the thresholds.

The Log Odds Form and Potential for Misinterpretation

Consider (12) again.

Taking the ratio of the response in two adjacent categories gives the odds of success at the threshold:

$$\frac{P\{X_{ni} = x\}}{P\{X_{ni} = x - 1\}} = \exp(\beta_n - \delta_{xi}). \qquad (17)$$

Taking the logarithm gives

$$\ln \frac{P\{X_{ni} = x\}}{P\{X_{ni} = x - 1\}} = \beta_n - \delta_{xi}. \qquad (18)$$

This log odds form of the model, while simple, eschews its richness and invites making up a response process that has nothing to do with the model. It does this because it can give the impression that there is an independent response at each threshold, an interpretation which incorrectly ignores that there is only one response among the categories and that the dichotomous responses at the thresholds are latent, implied, and never observed. This form loses, for example, the fact that the probability of a response in any category is a function of all thresholds. This can be seen from the normalizing constant denominator in (3), which contains all thresholds. Thus, the probability of a response in the first category is affected by the location of the last threshold.

Possibility of Reversed Thresholds in Data

Although the ordering of the thresholds is required in the data and the RM is compatible with such ordering, it is possible to have data in which the thresholds, when estimated, are not in the correct order. This can occur because there is only one response in one category, and there is no restriction on the distribution of those responses.

The relative distribution of responses for a single person across triplets of successive categories can be derived simply from (16) for pairs of successive categories:

$$\frac{P\{X_{ni} = x\}}{P\{X_{ni} = x - 1\}} = \exp(\beta_n - \delta_{xi}) \qquad (19)$$

and

$$\frac{P\{X_{ni} = x + 1\}}{P\{X_{ni} = x\}} = \exp(\beta_n - \delta_{x+1.i}). \qquad (20)$$

Therefore,

$$\frac{P\{X_{ni} = x\}}{P\{X_{ni} = x - 1\}} \frac{P\{X_{ni} = x\}}{P\{X_{ni} = x + 1\}} = \exp(\delta_{x+1,i} - \delta_{xi}). \qquad (21)$$

If $\delta_{x+1,i} > \delta_{xi}$, then $\delta_{x+1i} - \delta_{xi} > 0, \exp(\delta_{x+1,i} - \delta_{xi}) > 1$ then from (21),

$$[P\{X_{ni} = x\}]^2 > P\{X_{ni} = x - 1\}P\{X_{ni} = x + 1\}. \qquad (22)$$

Because each person responds only in one category of an item, there is no constraint in the responses to conform to (22); this is an empirical matter. The violation of order can occur even if the data fit the model statistically. It is a powerful property of the RM that the estimates of the thresholds can be obtained independently of the distribution of the person parameters, indicating that the relationship among the thresholds can be inferred as an intrinsic structure of the of the operational characteristics of the item.

Figure 3 shows the CCCs of an item in which the last two thresholds are reversed. It is evident that the threshold between Pass and Credit has a greater location than the threshold between Credit and Distinction. It means that if this is accepted, then the person who has 50% chance of being given a Credit or a Distinction has less ability than a person who has 50% chance of getting a Pass or a Credit. This clearly violates any a-priori principle of ordering of the categories. It means that there is a problem with the empirical ordering of the categories and that the successive categories do not reflect increasing order on the trait.

Other symptoms of the problem of reversed thresholds is that there is no region in Figure 3 in which the grade of Credit is most likely and that the region in which Credit should be assigned is undefined – it implies some kind of negative distance. Compatible with the paradigm of the RM, reversed threshold estimates direct a closer study of the possible reasons that the ordering of the categories are not working as intended.

Although lack of data in a sample in any category can result in estimates of parameters with large standard errors, the key factor in the estimates is the relationship amongst the categories of the implied probabilities of (22). These cannot be inferred directly from raw frequencies in categories in a sample. Thus, in the

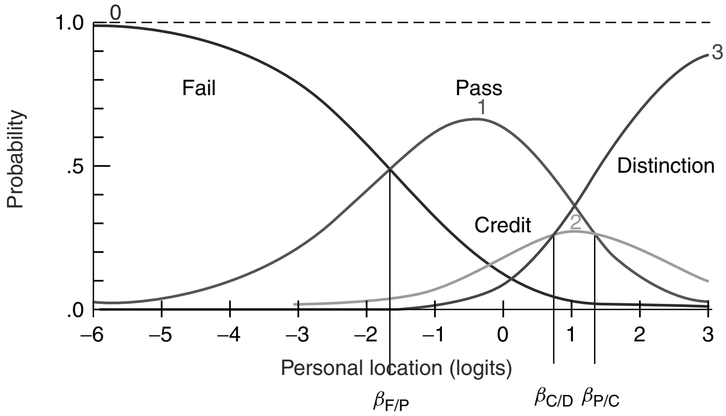


Figure 3 Category characteristic curves showing the probabilities of responses in each of four ordered categories when the thresholds are disordered

case of Figure 2, any *single* person whose ability estimate is between the thresholds identified by $\beta_{C/D}$ and $\beta_{P/C}$ will, simultaneously, have a higher probability of a getting a Distinction and a Pass than getting a Credit. This is not only incompatible with ordering of the categories, but it is *not a matter of the distribution of the persons in the sample of data analyzed.*

To consolidate this point, Table 3 shows the frequencies of responses of 1000 persons for two items each with 12 categories. These are simulated data, which fit the model to have correctly ordered thresholds. It shows that in the middle categories, the frequencies are very small compared to the extremes, and in particular, the score of 5 has a 0 frequency for item 1. Nevertheless, the threshold estimates shown in Table 4 have the required order. The method of estimation, which exploits the structure of responses among categories to span and adjust for the category with 0 frequency and conditions out the person parameters, is described in [8]. The reason that the frequencies in the middle categories are low or even 0 is that they arise from a bimodal distribution of person locations. It is analogous to having heights of a sample of adult males and females. This too would be bimodal and therefore heights somewhere in between the two modes would have, by definition, a low frequency. However, it would be untenable if the low frequencies in the middle heights would reverse the lines (thresholds) which define the units on the ruler. Figure 4 shows the frequency distribution of the estimated person parameters and confirms that it is bimodal. Clearly, given the distribution, there would be few cases in the middle categories with scores of 5 and 6.

Table 3 Frequencies of responses in two items with 12 categories

Item	0	1	2	3	4	5	6	7	8	9	10	11
I0001	81	175	123	53	15	0	8	11	51	120	165	86
I0002	96	155	119	57	17	5	2	26	48	115	161	87

Table 4 Estimates of thresholds for two items with low frequencies in the middle categories

		Threshold estimates										
Item	$\hat{\delta}_i$	1	2	3	4	5	6	7	8	9	10	11
1	0.002	−3.96	−2.89	−2.01	−1.27	−0.62	−0.02	0.59	1.25	2.00	2.91	4.01
2	−0.002-	−3.78	−2.92	−2.15	−1.42	−0.73	−0.05	0.64	1.36	2.14	2.99	3.94

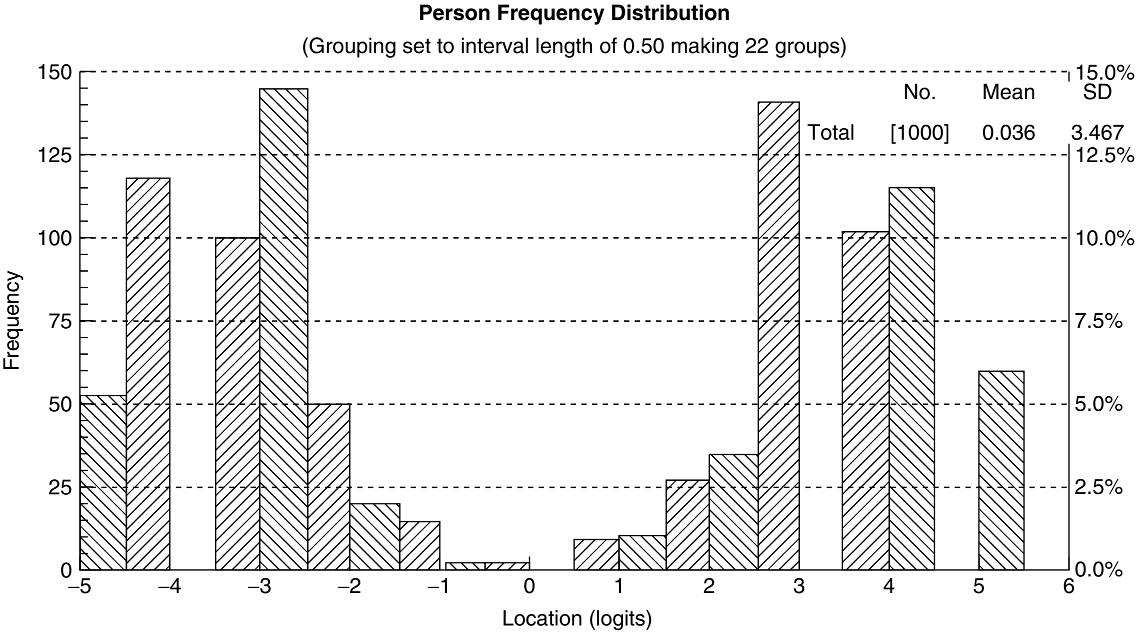


Figure 4 A bimodal distribution of locations estimates

The Collapsing of Adjacent Categories

A distinctive feature of the RM is that summing the probabilities of adjacent categories, produces a model which is no longer a RM. In particular, dichotomizing in that way is not consistent with the model. That is, taking (3), and forming

$$\begin{aligned} P\{X_{ni} = x\} + P\{X_{ni} = x + 1\} \\ = \frac{1}{\gamma_{ni}} \exp\left(-\sum_{k=1}^{x} \tau_{ki} + x(\beta_n - \delta_i)\right) \\ + \frac{1}{\gamma_{ni}} \exp\left(-\sum_{k=1}^{x+1} \tau_{ki} + x(\beta_n - \delta_i)\right) \quad (23) \end{aligned}$$

gives (23) which cannot be reduced to the form of (3). This result has been discussed in [4], [13] and was noted by Rasch [15]. It implies that collapsing categories is not arbitrary but an integral property of the data revealed through the model. This nonarbitrariness in combining categories contributes to the model providing information on the empirical ordering of the categories.

The Process Compatible with the Rasch Model

From the above outline of the structure of the RM, it is possible to summarize the response *process* that is compatible with it. The response process is one of *simultaneous ordered classification*. That is, the process is one of considering the property of an object, which might be a property of oneself or of some performance, relative to an item with two or more than two ordered categories, and deciding the category of the response.

The examples in Table 1 show that each successive category implies the previous category in the order *and in addition*, reflects more of the assessed trait. This is compatible with the Guttman structure. Thus, a response in a category implies that the latent response was a success at the lower of the two thresholds, and a failure at the greater of the two thresholds. *And this response determines the implied latent responses at all of the other thresholds.* This makes the response process a *simultaneous classification* process across the thresholds. The further implication is that when the manifest responses are used to estimate the thresholds, the threshold locations are themselves empirically defined *simultaneously* - that is, the estimates arise from data in which all the thresholds of an item were involved simultaneously in every response. This contributes further to the distinctive feature of the RM that it can be used to assess whether or not the categories are working in the intended ordering, or whether on this feature, the empirical ordering breaks down.

References

[1] Andersen, E.B. (1977). Sufficient statistics and latent trait models, *Psychometrika* **42**, 69–81.

[2] Andersen, E.B. & Olsen, L.W. (2000). The life of Georg Rasch as a mathematician and as a statistician, in *Essays in Item Response Theory*, A. Boomsma, M.A.J. van Duijn & T.A.B. Snijders, eds, Springer, New York, pp. 3–24.

[3] Andrich, D. (1978). A rating formulation for ordered response categories, *Psychometrika* **43**, 357–374.

[4] Andrich, D. (1995). Models for measurement, precision and the non-dichotomization of graded responses, *Psychometrika* **60**, 7–26.

[5] Andrich, D. (1996). Measurement criteria for choosing among models for graded responses, in *Analysis of Categorical Variables in Developmental Research*, Chap. 1, A. von Eye & C.C. Clogg, eds, Academic Press, Orlando, pp. 3–35.

[6] Andrich, D. (2004a). *Georg Rasch: Mathematician and Statistician. Encyclopaedia of Social Measurement*, Academic Press.

[7] Andrich, D. (2004b). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care* **42**, 7–16.

[8] Andrich, D. & Luo, G. (2003). Conditional estimation in the Rasch model for ordered response categories using principal components, *Journal of Applied Measurement* **4**, 205–221.

[9] Brogden, H.E. (1977). The Rasch model, the law of comparative judgement and additive conjoint measurement, *Psychometrika* **42**, 631–634.

[10] Engelhard Jr, G. & Wilson, M., eds (1996). *Objective Measurement: Theory into Practice*, Vol. 3, Norwood, Ablex Publishing.

[11] Fischer, G.H. & Molenaar, I.W., eds, (1995). *Rasch Models: Foundations, Recent Developments, and Applications*, Springer, New York.

[12] Guttman, L. (1950). The basis for scalogram analysis, in *Measurement and Prediction*, S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star & J.A. Clausen, eds, Wiley, New York, pp. 60–90.

[13] Jansen, P.G.W. & Roskam, E.E. (1986). Latent trait models and dichotomization of graded responses, *Psychometrika* **51**(1), 69–91.

[14] Rasch, G. (1961). On general laws and the meaning of measurement in psychology, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. IV, J. Neyman, ed., University of California Press, Berkeley, pp. 321–334.

[15] Rasch, G. (1966). An individualistic approach to item analysis, in *Readings in Mathematical Social Science*, P.F. Lazarsfeld & N.W. Henry, eds, Science Research Associates, Chicago, pp. 89–108.

[16] Smith Jr, E.V. & Smith, R.M., eds, (2004). *Introduction to Rasch Measurement*, JAM Press, Maple Grove.

[17] Van der Linden, W. & Hambleton, R., eds (1997). *Handbook of Modern Item Response Theory*, Springer, New York, pp. 139–152.

[18] Wright, B.D. (1980). Foreword to *Probabilistic Models for Some Intelligence and Attainment Tests*, G., Rasch, ed., 1960 Reprinted University of Chicago Press.

[19] Wright, B.D. (1997). A history of social science measurement, *Educational Measurement: Issues and Practice* **16**(4), 33–45.

[20] Wright, B.D. & Masters, G.N. (1982). *Rating Scale Analysis: Rasch Measurement*, MESA Press, Chicago.

(*See also* **Rasch Modeling**)

DAVID ANDRICH

Rater Agreement

Christof Schuster and David Smith

Volume 4, pp. 1707–1712

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Rater Agreement

Introduction

Human judgment is prone to error that researchers routinely seek to quantify, understand, and minimize. To quantify the quality of nominal judgments, agreement among multiple raters of the same target is examined using either global summary indices, such as Cohen's kappa (*see* **Rater Agreement – Kappa**), or by modeling important properties of the judgment process using **latent class analysis**. In this entry, we give an overview of the latent class analysis approach to nominal scale rater agreement data.

Latent class models of rater agreement assume there is an underlying or 'true' latent category to which each target belongs, and that clues to the nature of this *latent class*, and to the nature of the judgment process itself, can be found in the observed classifications of multiple raters. With targets belonging to only one of the latent classes, manifest disagreement among observed judgments requires that at least one judgment be erroneous. However, without knowledge of the true latent class, or a 'gold standard' indicator of the true latent class, correct and erroneous judgments cannot be distinguished. Even though individual target classifications cannot be directly established at the latent level, given certain assumptions are fulfilled, latent class analysis can nevertheless estimate overall misclassification probabilities.

Data Example

To illustrate the rater agreement models we discuss in this article, suppose that 212 patients were diagnosed according to four raters as either 'schizophrenic' or 'not schizophrenic.' For two categories and four raters, there are 16 possible rating profiles, which are given in the first four columns of Table 1. The frequency with which each profile occurred is given in the fifth column.[1] As the frequencies in this table show, all four judges agree on 111 of the 212 targets. In other words, in approximately 48% of the targets there is at least some disagreement among the judges.

The Latent Class Model

The event $A = i$, $i = 1, \ldots, I$, indicates a target assigned to the ith category by rater A. Similarly, $X = t$, $t = 1, \ldots, T$, indicates a target truly belongs to the tth latent class. The probabilities of these events are denoted as $P(A = i) = \pi_A(i)$ and $P(X = t) = \pi_X(t)$, respectively. The conditional probability (*see* **Probability: An Introduction**) that a rater will assign a target of the tth latent class to the ith category is denoted as $P(A = i|X = t) = \pi_{A|X}(i|t)$. To avoid unnecessarily complex notation, the illustrative latent class model example we present in this chapter (viz. Table 1) assumes four raters, denoted as A_1, A_2, A_3, and A_4.

Table 1 Rating profile frequencies, *n*, of psychiatric diagnoses according to four raters, A_1, A_2, A_3, A_4, where '1' = schizophrenic and '2' = not schizophrenic

A_1	A_2	A_3	A_4	n
1	1	1	1	12
1	1	1	2	10
1	1	2	1	4
1	1	2	2	8
1	2	1	1	10
1	2	1	2	6
1	2	2	1	7
1	2	2	2	8
2	1	1	1	0
2	1	1	2	1
2	1	2	1	1
2	1	2	2	8
2	2	1	1	8
2	2	1	2	15
2	2	2	1	15
2	2	2	2	99

Latent class analysis is based on the assumption of local (conditional) independence (*see* **Conditional Independence**). According to this assumption, multiple judgments of the same target are independent. Therefore, the raters and latent status joint probability factors as

$$\pi_{A_1A_2A_3A_4X}(i, j, k, l, t) = \pi_{A_1|X}(i|t)\pi_{A_2|X}(j|t)\pi_{A_3|X}(k|t)\pi_{A_4|X}(l|t)\pi_X(t). \tag{1}$$

This raters and latent status joint probability can be related to the rating profile probability by

$$\pi_{A_1A_2A_3A_4}(i, j, k, l) = \sum_{t=1}^{T} \pi_{A_1A_2A_3A_4X}(i, j, k, l, t). \tag{2}$$

This formula shows that the probabilities of the rating profiles, the left-side of (2), are obtained by collapsing the joint probability $\pi_{A_1A_2A_3A_4X}(i,\ j,\ k,\ l,\ t)$ over the levels of the latent class variable. Combining (1) and (2) yields

$$\pi_{A_1A_2A_3A_4}(i, j, k, l) = \sum_{t=1}^{T} \pi_{A_1|X}(i|t)\pi_{A_2|X}(j|t)\pi_{A_3|X}(k|t)\pi_{A_4|X}(l|t)\pi_X(t) \tag{3}$$

This equation relates the probabilities of the observed rating profiles to conditional response probabilities and the latent class probabilities. For this model to be identified, it is necessary, though not sufficient, for the degrees of freedom, calculated as $\mathrm{df} = I^R - (RI - R + 1)T$, where R denotes the number of raters, to be nonnegative.

Although the present review of latent class models only involves one categorical latent variable, the model is applicable also to more than one categorical latent variable. This is because the cells in a cross-classification can be represented equivalently as categories of a single nominal variable. Thus, if models are presented in terms of multiple latent variables – as is often the case when modeling rater agreement data – this is done only for conceptual clarity, not out of statistical necessity. We return to this issue in the next section.

Furthermore, it is often desirable to restrict the probabilities on the right-hand side of (3), either for substantive reasons or to ensure model identification. Probabilities can be fixed to known values, set equal to one another, or any combination of the two. The wide variety of rater agreement latent class models found in the literature emerges as a product of the wide variety of restrictions that can be imposed on the model probabilities.

Distinguishing Target Types

In the rater agreement literature, it has become increasingly popular to introduce a second latent class variable that reflects the common recognition that targets differ not only with respect to their status on the substantive latent variable of interest but also with respect to the target's prototypicality with respect to the latent class (e.g., [6]). These 'target type' models consider the ease with which targets can be classified, although 'ease' in this context is only metaphorical and should not necessarily be equated with any perceived experience on the part of the raters.

We will illustrate the operation of the target type latent class variable by considering three target types; obvious, suggestive, and ambiguous. Where Y is the target type latent variable, we arbitrarily define $Y = 1$ for obvious targets, $Y = 2$ for suggestive targets, and $Y = 3$ for ambiguous targets. The latent variables X and Y are assumed to be fully crossed. In other words, all possible combinations of levels of X and Y have a positive probability of occurrence. However, we do not assume populations that necessarily involve all three target types.

Obvious Targets. Targets belonging to the obvious latent class can be readily identified. These targets are often referred to as prototypes or 'textbook cases'. To formalize this idea, one restricts the conditional probabilities $\pi_{A|XY}(i|t, 1) = \delta_{it}$, where $\delta_{it} = 1$ if $i = t$ and $\delta_{it} = 0$ else. Because this restriction is motivated by characteristics of the targets, it applies to all raters in the same manner. In connection with the local independence assumption, and continuing our four judge example, one obtains

$$\pi_{A_1A_2A_3A_4|Y}(i, j, k, l|1) = \delta_{ijkl}\pi_{X|Y}(i|1), \tag{4}$$

where $\delta_{ijkl} = 1$ if $i = j = k = l$ and $\delta_{ijkl} = 0$ else.

Suggestive Targets. While targets belonging to the suggestive latent class are not obvious, their rate of correct assignment is better than chance. Suggestive targets possess features tending toward, but not decisively establishing, a particular latent status. In this sense, the suggestive targets are the ones for which the latent class model has been developed. Formally speaking, the suggestive class model is

$$\pi_{A_1A_2A_3A_4|Y}(i, j, k, l|2) = \sum_{t=1}^{T} \pi_{A_1|XY}(i|t, 2)\pi_{A_2|XY}(j|t, 2)\pi_{A_3|XY}(k|t, 2) \times \pi_{A_4|XY}(l|t, 2)\pi_{X|Y}(t|2). \tag{5}$$

Note that (5) is equivalent to (3).

Ambiguous Targets. Judgments of ambiguous targets are no better than random. In conditional probability terms, this means that ambiguous target judgments do not depend on the target's true latent class.

In other words, $\pi_{A|XY}(i|t,3) = \pi_{A|Y}(i|3)$. Together with the local independence assumption, this yields

$$\pi_{A_1A_2A_3A_4|Y}(i,j,k,l|3) = \pi_{A_1|Y}(i|3)\pi_{A_2|Y}(j|3)\pi_{A_3|Y}(k|3)\pi_{A_4|Y}(l|3). \tag{6}$$

Models Having a Single Target Type

Having delineated three different target types, it is possible to consider models that focus on a single target type only. Of the three possible model classes, – models for suggestive, obvious, or ambiguous targets – only models for suggestive targets are typically of interest to substantive researchers. Nevertheless, greater appreciation of the latent class approach can be gained by briefly considering the three situations in which only a single target type is present.

Suggestive Targets Only. Because latent class models for rater agreement are based on the premise that for each observed category there exists a corresponding latent class, it is natural to consider the general latent class model that has as many latent classes as there are observed categories. In other words, one can consider the general latent class model for which $T = I$. This model has been considered by, for example, Dillon and Mulani [2].[2]

Fitting this model to the data given in Table 1 produces an acceptable model fit. Specifically, one obtains $X^2 = 9.501$ and $G^2 = 10.021$ based on df $=$ 7, allowing for one boundary value. Boundary values are probabilities that are estimated to be either zero or one. For the hypothetical population from which this sample has been taken the prevalence of schizophrenia is estimated to be $\pi_X(1) = 0.2788$. Individual rater sensitivities, $\pi_{A_r|X}(1|1)$, for $r = 1,2,3,4$ are estimated as 1.0000, .5262, .5956, and .5381. Similarly, individual rater specificities, $\pi_{A_r|X}(2|2)$, are estimated as .9762, .9346, .8412, and .8401.

Obvious Targets Only. If only obvious targets are judged, there would be perfect agreement. Every rater would be able to identify the latent class to which each target correctly belongs. Of course, in this situation, statistical models would be unnecessary.

Ambiguous Targets Only. If a population consists only of ambiguous targets, the quality of the judgments would be no better than chance. Consequently, the judgments would be meaningless from a substantive perspective. Nevertheless, it is possible to distinguish at least two different modes either of which could underlie random judgments. First, raters could produce random judgments in accordance with category specific base-rates, that is, base-rates that could differ across raters. Alternatively, judgments could be random in the sense of their being evenly distributed among the categories.

Models Having Two Types of Targets

Allowing for two different types of targets produces three different model classes of the form

$$\pi_{A_1A_2A_3A_4}(i,j,k,l) = \pi_{A_1A_2A_3A_4|Y}(i,j,k,l|u)\pi_Y(u) + \pi_{A_1A_2A_3A_4|Y}(i,j,k,l|v)\pi_Y(v), \tag{7}$$

where $u, v = 1, 2, 3$. For models that include only obvious and ambiguous targets, $u = 1$ and $v = 3$; for obvious and suggestive targets only, $u = 1$ and $v = 2$; and for suggestive and ambiguous targets only, $u = 2$ and $v = 3$. Of course, in each of the three cases the two conditional probabilities on the right-hand side of (7) will be replaced with the corresponding expressions given in (4), (5), and (6). Because the resulting model equations are evident, they are not presented.

Obvious and Ambiguous Targets. Treating all targets as either obvious or ambiguous has been suggested by Clogg [1] and by Schuster and Smith [8]. Fitting this model to the data example yields a poor fit. Specifically, one obtains $X^2 = 22.6833$ and $G^2 = 28.1816$ based on df $= 9$. The proportion of obvious targets, $\pi_Y(1)$, is estimated as 44.80%. Among these, only 8.43% are estimated to be schizophrenic, that is, $\pi_{X|Y}(1|1) = .0843$. The rater specific probabilities of a schizophrenia diagnosis for the ambiguous targets, $\pi_{A_r|Y}(1|3)$, $r = 1,2,3,4$, are .4676, .2828, .4399, and .4122 respectively. Clearly, while raters one, three, and four seem to behave similarly when judging ambiguous targets, the second rater's behavior is different from the other three. Note that this model does not produce a prevalence estimate of schizophrenia for ambiguous targets.

Obvious and Suggestive Targets. Treating all targets as either obvious or suggestive has been considered by Espeland and Handelman [3]. Fitting this model to the data produces an acceptable model fit. Specifically, one obtains $X^2 = 6.5557$ and $G^2 = 7.1543$ based on df $= 5$, allowing for one boundary value. The proportion of obvious targets is estimated as 23.43%. Among these, only 5.22% are estimated to be schizophrenic, that is, $\pi_{X|Y}(1|1) = .0522$. Similarly, among the 76.57% suggestive targets one estimates 35.32% to be schizophrenic, that is, $\pi_{X|Y}(1|1) = .3532$. The individual rater sensitivities for the suggestive targets, $\pi_{A_r|XY}(1|1, 2)$, $r = 1, 2, 3, 4$, are 1.0000, .5383, .6001, .5070, and the corresponding specificities, $\pi_{A_r|XY}(2|2, 2)$, are .9515, .8994, .7617, and .7585. Overall, the values for the sensitivities and specificities are very similar to the model that involves only suggestive targets.

Suggestive and Ambiguous Targets. Treating all targets as either suggestive or ambiguous has been considered by Espeland and Handelman [3] and by Schuster and Smith [8]. Fitting this model to the data yields an acceptable model fit of $X^2 = 1.8695$ and $G^2 = 1.7589$ based on df $= 4$, allowing for three boundary values. The proportion of suggestive targets, $\pi_Y(2)$, is estimated to be 85.81% of which 28.30% are schizophrenic, that is, $\pi_{X|Y}(1|2) = .2830$. The individual rater sensitivities for the suggestive targets, $\pi_{A_r|XY}(1|1, 2)$, $r = 1, 2, 3, 4$, are 1.0000, .5999, .6373, .5082 for raters A_1 to A_4, and the corresponding specificities, $\pi_{A_r|XY}(2|2, 2)$, are .9619, .9217, .8711, and 1.0000. The rater specific probabilities of a schizophrenia diagnosis for the ambiguous targets are .2090, .0000, .3281, and .9999 for raters A_1 to A_4. It is remarkable how much the likelihood of a schizophrenia diagnosis for ambiguous targets differs for raters two and four. If the relatively large number of boundary values is not indicative of an inappropriate model, then it is as if rater 2 will not diagnose schizophrenia unless there is enough specific evidence of it, while rater 4 views ambiguity as diagnostic of schizophrenia. Note that this model cannot produce a prevalence estimate of schizophrenia for ambiguous targets.

Three Types of Targets

When the population of targets contains all three types, one obtains

$$\pi_{A_1A_2A_3A_4}(i, j, k, l) = \pi_{A_1A_2A_3A_4|Y}(i, j, k, l|1)\pi_Y(1) + \pi_{A_1A_2A_3A_4|Y}(i, j, k, l|2)\pi_Y(2) + \pi_{A_1A_2A_3A_4|Y}(i, j, k, l|3)\pi_Y(3), \tag{8}$$

where, as before, each of the three conditional probabilities on the right-hand side are replaced using (4–6). Although this model follows naturally from consideration of different targets, we are not aware of an application of this model in the literature. For the present data example the number of parameters of this model exceeds the number of rating profiles, and, therefore, the model is not identified.

Espeland and Handelman [3] have discussed this model assuming equally probable categories. In this case, the model can be fitted to the data in Table 1. However, for the present data set, the probability of ambiguous targets is estimated to be zero. Therefore, the model fit is essentially equivalent to the model involving only obvious and suggestive targets.

Discussion

Table 2 summarizes the goodness-of-fit statistics for the models fitted to the data in Table 1. When comparing the goodness-of-fit statistics of different models one has to keep in mind that differences between likelihood-ratio statistics follow a central chi-square distribution only if the models are nested. Thus, the only legitimate comparisons are between Model 1 and Model 3 ($\Delta G^2 = 2.8663$, $\Delta\text{df} = 3$, $p = .413$) and between Model 1 and Model 4 ($\Delta G^2 = 8.2617$, $\Delta\text{df} = 3$, $p = .041$). Clearly, insofar as these data are concerned allowing for obvious targets in addition to suggestive targets does not improve model fit considerably. However, allowing for ambiguous targets in addition to suggestive targets improves the model fit considerably.

Depending on the target type, additional restrictions can be imposed on the latent class model. In particular, for suggestive targets one can constrain the hit-rates, $\pi_{A|X}(i|i)$, to be equal across raters,

Table 2 Goodness-of-fit statistics for models fitted to the data in Table 1. The models are specified in terms of the target types involved

Model	Target-types	df	X^2	G^2
1	$Y = 2$	7	9.5010	10.0206
2	$Y = 1, Y = 3$	9	22.6833	28.1816
3	$Y = 1, Y = 2$	5	6.5557	7.1543
4	$Y = 2, Y = 3$	4	1.8695	1.7589

across categories, or across raters and categories simultaneously, see [2] or [8] for details. Alternatively, one can constrain error-rates across raters or targets.

For ambiguous targets the response probabilities can be constrained to be equal across raters, that is, $\pi_{A_r|Y}(i|3) = \pi_{A_q|Y}(i|3)$ for $r \neq q$. It is also possible to define random assignment more restrictively, such as requiring each category be equally probable. In this case, each of the four probabilities on the right-hand side of (6) would be replaced with $1/I$, where Is denotes the number of categories.

Of course, the restrictions for suggestive and ambiguous targets can be combined. In particular, if the rater panel has been randomly selected, one should employ restrictions that imply homogeneous rater margins, that is, $\pi_{A_r}(i) = \pi_{A_q}(i)$ for $r \neq q$. For random rater panels one could also consider symmetry models, that is, models that imply $\pi_{A_1A_2A_3A_4}(i, j, k, l)$ is constant for all permutations of index vector (i, j, k, l).

Finally, an alternative way in which a second type of latent variable can be introduced is to assume two different rater modes or states in which the raters operate [5], which is an idea closely related to the models proposed by Schutz [9] and Perrault and Leigh [7]. Raters in reliable mode judge targets correctly with a probability of 1.0, while raters in unreliable mode will 'guess' the category. This model is different from the target type models inasmuch as the representation of the rater mode will require a separate latent variable for each rater. In addition, the rater modes are assumed independent. Therefore, the model involves restrictions among the latent variables, which is not the case for traditional latent class models.

Notes

1. A similar data set has been presented by Young, Tanner, and Meltzer [10].
2. However, note that for a given number of raters, the number of latent classes that are identifiable may depend on the number of observed categories. For dichotomous ratings, model identification requires at least three raters. In cases of either three or four categories, four raters are required to ensure model identification (see [4]).

References

[1] Clogg, C.C. (1979). Some latent structure models for the analysis of likert-type data, *Social Science Research* **8**, 287–301.

[2] Dillon, W.R. & Mulani, N. (1984). A probabilistic latent class model for assessing inter-judge reliability, *Multivariate Behavioral Research* **19**, 438–458.

[3] Espeland, M.A. & Handelman, S.L. (1989). Using latent class models to characterize and assess relative error in discrete measurements, *Biometrics* **45**, 587–599.

[4] Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika* **61**(2), 215–231.

[5] Klauer, K.C. & Batchelder, W.H. (1996). Structural analysis of subjective categorical data, *Psychometrika* **61**(2), 199–240.

[6] Maxwell, A.E. (1977). Coefficients of agreement between observers and their interpretation, *British Journal of Psychiatry* **130**, 79–83.

[7] Perrault, W.D. & Leigh, L.E. (1989). Reliability of nominal data based on qualitative judgments, *Journal of Marketing Research* **26**, 135–148.

[8] Schuster, C. & Smith, D.A. (2002). Indexing systematic rater agreement with a latent class model, *Psychological Methods* **7**(3), 384–395.

[9] Schutz, W.C. (1952). Reliability, ambiguity and content analysis, *Psychological Review* **59**, 119–129.

[10] Young, M.A., Tanner, M.A. & Meltzer, H.Y. (1982). Operational definitions of schizophrenia: what do they identify? *Journal of Nervous and Mental Disease* **170**(8), 443–447.

Christof Schuster and David Smith

Rater Agreement – Kappa

Eun Young Mun

Volume 4, pp. 1712–1714

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Rater Agreement – Kappa

Whether there is agreement or consensus among raters or observers on their evaluation of the objects of interest is one of the key questions in the behavioral sciences. Social workers assess whether parents provide appropriate rearing environments for their children. School psychologists evaluate whether a child needs to be placed in a special class. If agreement among raters is low or absent, it begs many questions, including the validity of the guidelines or criteria, and the reliability of the raters involved in their judgment. Thus, it is important to establish rater agreement in the behavioral sciences. One way to measure rater agreement is to have two raters make independent observations on the same group of objects, to classify them into a limited number of categories, and then to see to what extent the raters' evaluations overlap. There are several measures to assess rater agreement for this type of situation. Of all measures, Cohen's kappa (κ, [1]) is one of the best known and most frequently used measures to assess the extent of agreement by raters in a summary statement for entire observations [3].

Cohen's kappa measures the strength of rater agreement against the expectation of independence of ratings. Independence of ratings refers to a situation where the judgment made by one rater or observer is independent of, or unaffected by, the judgment made by the other rater. Table 1 illustrates a typical cross-classification table of two raters (*see* **Contingency Tables**). Three classification categories by both raters are completely crossed, resulting in a square table of nine cells. Of the nine cells, the three diagonal cells from the left to the right, shaded, are called the *agreement cells*. The remaining six other cells are referred to as the *disagreement cells*. The numbers in the cells are normally frequencies, and are indexed by m_{ij}. Subscripts i and j index I rows and J columns. m_{11}, for example, indicates the number of observations for which both raters used Category 1. The last row and column in Table 1 represent the row and column totals. m_{1i} indicates the row total for Category 1 by Rater A, whereas m_{i1} indicates the column total for Category 1 by Rater B. N is for the total number of objects that are evaluated by the two raters.

Cohen's kappa is calculated by

$$\hat{\kappa} = \frac{\sum p_{ii} - \sum p_{i.}p_{.i}}{1 - \sum p_{i.}p_{.i}},$$

where $\sum p_{ii}$ and $\sum p_{i.}p_{.i}$ indicate the *observed* and the *expected* sample proportions of agreement based on independence of ratings, respectively. Based on Table 1, the observed proportion of agreement can be calculated by adding frequencies of all agreement cells and dividing the total frequency of agreement by the number of total observations, $\sum p_{ii} = \sum m_{ii}/N = (m_{11} + m_{22} + m_{33})/N$. The expected sample proportion of agreement can be calculated by summing the products of the row and the column sums for each category, and dividing the sum by the squared total number of observations, $\sum p_{i.}p_{.i} = \sum m_{i.}m_{.i}/N^2 = (m_{1i}m_{i1} + m_{2i}m_{i2} + m_{3i}m_{i3})/N^2$.

The numerator of Cohen's kappa formula suggests that when the observed proportion of cases in which the two independent raters agree is greater than the expected proportion, then kappa is positive. When the observed proportion of agreement is less than the expected proportion, Cohen's kappa is negative. In addition, the difference in magnitude between the observed and the expected proportions is factored into Cohen's kappa calculation. The greater the

Table 1 A typical cross-classification of rater agreement

		Rater B			
		Category 1	Category 2	Category 3	Row sum
Rater A	Category 1	m_{11}	m_{12}	m_{13}	m_{1i}
	Category 2	m_{21}	m_{22}	m_{23}	m_{2i}
	Category 3	m_{31}	m_{32}	m_{33}	m_{3i}
	Column Sum	m_{i1}	m_{i2}	m_{i3}	N

Table 2 Cross-classification of rater agreement when Cohen's kappa = 0.674

		Psychologist B			
		Securely attached	Resistant or ambivalent	Avoidant	Row sum
Psychologist A	Securely Attached	64	4	2	70
	Resistant or Ambivalent	5	4	1	10
	Avoidant	1	2	17	20
	Column Sum	70	10	20	100

difference between the two proportions, the higher the absolute value of Cohen's kappa. The difference is scaled by the proportion possible for improvement. Namely, the denominator of the formula indicates the difference between the perfect agreement (i.e., Cohen's kappa $= 1$) and the expected proportion of agreement. In sum, Cohen's kappa is a measure of the proportion of agreement relative to that expected based on independence. κ can also be characterized as a measure of *proportionate reduction in error* (PRE; [2]). A κ value multiplied by 100 indicates the percentage by which two raters' agreement exceeds the expected agreement from chance.

Perfect agreement is indicated by Cohen's kappa $= 1$. Cohen's kappa $= 0$ if ratings are completely independent of each another. The higher Cohen's kappa, the better the agreement. Although it is rare, it is possible to have a negative Cohen's kappa estimate. A negative kappa estimate indicates that observed agreement is worse than expectation based on chance. Cohen's kappa can readily be calculated using a general purpose statistical software package and a significance test is routinely carried out to see whether κ is different from zero. In many instances, however, researchers are more interested in the magnitude of Cohen's kappa than in its significance. While there are more than one suggested guidelines for acceptable or good agreement, in general, a $\hat{\kappa}$ value less than 0.4 is considered as fair or slight agreement. A $\hat{\kappa}$ value greater than 0.4 is considered as good agreement (see [3] for more information).

Cohen's kappa can be extended to assess agreement by more than two raters. It can also be used to assess ratings or measurements carried out for multiple occasions across time. When more than two raters are used in assessing rater agreement, the expected proportion of agreement can be calculated by utilizing a standard main effect loglinear analysis.

Data Example

Suppose two psychologists observed infants' reactions to a separation and reunion situation with their mothers. Based on independent observations, psychologists A and B determined the attachment styles of 100 infants. Typically 70%, 10%, and 20% of infants are classified as *Securely Attached, Resistant or Ambivalent*, and *Avoidant*, respectively. Table 2 shows an artificial data example. The proportion of *observed* agreement is calculated at 0.85 (i.e., $(64 + 4 + 17)/100 = 0.85$), and the *expected* agreement is calculated at 0.54 (i.e., $(70 * 70 + 10 * 10 + 20 * 20)/10\,000 = 0.54$). The difference of 0.31 in proportion between the observed and expected agreement is compared against the maximum proportion that can be explained by rater agreement. $\hat{\kappa} = (0.85 - 0.54)/(1 - 0.54) = 0.674$, with standard error $= 0.073$, $z = 8.678$, $p < 0.01$. Thus, we conclude that two psychologists agree to 67.4% more than expected by chance, and that the agreement between two psychologists is significantly better than chance.

References

[1] Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**, 37–46.

[2] Fleiss, J.L. (1975). Measuring agreement between two judges in the presence or absence of a trait, *Biometrics* **31**, 651–659.

[3] von Eye, A. & Mun, E.Y. (2005). *Analyzing Rater Agreement: Manifest Variable Methods*, Lawrence Erlbaum, Mahwah.

Eun Young Mun

Rater Agreement – Weighted Kappa

Eun Young Mun

Volume 4, pp. 1714–1715

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Rater Agreement – Weighted Kappa

One of the characteristics of Cohen's kappa (κ, [1]) (*see* **Rater Agreement – Kappa**) is that any discrepancy between raters is equally weighted as zero. On the other hand, any agreement means *absolute agreement* between raters, and is equally weighted as one. Thus, the distinction between agreement and disagreement is categorical. For example, Table 1 from the article on Cohen's kappa (*see* **Rater Agreement – Kappa**) shows three agreement cells and six disagreement cells. In deriving Cohen's kappa, all six disagreement cells are treated equally. However, there could be situations in which the discrepancy between adjacent categories such as Categories 1 and 2 can be considered as partial agreement as opposed to complete disagreement. In those situations, Cohen's weighted kappa [2] can be used to reflect the magnitude of agreement. Thus, Cohen's weighted kappa is used when categories are measured by ordinal (or interval) scales. What we mean by ordinal is that categories reflect orderly magnitude. For example, categories of *Good, Average*, and *Bad* reflect a certain order in desirability. The disagreement by raters between *Good* and *Average*, or *Average* and *Bad* can be considered as less of a problem than the discrepancy between *Good* and *Bad*, and furthermore, the disagreement by one category or scale can sometimes be considered as partial agreement.

Cohen's weighted kappa is calculated by

$$\hat{\kappa}_{\text{w}} = \frac{\sum \omega_{ij} p_{ij} - \sum \omega_{ij} p_{i.} p_{.j}}{1 - \sum \omega_{ij} p_{i.} p_{.j}}, \qquad (1)$$

where the ω_{ij} are the weights, and $\sum \omega_{ij} p_{ij}$ and $\sum \omega_{ij}\ p_{i.}\ p_{.j}$ are the *weighted observed* and the *weighted expected* sample proportions of agreement, based on the assumption of independence of ratings, respectively (*see* **Rater Agreement – Kappa**). Subscripts i and j index I rows and J columns. The weights take a value between zero and one, and they should be ratios. For example, the weight of one can be assigned to the agreement cells (i.e., $\omega_{ij} = 1$, if $i = j$.), and 0.5 can be given to the partial agreement cells of adjacent categories (i.e., $\omega_{ij} = 0.5$, if $|i - j| = 1$.). And zero for the disagreement cells that are off by more than one category (i.e., $\omega_{ij} = 0$, if $|i - j| > 1$.). Thus, all weights range from zero to one, and the weights of 1, 0.5, and 0 are ratios of one another (see Table 1). Cohen's weighted kappa tends to be higher than Cohen's unweighted kappa because weighted kappa takes into account partial agreement between raters (see [3] for more information).

Data Example

Suppose two teachers taught a course together and independently graded 100 students in their class. Table 2 shows an artificial data example. Students were graded on an ordinal scale of A, B, C, D, or F grade. The five diagonal cells, shaded, indicate *absolute agreement* between two teachers. These cells get a weight of one. The adjacent cells that differ just by one category get a weight of 0.75 and the adjacent cells that differ by two and three categories get weights of 0.50 and 0.25, respectively. The proportions of the *weighted observed* agreement and the *weighted expected* agreement are calculated at 0.92 and 0.71, respectively. Cohen's weighted kappa is calculated at 0.72 (e.g., $\hat{\kappa}_{\text{w}} = (0.92 - 0.71)/(1 - 0.71)$) with standard error = 0.050, $z = 14.614$, $p < 0.01$. This value is higher than the value for Cohen's unweighted kappa = 0.62 with standard error = 0.061, $z = 10.243$, $p < 0.01$. The 10% increase in agreement reflects the partial weights

Table 1 An example of weights

		Rater B		
		Category 1	Category 2	Category 3
Rater A	Category 1	1(1)	0.5(0)	0(0)
	Category 2	0.5(0)	1(1)	0.5(0)
	Category 3	0(0)	0.5(0)	1(1)

Note: The numbers in parenthesis indicate weights of Cohen's unweighted kappa.

Table 2 Two teachers' ratings of students' grades (weights in parenthesis)

		Teacher B					
		A	B	C	D	F	Row sum
Teacher A	A	14 (1)	2 (0.75)	1 (0.50)	0 (0.25)	0 (0)	17
	B	4 (0.75)	25 (1)	6 (0.75)	1 (0.50)	0 (0.25)	36
	C	1 (0.50)	3 (0.75)	18 (1)	5 (0.75)	0 (0.50)	27
	D	0 (0.25)	1 (0.50)	3 (0.75)	14 (1)	1 (0.75)	19
	F	0 (0)	0 (0.25)	0 (0.50)	0 (0.75)	1 (1)	1
	Column sum	19	31	28	20	2	N = 100

given to the adjacent cells by one, two, and three categories.

References

[1] Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**, 37–46.

[2] Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, *Psychological Bulletin* **70**, 213–220.

[3] von Eye, A. & Mun, E.Y. (2005). *Analyzing Rater Agreement: Manifest Variable Methods*, Lawrence Erlbaum, Mahwah.

Eun Young Mun

Rater Bias Models

Kimberly J. Saudino

Volume 4, pp. 1716–1717

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Rater Bias Models

Rater bias occurs when behavioral ratings reflect characteristics of the rater in addition to those of the target. As such, rater bias is a type of method variance (i.e., systematically contributes variance to behavioral ratings that are due to sources other than the target). Rater personality, response styles (e.g., the tendency to rate leniently or severely), normative standards, implicit theories (e.g., stereotypes, halo effects) and even mood can influence ratings of a target's behaviors.

In phenotypic research, rater biases are of concern because they result in disagreement between raters while at the same time inflating correlations between variables that are rated by the same rater [2]. In twin studies (*see* **Twin Designs**), the effects of rater biases are more complex. When both members of a twin pair are assessed by the *same* rater, rater biases act to inflate estimates of **shared environmental** variance. That is, there is covariance between the biases that affect the ratings of each twin such that the rater tends to consistently overestimate or underestimate the behavior of *both* cotwins. This consistency across cotwins would act to inflate the similarity of both monozygotic (MZ) and dizygotic (DZ) twins. Thus, the correlation between cotwins is due in part to bias covariance. Since bias covariance would not be expected to differ according to twin type, its net effect will be to result in overestimates of shared environmental variance. However, if each member of a twin pair is assessed by a *different* informant, rater biases will result in overestimates of **nonshared environmental** variance. Here, there is no bias covariance between the ratings of each twin – different raters will have different biases that influence their behavioral ratings. Thus, rater bias will be specific to each twin and will contribute to differences between cotwins and, consequently, will be included in estimates of nonshared environment in quantitative genetic analyses.

Rater bias can be incorporated into quantitative genetic models [e.g., [1, 3]]. According to the Rater Bias model, observed scores are a function of the individual's latent phenotype, rater bias, and unreliability. The basic model requires at least two raters, each of whom has assessed both cotwins. Under this model, it is assumed that raters agree because they are assessing the same latent phenotype (i.e., what is common between raters is reliable trait variance). This latent phenotype is then decomposed into its genetic and environmental components. Disagreement between raters is assumed to be due to rater bias and unreliability. The model includes latent bias factors for each rater that account for bias covariance between the rater's ratings of each twin (i.e., what is common *within* a rater *across* twins is bias). Unreliability is modeled as rater-specific, twin-specific variances. Thus, this model decomposes the observed variance in ratings into reliable trait variance, rater bias, and unreliability; and allows for estimates of genetic and environmental influences on the reliable trait variance independent of bias and error [1].

As indicated above, the Rater Bias model assumes that rater bias and unreliability are the only reasons why raters disagree, but this may not be the case. Different raters might have different perspectives or knowledge about the target's behaviors. Therefore, it is important to evaluate the relative fit of the Rater Bias model against a model that allows raters to disagree because each rater provides different but valid information regarding the target's behavior. Typically, the Rater Bias model is compared to the Psychometric model (also known at the **Common Pathway model**). Like the Rater Bias model, this model suggests that correlations between raters arise because they are assessing a common phenotype that is influenced by genetic and/or environmental influences; however, this model also allows genetic and environmental effects specific to each rater. Thus, behavioral ratings include a common phenotype (accounting for agreement between raters) and specific phenotypes unique to each rater (accounting for disagreement between raters). As with the Rater Bias model, the common phenotype represents reliable trait variance. Because neither rater bias nor unreliability can result in the systematic effects necessary to estimate genetic influences, the specific genetic effects represent real effects that are unique to each rater [4]. Specific shared environmental influences may, however, be confounded by rater biases. When the Psychometric model provides a relatively better fit to the data than the Rater Bias model and rater-specific genetic variances estimated, it suggests that rater differences are not simply due to rater bias (i.e., that the raters to some extent assess different aspects of the target's behaviors).

References

[1] Hewitt, J.K., Silberg, J.L., Neale, M.C., Eaves, L.J. & Erikson, M. (1992). The analysis of parental ratings of children's behavior using LISREL, *Behavior Genetics* **22**, 292–317.

[2] Hoyt, W.T. (2000). Rater bias in psychological research: when is it a problem and what can we do about it? *Psychological Methods* **5**, 64–86.

[3] Neale, M.C. & Stevenson, J. (1989). Rater bias in the EASI temperament survey: a twin study, *Journal of Personality and Social Psychology* **56**, 446–455.

[4] van der Valk, J.C., van den Oord, E.J.C.G., Verhulst, F.C. & Boomsma, D.I. (2001). Using parent ratings to study the etiology of 3-year-old twins' problem behaviors: different views or rater bias? *Journal of Child Psychology and Psychiatry* **42**, 921–931.

KIMBERLY J. SAUDINO

Reactivity

PATRICK ONGHENA

Volume 4, pp. 1717–1718

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Reactivity

Reactivity is a threat to the construct validity of observational and intervention studies that refers to the unintended reaction of participants on being observed, or, more in general, on being in a study. It can entail a threat to the construct validity of outcomes or a threat to the construct validity of treatments [4, 5, 12].

Reactivity as a threat to the *outcome* construct validity may be involved if some sort of reactive or obtrusive assessment procedure is used, that is, an assessment procedure for which it is obvious to the participants that some aspect of their functioning is being observed or measured [13, 14]. Reactivity can occur both with respect to pretests and with respect to posttests. For example, in self-report pretests, participants may present themselves in a way that makes them eligible for a certain treatment [1]. Posttests can become a learning experience if certain ideas presented during the treatment 'fall into place' while answering a posttreatment questionnaire [3]. Pretest as well as posttest reactivity yields observations or measurements that tap different or more complex constructs (including participant perceptions and expectations) than the constructs intended by the researcher.

Reactivity as a threat to the *treatment* construct validity may be involved if participants are very much aware of being part of a study and interpret the research or treatment setting in a way that makes the actual treatment different from the treatment as planned by the researcher [9, 11]. Such confounding treatment reactivity can be found in the research literature under different guises: hypothesis guessing within experimental conditions [10], **demand characteristics** [6], **placebo effects** [2], and evaluation apprehension [7]. A closely related risk for treatment construct invalidity is formed by (*see* **Expectancy Effect by Experimenters**) [8], but here the prime source for bias are the interpretations of the researcher himself, while in reactivity, the interpretations of the participants are directly at issue.

Finally, it should be remarked that reactivity is a validity threat to both **quasi-experiments** and 'true' experiments. Random assignment procedures clearly provide no solution to reactivity problems. In fact, the random assignment itself might be responsible for changes in the measurement structure or meaning of the constructs involved, even before the intended treatment is brought into action [1]. In-depth discussion and presentation of methods to obtain unobtrusive measures and guidelines to conduct nonreactive research can be found in [9], [13], and [14].

References

[1] Aiken, L.S. & West, S.G. (1990). Invalidity of true experiments: self-report pretest biases, *Evaluation Review* **14**, 374–390.

[2] Beecher, H.K. (1955). The powerful placebo, *Journal of the American Medical Association* **159**, 1602–1606.

[3] Bracht, G.H. & Glass, G.V. (1968). The external validity of experiments, *American Educational Research Journal* **5**, 437–474.

[4] Campbell, D.T. & Stanley, J.C. (1963). *Experimental and Quasi-Experimental Designs for Research*, Rand-McNally, Chicago.

[5] Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Rand-McNally, Chicago.

[6] Orne, M.T. (1962). On the social psychology of the psychological experiment, *American Psychologist* **17**, 776–783.

[7] Rosenberg, M.J. (1965). When dissonance fails: on elimination of evaluation apprehension from attitude measurement, *Journal of Personality and Social Psychology* **1**, 28–42.

[8] Rosenthal, R. (1966). *Experimenter Effects in Behavioral Research*, Appleton-Century-Crofts, New York.

[9] Rosenthal, R. & Rosnow, R.L. (1991). *Essentials of Behavioral Research: Methods and Data Analysis*, McGraw-Hill, New York.

[10] Rosenzweig, S. (1933). The experimental situation as a psychological problem, *Psychological Review* **40**, 337–354.

[11] Rosnow, R.L. & Rosenthal, R. (1997). *People Studying People: Artifacts and Ethics in Behavioral Research*, Freeman, New York.

[12] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston.

[13] Webb, E.J., Campbell, D.T., Schwartz, R.D. & Sechrest, L. (1966). *Unobtrusive Measures: Nonreactive Research in the Social Sciences in the Social Sciences*, Rand McNally Chicago.

[14] Webb, E.J., Campbell, D.T., Schwartz, R.D., Sechrest, L. & Grove, J.B. (1981). *Nonreactive Measures in the Social Sciences*, 2nd Edition, Houghton Mifflin, Boston.

PATRICK ONGHENA

Receiver Operating Characteristics Curves

Daniel B. Wright

Volume 4, pp. 1718–1721

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Receiver Operating Characteristics Curves

The receiver operating characteristics curve, better known as ROC and sometimes called *receiver operating curve* or *relative operating characteristics*, is the principal graphical device for **signal detection theory (SDT)**. While ROCs were originally developed in engineering and psychology [3], they have become very common in medical diagnostic research (see [6] for a recent textbook, and also journals such as *Medical Decision Making, Radiology, Investigative Radiology*).

Consider an example: I receive about 100 emails a day, most of which are unsolicited junk mail. When each email arrives, it receives a 'spamicity' value from an email filter. Scores close to 1 mean that the filter predicts the email is 'spam'; scores close to 0 predict the email is nonspam ('ham'). I have to decide above what level of spamicity, I should automatically delete the emails. For illustration, suppose the data for 1000 emails and two different filters are those shown in Table 1.

The two most common approaches for analyzing such data are **logistic regression** and SDT. While in their standard forms, these are both types of the **generalized linear model**, they have different origins and different graphical methods associated with them. However, the basic question is the same: what is the relationship between spamicity and whether the email is spam or ham? ROCs are preferred when the decision criterion is to be determined and when the decision process itself is of interest, but when discrimination is important researchers should choose the logistic regression route. Because both these approaches are based on similar methods [1], it is sometimes advisable to try several graphical methods and use the ones which best communicate the findings.

There are two basic types of curves: empirical and fitted. Empirical curves show the actual data (sometimes with slight modifications). Fitted curves are based on some model. The fitted curves vary on how much they are influenced by the data versus the model. In general, the more parameters estimated in the model, the more influence the data will have. Because of this, some 'fitted' models are really just empirical curves that have been smoothed (*see* **Kernel Smoothing**) (see [5] and [6] for details of the statistics underlying these graphs).

The language of SDT is explicitly about accuracy and focuses on two types: sensitivity and specificity. Sensitivity means being able to detect spam when the email is spam; and specificity means just saying an email is spam if it is spam. In psychology, these are usually referred to as hits (or true positives) and correct rejections. The opposite of specificity is the false alarm rate: the proportion of time that the filter predicts that real email. Suppose the data in Table 1 were treated as predicting ham if spamicity is 0.5 or less and predicting spam if it is above. Table 2 shows the breakdown of hits and false alarms by whether an email is or is not spam, and whether the filter decrees it as spam or ham. Included also are the calculations for hit and false alarm rates. The filters only provide a spamicity score. I have to decide above what level of spamicity the email should be deleted. The choice of criterion is important. This is dealt with later and is the main advantage of ROCs over other techniques.

It is because all the values on one side of a criterion are classified as either spam or ham that ROCs are cumulative graphs. Many statistical packages

Table 1 The example data used throughout this article. Each filter had 1000 emails, about half of which were spam. The filter gave each email a spamicity rating

	Spamicity ratings										
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Filter one											
Email is ham	47	113	30	122	104	41	21	20	4	1	0
Email is spam	0	2	3	10	78	66	27	114	3	28	166
Filter two											
Email is ham	199	34	0	44	11	40	17	8	21	66	63
Email is spam	51	37	3	20	21	45	21	11	17	74	197

Table 2 The decision table if the criterion to delete emails was that the spamicity scores be 0.6 or above. While the two filters have similar hit rates, the first filter has a much lower false alarm rate (i.e., better specificity)

	Filter 1			Filter 2			
	<= 0.5	> 0.6	Rate	<= 0.5	> 0.6	Rate	Total
Is ham	457 correct rejection	46 false alarm	9%	328 correct rejection	175 false alarm	35%	503
Is spam	159 miss	338 hit	68%	177 miss	320 hit	64%	497
Total	616	384		505	495		1000

have cumulative data transformation functions (for example, in SPSS the CSUM function), so this can be done easily in mainstream packages. In addition, good freeware is available (for example, ROCKIT, RscorePlus), macros have been written for some general packages (for example, SAS and S-Plus), and other general packages contain ROC procedures (for example, SYSTAT). For each criterion, the number of hits and false alarms is divided by the number of spam and ham emails, respectively.

Figures 1(a) and 1(b) are the standard empirical ROCs and the fitted binormal curves (using RscorePlus [2]). The program assumes ham and spam vary on some dimension of spaminess (which is related to, but not the same as, the variable spamicity) and that these distributions of ham and spam are normally distributed on this dimension. The normal distribution assumption is there for historical reasons though nowadays researchers often use the logistic distribution, which yields nearly identical results and is simpler mathematically; the typical package offers both these alternatives plus others. In psychology, usually normality is assumed largely because Swets [3] showed that much psychological data fits with this assumption (and therefore also with the logistic distribution). In other fields like medicine, this assumption is less often made [6]. ROCs usually show the concave pattern of Figures 1(a) and

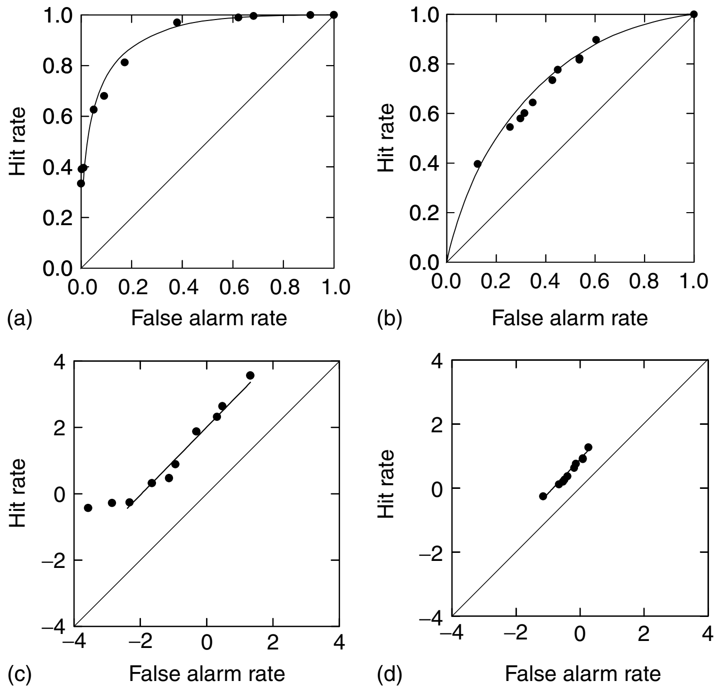


Figure 1 (a) and (b) plot the empirical hit rate and false alarm rates with fitted binormal model. Figures (c) and (d) show these graphs after they have been normalized so that the fitted line is straight

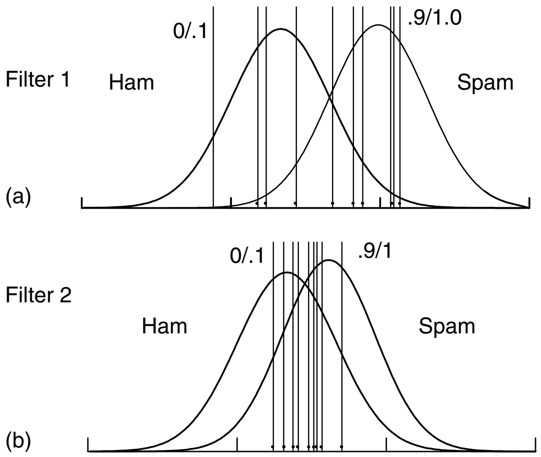


Figure 2 The models for filters (a) and (b) based on the fitted ROCs. The vertical lines show the 10 different response criteria

1(b). The diagonal lines stand for chance responding, and the further away from the diagonal the more diagnostic the variable spamicity is. Thus, comparing these two ROCs shows that the first filter is better.

The data are more easily seen if they are transformed so that the predicted lines are straight. These are sometimes called *normalized*, *standardized*, or *z-ROCs*. They can be calculated with the inverse normal function in most statistics packages and are available in most SDT software. These are shown in Figures 1(c) and 1(d). The straight lines in these graphs provide two main pieces of information. First, if the slopes are near 1, which they are for both these graphs, then they suggest the distributions for ham and spam have the same variance. The distance from the line to the diagonal is a measure of diagnosticity. If the slope of the line is 1, then this distance is the same at all points on the line and it corresponds to the SDT statistic d'. If the slope is not 1, then the distance between the diagonal and the line depends on the particular decision criterion. Several statistics are available for this (see [5] and [6]).

The fitted models in Figures 1(a–d) can be shown as normal distributions. Figures 2(a) and 2(b) indicate how well the ham and the spam can be separated by the filters. For the first filter, the spam distribution is about two standard deviations away from the ham distribution, while it is only about one standard deviation adrift for the second filter. The decision criteria are also included on these graphs. For example, the far left criterion is at the cut-off between 0.0 and 0.1 on the spamicity variable. As can be seen, for the second filter the criteria are all close together, which means the users would have less choice about where along the dimension they could choose a criterion. These graphs are useful ways of communicating the findings.

An obvious question is whether the normal distribution assumption is valid. There are statistical tests that look at this, but it can also be examined graphically. Note, however, that as these graphs are cumulative, the data are not independent. Consequently, conducting standard regressions for the data in Figures 1(c) and 1(d) is not valid, and in these cases, the analysis should be done on the noncumulative data or using specialized packages like those cited earlier.

More advanced and less restrictive techniques are discussed in [6], but are relatively rare. The more common procedure is simply to draw straight lines between each of the observed points. This is called the *trapezoid method* because the area below the line between each pair of points is a trapezoid. Summing these trapezoids gives a measure called A. This tends to underestimate the area under real ROC curves. The values of A can range between 0 and 1 with chance discrimination as 0.5. The first filter has $A = 0.92$ and the second has $A = 0.73$.

An important characteristic of SDT is how the relative values of sensitivity and specificity are calculated and used to determine a criterion. In the spam example, having low specificity would mean many real emails (ham) are labelled as spam and deleted. Arguably, this is more problematic than having to delete a few unsolicited spam. Therefore, if my filter automatically deletes messages, I would want the criterion to be high because specificity is more important than sensitivity.

To decide the relative value of deleting ham and not deleting spam, an equation from expected utility theory needs to be applied (slightly adapting the notation from [3], p.127):

$$S_{\text{opt}} = \frac{P(\text{ham})}{P(\text{spam})} \times \frac{V_{\text{CR}} - V_{\text{FA}}}{V_{\text{Hit}} - V_{\text{miss}}} \qquad (1)$$

where S_{opt} is the optimal slope, $P(\text{ham})$ is the probability that an item will be ham, $P(\text{spam})$ is 1 – P(ham), V_{CR} is the value of correctly not identifying ham as spam (this will be positive), V_{FA} is the value of incorrectly identifying ham as spam (negative), V_{Hit} is the value of correctly identifying spam (positive), and V_{miss} is the value of not identifying spam (negative). (It is worth noting here that a separate study is needed to estimate these utilities unless the minimum sensitivity or specificity is set externally; in such cases, simply go to this value on the ROC.) Thus, the odds value (*see* **Odds and Odds Ratios**) of ham is directly proportional to the slope. It is important to realize how important this baseline is for deciding the decision criterion. Often, people do not consider the baseline information when making decisions (see [4]).

Once S_{opt} is found, if one of the standard fitted ROC curves is used, then the optimal decision point is where the curve has this slope. For more complex fitted curves and empirical curves, start in the upper left-hand corner of the ROC with a line of slope S_{opt} and move towards the opposite corner. The point where the line first intersects the ROC shows where the optimal decision criterion should be. Because there are usually only a limited number of possible decision criteria, the precision of this method is usually adequate to identify the optimal criterion.

This discussion only touches the surface of an exciting area of contemporary statistics. This general procedure has been expanded to many different experimental designs (see [2] and [5]), and has been generalized for **meta-analyses**, correlated and biased data, robust methods, and so on [6].

References

[1] DeCarlo, L.T. (1998). Signal detection theory and generalized linear models, *Psychological Methods* **3**, 186–205.

[2] Harvey Jr, L.O. (2003). *Parameter Estimation of Signal Detection Models: RscorePlus User's Manual.* Version 5.4.0, (`http://psych.colorado.edu/~lharvey/`, as at 17.08.04).

[3] Swets, J.A. (1996). *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*, Lawrence Erlbaum.

[4] Tversky, A. & Kahneman, D. (1980). Causal schemes in judgements under uncertainty, in *Progress in Social Psychology*, M. Fishbein, ed., Lawrence Erlbaum, Hillsdale.

[5] Wickens, T.D. (2002). *Elementary Signal Detection Theory*, Oxford University Press, New York.

[6] Zhou, X.-H., Obuchowski, N.A. & McClish, D.K. (2002). *Statistical Methods in Diagnostic Medicine*, John Wiley & Sons, New York.

DANIEL B. WRIGHT

Recursive Models

Lilia M. Cortina

Volume 4, pp. 1722–1723

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Recursive Models

In behavioral research, statistical models are often *recursive*, meaning that causality flows only in one direction. In other words, these models only include unidirectional effects (e.g., variable A influences variable B, which in turn influences variable C; see Figure 1).

These cross-sectional recursive models do not include circular effects or *reciprocal causation* (e.g., variable A influences variable B, which in turn influences variable A; see Figure 2(a)), nor do they permit feedback loops (variable A influences variable B, which influences variable C, which loops back to influence variable A; see Figure 2(b)).

In addition, recursive models do not allow variables that are linked in a causal chain to have correlated disturbance terms (also known as 'error terms' or 'residuals'). If all of these criteria for being recursive are met, then the model is *identified* – that is, a unique value can be obtained for each free parameter from the observed data, yielding a single solution for all equations underlying the model (*see* **Identification**) [1, 4].

Distinctions are sometimes made between models that are 'recursive' versus 'fully recursive'. In a *fully recursive* model (or a 'fully saturated recursive model'), each variable directly influences all other variables that follow it in the causal chain; Figure 3(a) displays an example. Fully recursive models are *exactly identified* (or 'just identified'). It is impossible to disconfirm these models, because they will fit any set of observations perfectly. Moreover, fully recursive models often lack parsimony, being as complex as the observed relationships [3]. For these reasons, fully recursive models are most useful in the context of exploratory data analysis, but they are suboptimal for the testing of theoretically derived predictions.

Models that are recursive (but not fully so), on the other hand, omit one or more direct paths in the causal chain. In other words, some variables only influence other variables indirectly. In Figure 3(b), for instance, variable A influences variable D only by way of variables B and C (i.e., variables B and C *mediate* the effects of A on D). This amounts to a more restrictive model (i.e., constraining the direct relationship from variable A to variable D to zero) than in the fully recursive case, so these not-fully-recursive models are more impressive when they fit well [2, 3]. Recursive models in the behavioral sciences typically fall into this 'not fully saturated' category.

In contrast to recursive models, *nonrecursive* models include bidirectional or feedback effects (displayed in Figures 2(a) and 2(b), respectively), and/or they contain correlated disturbances for variables that are part of the same causal chain. Nonrecursive models are intuitively appealing in the behavioral sciences, given that many phenomena in the real world would seem to have mutual influences or feedback loops. However, a major drawback of nonrecursive models is that estimation can be difficult, especially with cross-sectional data. These models are often *underidentified*, meaning that there is more than one

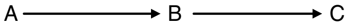


Figure 1 Graphical representation of a basic recursive model

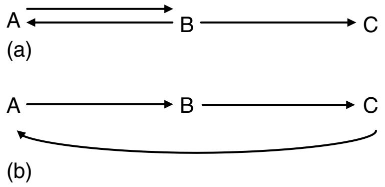


Figure 2 Examples of nonrecursive models; (a) nonrecursive model depicting reciprocal causation between variables A and B and (b) nonrecursive model depicting a feedback loop between variables A and C

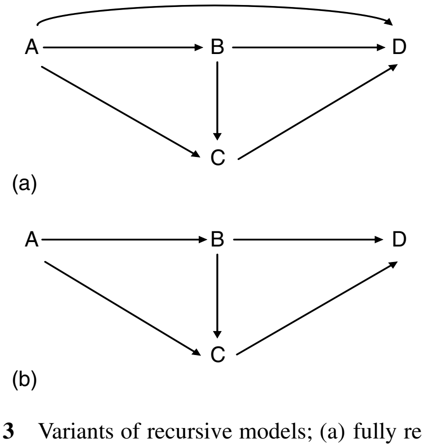


Figure 3 Variants of recursive models; (a) fully recursive model and (b) (not fully) recursive model

possible value for one or more parameters (*see* **Identification**); that is, multiple estimates fit the data equally well, making it impossible to arrive at a single unique solution [2]. For this reason, nonrecursive models are less common than recursive models in behavioral research.

References

[1] Bollen, K.A. (1989). *Structural Equations with Latent Variables*, John Wiley & Sons, New York.

[2] Klem, L. (1995). Path analysis, in *Reading and Understanding Multivariate Statistics*, L.G. Grimm & P.R. Yarnold, eds, American Psychological Association, Washington, pp. 65–98.

[3] MacCallum, R.C. (1995). Model specification: procedures, strategies, and related issues, in *Structural Equation Modeling: Concepts, Issues, and Applications*, R.H. Hoyle, ed., Sage Publications, Thousand Oaks, pp. 16–36.

[4] Pedhazur, E.J. (1997). *Multiple Regression in Behavioral Research: Explanation and Prediction*, 3rd Edition, Wadsworth Publishers.

(*See also* **Structural Equation Modeling: Overview**)

LILIA M. CORTINA

Regression Artifacts

David A. Kenny

Volume 4, pp. 1723–1725

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Regression Artifacts

Regression toward the mean has a long and somewhat confusing history. The term was invented by **Galton** [3] in 1886 when he noted that tall parents tended to have somewhat shorter children. Its presence is readily apparent in everyday life. Movie sequels tend to be worse than the original movies, married couples tend to experience a decline in marital satisfaction over time, and therapy usually results in improvement in symptoms.

The formal definition of regression toward the mean is as follows: Given a set of scores measured on one variable, X, it is a mathematical necessity that the expected value of a score on another variable, Y, will be closer to the mean, when both X and Y are measured in standard deviation units. As an example, if a person were two standard deviation units above the mean on intelligence, the expectation would be that the person would be less than two standard deviation units above the mean on any other variable. Typically, regression toward the mean is presented in terms of the same variable measured at two times, but it applies to any two variables measured on the same units or persons. As discussed by Campbell and Kenny [1], regression toward the mean does not depend on the assumption of linearity, the level of measurement of the variables (i.e., the variables can be dichotomous), or measurement error. Given a less than perfect correlation between X and Y, regression toward the mean is a mathematical necessity. It is not something that is inherently biological or psychological, although it has important implications for both biology and psychology.

Regression toward the mean applies in both directions, from Y to X as well as from X to Y. For instance, Galton [3] noticed that tall children tended to have shorter parents. Regression toward the mean does not imply increasing homogeneity over time because it refers to expected or predicted scores, not actual scores. On average, scores regress toward the mean, but some scores may regress away from the mean and some may not change at all.

Campbell and Kenny [1] discuss several ways to illustrate graphically regression toward the mean. The simplest approach is a **scatterplot**. An alternative is a *pair-link diagram.* In such a diagram, there are two vertical bars, one for X and one for Y. An individual unit is represented by a line that goes from one bar to the other. A particularly useful method for displaying regression toward the mean is a *Galton squeeze diagram* [1], which is shown in Figure 1. The left axis represents the scores on one measure, the pretest, the right axis represents the means on the second measure, the posttest. The line connecting the two axes represents the score on one measure and the mean score of those on the second measure. Regression toward the mean is readily apparent.

Because regression toward the mean refers to standard scores and an entire distribution of scores, its implications are less certain for raw scores. Consider, for instance, a measure of educational ability on which children are measured. All children receive some sort of intervention, and all are remeasured on that same variable. If the children improve at the second testing (i.e., the mean of the second testing is greater than the mean on the first testing), can we attribute that improvement to regression toward the mean or to the intervention? Without further information, we would be unable to answer the question definitively. If the children in the sample were below the mean in that population at time one and the mean and variance of the scores in the population was the same at both times without any intervention, then it is likely that the change is due to regression toward the mean. Of course, an investigator is not likely to know whether the children in the study are below the population mean and whether the population mean and standard deviation are unchanging.

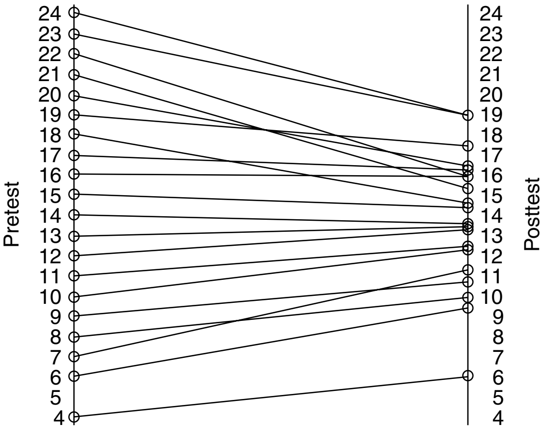


Figure 1 A Galton squeeze diagram illustrating regression toward the mean

Because raw scores do not necessarily regress toward the mean, some (e.g., [5]) have argued that regression toward the mean is not a problem in interpreting change. However, it should be realized that regression toward the mean is omnipresent with standard scores and therefore it has implications, uncertain though they are, for the interpretation of raw score change. Certainly, a key signal that regression toward the mean is likely to be a problem is when there is selection of extreme scores from a distribution.

Further complications also arise when two populations are studied over time. Consider two populations whose means and standard deviations differ from each other but their means and standard deviations do not change over time. If at time one, a person from each of the two populations is selected and these two persons have the very same score, at time two the two persons would not be expected to have the same score because each is regressing toward a different mean. It might be that one would be improving and the other worsening. Furby [2] has a detailed discussion of this issue.

Gilovich [4] and others have discussed how it is that lay people fail to take into account regression toward the mean in everyday life. For example, some parents think that punishment is more effective than reward. However, punishment usually follows bad behavior and, given regression toward the mean, the expectation is for improvement. Because reward usually follows good behavior, the expectation is a decline in good behavior and an apparent ineffectiveness of reward. Also, Harrison and Bazerman [6] discuss the often unnoticed effects of regression toward the mean in organizational contexts.

Regression toward the mean has important implications in prediction. In situations in which one has little information to make a judgment, often the best advice is to use the mean value as the prediction. In essence, the prediction is regressed to the mean.

In summary, regression toward the mean is a universal phenomenon. Nonetheless, it can be difficult to predict its effects in particular applications.

References

[1] Campbell, D.T. & Kenny, D.A. (1999). *A Primer of Regression Artifacts*, Guilford, New York.

[2] Furby, L. (1973). Interpreting regression toward the mean in developmental research, *Developmental Psychology* **8**, 172–179.

[3] Galton, F. (1886). Regression toward mediocrity in hereditary stature, *Journal of the Anthropological Institute of Great Britain and Ireland* **15**, 246–263.

[4] Gilovich, T. (1991). *How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life*, Free Press, New York.

[5] Gottman, J.M., ed. (1995). *The Analysis of Change*, Lawrence Erlbaum Associates, Mahwah.

[6] Harrison, J.R. & Bazerman, M.H. (1995). Regression toward the mean, expectation inflation, and the winner's curse in organizational contexts, in *Negotiation as a Social Process: New Trends in Theory and Research*, R.M. Kramer & D.M. Messick, eds, Sage Publications, Thousand Oaks, pp. 69–94.

DAVID A. KENNY

Regression Discontinuity Design

William R. Shadish and Jason K. Luellen

Volume 4, pp. 1725–1727

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Regression Discontinuity Design

Thistlewaite and Campbell [9] published the first example of a regression discontinuity design, the only **quasi-experimental design** that can yield unbiased estimates of the treatment effects [3, 4, 6, 7]. More extensive treatments of this design appear in [2], [8], [10], [11], and [12].

With the regression discontinuity design, the experimenter assigns participants to treatment conditions using a cutoff score on an assignment variable that is measured prior to treatment and is at least an ordinal measurement scale. The assignment variable is often a measure of need or merit, meaning that those who need a treatment or merit a reward the most are also the most likely to receive it. This feature of the design provides an ethical alternative when objections occur to randomly assigning needy or meritorious participants to a treatment, or to randomly depriving others of those same treatments.

The simplest design places units scoring on one side of the cutoff into the treatment condition, and those on the other side into the control condition. This design is diagrammed as shown in Table 1:

Table 1 A diagram of the regression discontinuity design

O_A	C	X	O
O_A	C		O

where O_A is the assignment variable, C indicates that participants are assigned to conditions using a cutoff score, X denotes treatment, O is a posttest observation, and the position of these letters from left to right indicates the time sequence in which they occur. If j is a cutoff score on O_A, then any participant scoring greater than or equal to j is in one group, and anything less than j is in the other. For example, suppose an education researcher implements a treatment program to improve math skills of third graders, but resource restrictions prohibited providing the treatment to all third-graders. Instead, the researcher could administer all third-graders a test of math achievement, and then assign those below a cutoff score on that test to the treatment, with those above being in the control group. If we draw a **scatterplot** of the relationship between scores on the assignment variable (horizontal axis) and scores on a math outcome measure (vertical axis), and if the training program improved math skills, the points in the scatterplot on the treatment side of the cutoff would be displaced upwards to reflect higher posttest math scores. And a regression line (*see* **Regression Models**) would show a discontinuity at the cutoff, that is, the size of the treatment effect. If the treatment had no effect, a regression line would not have this discontinuity. Shadish et al. [8] present many real-world studies that used this design. Many variants of the basic regression discontinuity design exist that allow comparing more than two conditions, that combine regression discontinuity with random assignment or with a quasi-experiment, or that improve statistical **power** [8, 10].

In most other quasi-experiments, it is difficult to rule out plausible alternative explanations for observed treatment effects – what Campbell and Stanley [1] call threats to **internal validity**. Threats to internal validity are less problematic with the regression discontinuity design, especially when the change in the regression line occurs at the cutoff and is large. Under such circumstances, it is difficult to conceive of reasons other than treatment that such a change in the outcome measure would occur for those immediately to one side of the cutoff, but not for those immediately to the other side.

Statistical regression (*see* **Regression to the Mean**) may seem to be a plausible threat to internal validity with the regression discontinuity design. That is, because groups were formed from the extremes of the distribution of the assignment variable, a participant scoring high on the assignment variable will likely not score as high on the outcome measure, and a participant scoring low on the assignment variable will likely not score as low on the outcome measure. But this regression does not cause a discontinuity in the regression line at the cutoff; it simply causes the regression line to turn more horizontal.

Selection is not a threat even though groups were selected to be different. Selection can be controlled statistically because the selection mechanism is fully known and measured. Put intuitively, the small difference in scores on the assignment variable for participants just to each side of the cutoff is not likely to account for a large difference in their scores on the outcome measure at the cutoff. *History* is a plausible threat if other interventions use the same cutoff for the same participants, which is usually unlikely. Because both groups are administered the same test measures, *testing*, per se, would not likely result in differences between the two groups. For the threat of *instrumentation* to apply, changes to the measurement instrument would need to occur exactly at the cutoff value of the assignment variable. *Attrition*, when correlated with treatment assignment, is probably the most likely threat to internal validity with this design.

In the regression discontinuity design, like a randomized experiment, the selection process is completely known and perfectly measured – the condition that must be met in order to successfully adjust for selection bias and obtain unbiased estimates of treatment effects. The selection process is completely known, assignment to conditions is based solely on whether a participant's score on the assignment variable is above or below the cutoff. No other variables, hidden or observed, determine assignment. Selection is perfectly measured because the pretest is strictly used to measure the selection mechanism. For example, if IQ is the assignment variable, although IQ scores imperfectly measure global intelligence, they have no error as a measure of how participants got into conditions.

The design has not been widely utilized because of a number of practical problems in implementing it. Two of those problems are unique to the regression discontinuity design. Treatment effect estimates are unbiased only if the functional form of the relationship between the assignment variable and the outcome measure is correctly specified, including nonlinearities and interactions. And statistical power of the regression discontinuity design is always lower than a comparably sized randomized experiment. Other problems are shared in common with a randomized experiment. Treatment assignment must adhere to the cutoff, just as assignment must adhere to a randomized selection process. In both cases, treatment professionals are not allowed to use judgment in assigning treatment. Cutoff-based assignments may be difficult to implement when the rate of participants entering the study is too slow or too fast [5]. Treatment crossovers may occur when participants assigned to treatment do not take it, or participants assigned to control end up being treated. Despite these difficulties, the regression discontinuity design holds a special place among cause-probing methods and deserves more thoughtful consideration when researchers are considering design options.

References

[1] Campbell, D.T. & Stanley, J.C. (1963). *Experimental and Quasi-Experimental Designs for Research*, Rand-McNally, Chicago.

[2] Cappelleri, J.C. (1991). *Cutoff-Based Designs in Comparison and Combination with Randomized Clinical Trials*, Unpublished doctoral dissertation, Cornell University, Ithaca, New York.

[3] Goldberger, A.S. (1972a). *Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations*, (Discussion paper No. 123), University of Wisconsin, Institute for Research on Poverty, Madison.

[4] Goldberger, A.S. (1972b). *Selection Bias in Evaluating Treatment Effects: The Case of Interaction*, (Discussion paper), University of Wisconsin, Institute for Research on Poverty, Madison.

[5] Havassey, B. (1988). *Efficacy of Cocaine Treatments: A Collaborative Study*, (NIDA Grant Number DA05582), University of California, San Francisco.

[6] Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading.

[7] Rubin, D.B. (1977). Assignment to treatment group on the basis of a covariate, *Journal of Educational Statistics* **2**, 1–26.

[8] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton-Mifflin, Boston.

[9] Thistlewaite, D.L. & Campbell, D.T. (1960). Regression-discontinuity analysis: an alternative to the ex post facto experiment, *Journal of Educational Psychology* **51**, 309–317.

[10] Trochim, W.M.K. (1984). *Research Design for Program Evaluation: The Regression-Discontinuity Approach*, California, Sage Publications.

[11] Trochim, W.M.K. (1990). The regression discontinuity design, in *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data*, L. Sechrest, E. Perrin & J. Bunker, eds, Public Health Service, Agency for Health Care Policy and Research, Rockville, pp. 199–140.

[12] Trochim, W.M.K. & Cappelleri, J.C. (1992). Cutoff assignment strategies for enhancing randomized clinical trials, *Controlled Clinical Trials* **13**, 190–212.

William R. Shadish and Jason K. Luellen

Regression Model Coding for the Analysis of Variance

Patricia Cohen

Volume 4, pp. 1727–1729

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Regression Model Coding for the Analysis of Variance

Using a **multiple linear regression** (MR) program to carry out **analysis of variance** (ANOVA) has certain advantages over use of an ANOVA program. They use the same statistical model (ordinary least squares, OLS), and thus make the same assumptions (normal distribution of population residuals from the model). Ordinarily, one's conclusions from an ANOVA will be the same as those from MR for the same data. However, with MR, you can code directly for tests of study hypotheses, whereas ANOVA can require a two-step process, beginning with tests of the significance of model 'factors', followed by specific tests relevant to your hypotheses. Second, MR permits using as many covariates as needed for substantive questions, including potential interactions among covariates. Third, MR permits tests of interactions of covariates with research factors, whereas **analysis of covariance** assumes such interactions to be absent in the population. Fourth, MR allows elimination of improbable and uninterpretable higher-order interactions in complex designs, improving the statistical power of the other tests. Finally, MR allows alternative treatments of unequal group ns, whereas standard ANOVA programs select a particular one (*see* **Type I, Type II and Type III Sums of Squares**). However, OLS MR programs handle **repeated measures analysis of variance** awkwardly at best.

How do you do it? An ANOVA study design consists of one or more research 'factors'. Each factor has two or more groups (categories) with participants in each group. Thus, one research factor (A) can consist of the three levels of some experimentally induced independent variable (IV). Another factor (B) can consist of four different ages, ethnicities, or litters of origin, and a third factor (C) can consist of two different testing times during the day. Thus, this study would consist of 3(A) × 4(B) × 2(C) = 24 different combinations or conditions; each combination includes some fraction of the total N participants on whom we have measured a dependent variable (DV). In an ANOVA, these three factors would be tested for significant differences among group means on the DV by the following F tests:

A with $g_A - 1 = 2$ *df* (degrees of freedom)
B with $g_B - 1 = 3$ *df*,
C with $g_C - 1 = 1$ *df*,
A × B interaction with $2\ df \times 3\ df = 6\ df$,
A × C interaction with $2\ df \times 1\ df = 2\ df$,
B × C interaction with $3\ df \times 1\ df = 3\ df$,
A × B × C interaction with $2\ df \times 3\ df \times 1\ df = 6\ df$, with a total of $2 + 3 + 1 + 6 + 2 + 3 + 6 = 23\ df$.

In a MR analysis of the same data, each of these 23 *df* is represented by a unique independent variable. These variables are usually created by coding each research factor and their interactions using one or more of the following coding systems.

Dummy variable coding: To dummy variable code each factor with g groups, each of the $g - 1$ groups is coded 1 on one and only one of the coded variables representing the factor and 0 on all the others. The 'left out' group is coded 0 on all $g - 1$ variables. When these $g - 1$ variables are entered into the regression analysis, (counterintuitively) the regression coefficient for each of the variables reflects the mean difference between the group coded 1 and the 'left out' group consistently coded 0. Thus, a statistical test of the coefficient provides a test for that difference. For this reason, dummy variable coding is most appropriate when there is a group that is logical to compare with all other groups in the research factor. The F-value for the contribution to R^2 of the $g - 1$ variables is precisely the same as it would have been for the ANOVA. However, the statistical **power** for the comparison of certain groups with the control group can be greater when the hypothesis of a difference from the control group is weak for some other groups. This occurs because the tests of group-control comparisons will not necessarily require an overall significant F test prior to the individual comparisons.

Effects coding of the $g - 1$ variables results in a contrast of each group's mean with the mean of the set of group means. This coding method is similar to dummy variable coding with an important difference and a very different interpretation. Instead of one group being selected as the reference group and coded 0 on all $g - 1$ variables, one group is coded -1 rather than 0 on every variable. This group should be selected to be the group of *least* interest because now the $g - 1$ regression coefficients contrast each of the other group means with the mean of the g group

Table 1 Alternative codes for three variables representing a four-group factor

Research factor	Dummy variable	Effects	Contrasts A	Contrasts B
Group L	1, 0, 0	1, 0, 0	+0.5, −0.5, 0	+0.5, +0.5, +0.5
Group M	0, 1, 0	−1, −1, −1 (least important)	+0.5, +0.5, 0	+0.5, −0.5, −0.5
Group N	0, 0, 1	0, 1, 0	−0.5, 0, −0.5	−0.5, +0.5, −0.5
Group P	0, 0, 0 (reference)	0, 0, 1	−0.5, 0, +0.5	−0.5, −0.5, +0.5

means. No explicit comparison of the '−1' group is represented in the regression coefficients, although it makes the usual contribution to the overall test of the variance in the group means, which is exactly equivalent to any of the other coding methods. (See Table 1 for dummy and effects codes for a four-group factor).

Contrast coding is an alternative method of coding the $g - 1$ variables representing the research factor. Contrast coding is actually not a single method, but a set of methods designed to match at least some of the study hypotheses. Thus, again, these tests are carried out directly in MR, whereas an ANOVA will typically provide an omnibus F test of a research factor that will need to be followed up with planned comparisons. Let us compare two alternative contrast codes representing a four-group research factor, comprised of groups L, M, N, and P, for which we will need three variables.

Suppose our primary interest is in whether groups L and M are different from groups N and P. In order for the regression coefficient to reflect this contrast, we will make the difference between the variable codes for these groups = 1. Thus, we code L and M participants = 1.0 and N and P participants = 0 on the first IV_1. At this point, we invoke the first rule of contrast coding: let the sum of the codes for each variable = 0, and recode L = M = +0.5 and N = P = −0.5. Our second major interest is in whether there are differences between L and M. Therefore, for IV_2 : L = −0.5 and M = +0.5, and, since we want to leave them out of consideration, N = P = 0.

Our third major interest is in whether there are differences between N and P. Therefore, for IV_3 : N = −0.5, P = +0.5, and L = M = 0. At this point, we invoke the second rule of contrast coding: let the sum of the products of each pair of variables = 0. The product of the codes for IV_1 and IV_2 is $(0.5 \times -0.5) = -0.25$ for group L, $(0.5 \times 0.5) = +0.25$ for M, and 0 for N and P. The code product of IV_1 and IV_3 is $(\pm 0.5 \times 0) = 0$ for L and M, $(-0.5 \times -0.5) = 0.25$ for N and $(-0.5 \times 0.5) = -0.25$ for P which also sum to 0. The IV_2 – IV_3 products similarly sum to 0. Finally, although the tests of statistical significance do not require it, a third rule also is useful: for each variable, let the difference between the negative weights and the positive weights = 1. Under these conditions, with these three variables in the regression equation predicting the dependent variable, the first coefficient equals the difference between combined groups L and M as compared to groups N and P, with its appropriate standard error (*SE*) and significance (t or F) test. The second regression coefficient equals the difference between groups L and M, with its appropriate *SE* and test, and the third coefficient equals the difference between groups N and P.

Suppose our study had a different content that made a different set of contrasts appropriate to our hypotheses. Perhaps the first contrast was the same L and M versus N and P as above, but the second was more appropriately a contrast of L and N (= +0.5) with M and P (= −0.5). The third variable that will satisfy the code rules would combine L = P = +0.5 and M = N = −0.5 (see Table 1).

A given study can use whatever combination of dummy variable, effects, and contrast codes that fits the overall hypotheses for different research factors. The coded variables representing the factors are then simply multiplied to create the variables that will represent the interactions among the factors (*see* **Interaction Effects**). Thus, if a given study participant had 1, 0, 0 on the three dummy variables representing the first factor and −0.5, −0.5, 0.25 on three contrast variables representing the second research factor, there would be nine variables representing the interaction between these factors. This participant would be scored −0.5, −0.5, 0.25, 0, 0, 0, 0, 0, 0. Because the first factor is dummy variable coded, all interactions are assessed as differences on the second factor contrasts between a particular group coded 1 and the reference group. The first three variables, therefore, reflect these differences for the group including this participant and the reference group on the three contrasts represented by the codes for the second research factor. The next six variables provide the same information for the other two groups coded 1 on the first factor variables 2 and 3, respectively.

Suppose that sample sizes in study cells representing the different combinations of research factors are not all equal. The analysis used in computer ANOVA programs in this case is equivalent to MR using effects codes, with regression coefficients contrasting each group's mean with the equally weighted mean of the *g* group means. Alternatively, rarely the investigator wishes the statistical tests to take the different cell sizes into account because their numbers appropriately represent the population of interest. Such a 'weighted means' set of comparisons involves using ratios of group sample sizes as coefficients (see Cohen, Cohen, West, & Aiken, 2003, p. 329 for details). (*See* **Type I, Type II and Type III Sums of Squares**.)

Repeated measure ANOVAs have multiple error terms, each of which would need to be identified in a separate MR using different functions of the original DV. Thus, MR is not generally employed. If the advantages of a regression model are desired, repeated DVs are better handled by multilevel regression analysis, in which, however, the statistical model is Maximum Likelihood rather than OLS. One advantage of this method is that it does not require that every trial is available for every individual. Codes of 'fixed effects' in such cases may employ the same set of options as described here. 'Nested' ANOVA designs in which groups corresponding to one research factor (B) are different for different levels of another research factor (A) are also currently usually managed in multilevel MR.

PATRICIA COHEN

Regression Models

Michael S. Lewis-Beck

Volume 4, pp. 1729–1731

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Regression Models

Regression is the most widely used procedure for analysis of observational, or nonexperimental, data in the behavioral sciences. A regression model offers an explanation of behavior in the form of an equation, tested against systematically gathered empirical cases. The simplest regression model may be written as follows:

$$Y = \mathrm{a} + \mathrm{b}X + \mathrm{e} \tag{1}$$

where $Y =$ the dependent variable (the phenomenon to be explained) and $X =$ the independent variable (a factor that helps explain, or at least predict, the dependent variable). The regression coefficient, b, represents the link between X and Y. The constant value, a, is always added to the prediction of Y from X, in order to reduce the level of error. The last term, e, represents the error that still remains after predicting Y.

The regression model of Eq. 1 is bivariate, with one independent variable and one dependent variable. A regression model may be multivariate, with two or more independent variables and one dependent variable. Here is a **multiple regression model**, with two independent variables:

$$Y = \mathrm{a} + \mathrm{b}X + \mathrm{c}Z + \mathrm{e} \tag{2}$$

where $Y =$ the dependent variable, X and $Z =$ independent variables, a = the constant, b and c = regression coefficients, and e = error. This multiple regression model has advantages over the bivariate regression model, first, because it promises a more complete account of the phenomenon under study, and second, because it more accurately estimates the link between X and Y.

The precise calculations of the link between the independent and dependent variables come from application of different estimation techniques, most commonly that of ordinary least squares (OLS) (*see* **Least Squares Estimation**). When researchers 'run regressions', the assumption is that the method of calculation is OLS unless otherwise stated. The least squares principle minimizes the sum of squared errors in the prediction of Y, from a line or plane. Since the principle is derived from the calculus, the sum of these squared errors is guaranteed 'least' of all possibilities, hence the phrase 'least squares'. If a behavioral scientist says 'Here is my regression model and here are the estimates', most likely the reference is to results from a multiple regression equation estimated with OLS.

Let us consider a hypothetical, but not implausible, example. Suppose a scholar of education policy, call her Dr. Jane Brown, wants to know why some American states spend more money on secondary public education than others. She proposes the following regression model, which explains differential support for public secondary education as a function of median income, elderly population, and private schools in the state:

$$\mathrm{Y} = \mathrm{a} + \mathrm{bX} + \mathrm{cZ} + \mathrm{dQ} + \mathrm{e}, \tag{3}$$

where Y = the per pupil dollar expenditures (in thousands) for secondary education by the state government, X = median dollar income of those in the state workforce; Z = people over 65 years of age (in percent); Q = private school enrollment in high schools of the state (scored 1 if greater than 15%, 0 otherwise).

Her hypotheses are that more income means more expenditure, that is, $\mathrm{b} > 0$; more elderly means less expenditure, that is, $\mathrm{c} < 0$; and a substantial private school enrollment means less expenditure, that is, $\mathrm{d} < 0$. To test these hypotheses, she gathers the variable scores on the 48 mainland states from a 2003 statistical yearbook, enters these data into the computer and, using a popular statistical software package, estimates the equation with OLS. Here are the results she gets for the coefficient estimates, along with some common supporting statistics:

$$\begin{gathered}\mathrm{Y} = 1.31 + .50^{*}\mathrm{X} - 0.29\mathrm{Z} - 4.66^{*}\mathrm{Q} + \mathrm{e} \\ (0.44) \quad (3.18) \quad (1.04) \quad (5.53) \\ \mathrm{R}^2 = 0.55 \qquad \mathrm{N} = 48 \end{gathered} \tag{4}$$

where Y, X, Z, Q and e are defined and measured as with Eq. 3; the estimates of coefficients a, b, c, and d are presented; e is the remaining error; the numbers in parentheses below these coefficients are absolute t-ratios; the $*$ indicates statistical significance at the 0.05 level; the R^2 indicates the coefficient of multiple determination; N is the size of the sample of 48 American states, excluding Alaska and Hawaii.

These findings suggest that, as hypothesized, income positively affects public high school spending, while the substantial presence of private schools negatively affects it. This assertion is based on the signs of the coefficients b and d, respectively, + and −, and on their statistical significance at 0.05. This level of statistical significance implies that, if Dr. Brown claims income matters for education, the claim is probably right, with 95% certainty. Put another way, if she did the study 100 times, only 5 of those times would she not be able to conclude that income related to education. (Note that the t-ratios of both coefficients exceed 2.0, a common rule-of-thumb for statistical significance at the 0.05 level.) Further, contrary to hypothesis, a greater proportion of elderly in the state does not impact public high school spending. This assertion is based on the lack of statistical significance of coefficient c, which says we cannot confidently reject the possibility that the percentage elderly in a state is not at all related to these public expenditures. Perhaps, for instance, with somewhat different measures, or a sample from another year, we might get a very different result from the negative sign we got in this sample.

Having established that there is a link between income and spending, what is it exactly? The coefficient $b = 0.50$ suggests that, on average, a one-unit increase in X (i.e., $1000 more in income) leads to a 0.50 unit increase in Y (i.e., a $500 increase in education expenditure). So we see that an income dollar translates strongly, albeit not perfectly, into an education dollar, as Dr. Brown expected. Further, we see that states that have many private schools (over 15%) are much less likely to fund public high school education for, on average, it is $4.66 \times 1000 = \$4660$ per pupil less in those states.

How well does this regression model, overall, explain state differences in public education spending? An answer to this question comes from the R^2, a statistic that reports how well the model fits the data. Accordingly, the R^2 here indicates that 55% of the variation in public high school expenditures across the states can be accounted for. Thus, the model tells us a lot about why states are not the same on this variable. However, almost half the variation $(1 - 0.55 = 0.45)$ is left unexplained, which means an important piece of the story is left untold. Dr. Brown should reconsider her explanation, perhaps including more variables in the model to improve its theoretical specification.

With a classical regression model, the OLS coefficients estimate real-world connections between variables, assuming certain assumptions are met. The assumptions include no specification error, no measurement error, no perfect multicollinearity, and a well-behaved error term. When these assumptions are met, the least squares estimator is BLUE, standing for Best Linear Unbiased Estimator. Among other things, this means that, on average, the estimated coefficients are true. Once assumptions are violated, they may be restored by variable transformation, or they may not. For example, if, in a regression model, the dependent variable is dichotomous (say, values are 1 if some property exists and 0 otherwise), then no transformation will render the least squares estimator BLUE. In this case, an alternative estimator, such as **maximum likelihood** (MLE), is preferred.

The need to use MLE, usually in order to achieve more efficient estimation, leads to another class of regression models, which includes logistic regression (when the dependent variable is dichotomous), polytomous **logistic regression** (when the dependent variable is ordered), or poisson regression (*see* **Generalized Linear Models (GLM)**)(when the dependent variable is an event count). Other kinds of regression models, which may use a least squares solution, are constructed to deal with other potential assumption violations, such as weighted least squares (to handle heteroskedasticity), local regression (to inductively fit a curve), censored regression (to deal with truncated observations on the dependent variable), seemingly unrelated regression (for two equations related through correlated errors but with different independent variables), spatial regression (for the problem of geographic autocorrelation), spline regression when there are smooth turning points in a line (*see* **Scatterplot Smoothers**), or stepwise regression (for selecting a subset of independent variables that misleadingly appears to maximize explanation of the dependent variable). All these variations are regression models of some sort, united by the notion that a dependent variable, Y, can be accounted for some independent variable(s), as expressed in an equation where Y is a function of some X(s).

Further Reading

Draper, N.R. & Smith, H. (1998). *Applied Regression Analysis*, 3rd Edition, Lawrence Erlbaum, Mahwah.

Fox, J. (2000). *Nonparametric Simple Regression: Smoothing Scatterplots*, Sage Publications, Thousand Oaks.

Kennedy, P. (2003). *A Guide to Econometrics*, 5th Edition, MIT Press, Cambridge.

Lewis-Beck, M.S. (1980). *Applied Regression: An Introduction*, Sage Publications, Beverly Hills.

Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables*, Sage Publications, Thousand Oaks.

MICHAEL S. LEWIS-BECK

Regression to the Mean

RANDOLPH A. SMITH

Volume 4, pp. 1731–1732

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Regression to the Mean

Francis Galton discovered regression to the mean in the 1800s; he referred to it as a 'tendency toward mediocrity' [1, p. 705]. In 1877, his work with sweet peas revealed that large parent seeds tended to produce smaller seeds and that small parent seeds tended to produce larger seeds [3]. Later, Galton confirmed that the same principle operated with human height: Tall parents tended to produce children who were shorter and short parents tended to have children who were taller. In all these examples, subsequent generations moved closer to the mean (regressed toward the mean) than the previous generation. These observations led to the definition of regression to the mean: extreme scores tend to become less extreme over time. Regression is always in the direction of a population mean [2].

Regression to the mean may lead people to draw incorrect causal conclusions. For example, a parent who uses punishment may conclude that it is effective because a child's bad behavior becomes better after a spanking. However, regression to the mean would predict that the child's behavior would be better shortly after bad behavior.

In the context of measurement, regression to the mean is probably due to measurement error. For example, if we assume that any measurement is made up of a true score + error, a high degree of error (either positive or negative) on one measurement should be followed by a lower degree of error on a subsequent measurement, which would result in a second score that is closer to the mean than the first score. For example, if a student guesses particularly well on a multiple-choice test, the resulting score will be high. On a subsequent test, the same student will likely guess correctly at a lower rate, thus resulting in a lower score. However, the lower score is due to error in measurement rather than the student's knowledge base.

Regression to the mean can be problematic in experimental situations; Cook and Campbell [2] listed it ('statistical regression', p. 52) as a threat to internal validity. In a repeated measures or prepost design, the experimenter measures the participants more than once (*see* **Repeated Measures Analysis of Variance**). Regression to the mean would predict that low scorers would tend to score higher on the second measurement and that high scorers would tend to score lower on the second measurement. Thus, a change in scores could occur as a result of a statistical artifact rather than because of an independent variable in an experiment. This potential problem becomes even greater if we conduct an experiment in which we use pretest scores to select our participants. If we choose the high scorers in an attempt to decrease their scores (e.g., depression or other psychopathology) or if we choose low scorers in an attempt to increase their scores (e.g., sociability, problem-solving behavior), regression to the mean may account for at least some of the improvement we observe from pre- to posttest [4].

Although regression to the mean is a purely statistical phenomenon, it can lead people to draw incorrect conclusions in real life and in experimental situations.

References

[1] Bartoszyński, R. & Niewiadomska-Bugaj, M. (1996). *Probability and Statistical Inference*, Wiley, New York.

[2] Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings*, Houghton Mifflin, Boston.

[3] Heyde, C.C. & Seneta, E. (2001). *Statisticians of the Centuries*, Springer, New York.

[4] Hsu, L.M. (1995). Regression toward the mean associated with measurement error and the identification of improvement and deterioration in psychotherapy, *Journal of Consulting and Clinical Psychology* **63**, 141–144.

RANDOLPH A. SMITH

Relative Risk

Brian S. Everitt

Volume 4, pp. 1732–1733

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Relative Risk

Quantifying risk and assessing risk involve the calculation and comparison of probabilities, although most expressions of risk are compound measures that describe both the probability of harm and its severity (*see* **Risk Perception**). The way risk assessments are presented can have an influence on how the associated risks are perceived. You might, for example, be worried if you heard that occupational exposure at your place of work doubled your risk of serious disease compared to the risk working at some other occupation entailed. You might be less worried, however, if you heard that your risk had increased from one in a million to two in a million. In the first case, it is relative risk that is presented, and in the second, it is an absolute risk.

Relative risk is generally used in medical studies investigating possible links between a risk factor and a disease. Formally relative risk is defined as

$$\text{Relative risk} = \frac{\text{incidence rate among exposed}}{\text{incidence rate among nonexposed}}. \tag{1}$$

Thus, a relative risk of five, for example, means that an exposed person is five times as likely to have the disease than a person who is not exposed.

Relative risk is an extremely important index of the strength of the association (*see* **Measures of Association**) between a risk factor and a disease (or other outcome of interest), but it has no bearing on the probability that an individual will contract the disease. This may explain why airline pilots who presumably have relative risks of being killed in airline crashes that are of the order of a thousandfold greater than the rest of us occasional flyers can still sleep easy in their beds. They know that the absolute risk of their being the victim of a crash remains extremely small.

Brian S. Everitt

Reliability: Definitions and Estimation

Alan D. Mead

Volume 4, pp. 1733–1735

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Reliability: Definitions and Estimation

Reliability refers to the degree to which test scores are free from error. Perfectly reliable scores would contain no error and completely unreliable scores would be composed entirely of error. In **classical test theory**, reliability is defined precisely as the ratio of the true score variance to the observed score variance and, equivalently, as one minus the ratio of error score variance to observed score variance:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}, \tag{1}$$

where $\rho_{XX'}$ is the reliability, σ_X^2 is the observed score variance, σ_E^2 is the error score variance, and σ_T^2 is the true score variance (see [2, 4]). If $\sigma_E^2 = 0$, then $\rho_{XX'} = 1$, while if $\sigma_E^2 = \sigma_X^2$ (i.e., error is at its maximum possible value), then $\rho_{XX'} = 0$. The symbol for reliability is the (hypothetical) correlation of a test score, X, with an independent administration of itself, X'.

Estimates of Reliability

True and error scores are not observable, so reliability must be estimated. Several different kinds of estimates have been proposed. Each estimate may suffer from various sources of error and, thus, no single estimate is best for all purposes. **Generalizability theory** may be used to conduct a generalizability study that can be helpful in understanding various sources of error, including interacting influences [1].

Test-retest reliability estimates require administration of a test form on two separate occasions separated by a period of time (the 'retest period'). The correlation of the test scores across the two administrations is the reliability estimate. The retest period may be quite short (e.g., minutes) or quite long (months or years). Sometimes, different retest periods are chosen, resulting in multiple test-retest estimates of reliability. For most tests, reliabilities decrease with longer retest periods, but the *rate of decline* decreases as the retest period becomes longer.

Test-retest reliability estimates may be biased because the same form is used on two occasions and test-takers may recall the items and answers from the initial testing ('memory effects'). Also, during the retest period, it is possible that the test-taker's true standing on the test could change (due to learning, maturation, the dynamic nature of the assessed trait, etc.). These 'maturation effects' might be a cause for the reliability estimate to underestimate the reliability at a single point in time. As a practical matter, test-retest reliability estimates entail the costs and logistical problems of two independent administrations.

Parallel forms estimates of reliability require the preparation of parallel forms of the test. By definition, perfectly parallel forms have equal reliability but require different questions. Both forms are administered to each test-taker and the correlation of the two scores is the estimate of reliability.

Parallel forms reliability estimates eliminate the 'memory effects' concerns that plague test-retest estimates but there still may be 'practice effects' and if the two forms are not administered on the same occasion, 'maturation effects' may degrade the estimate. As a partial answer to these concerns, administration of forms is generally counterbalanced. However, perhaps the most serious problem with this estimate can occur when the forms are substantially nonparallel. Reliability will be misestimated to the degree that the two forms are not parallel.

Split-half reliability estimates are computed from a single administration by creating two (approximately) parallel halves of the test and correlating them. This represents the reliability of half the test, and the *Spearman–Brown formula* is used to 'step up' the obtained correlation to estimate the reliability of the entire form.

Split-half reliability estimates retain many of the same strengths and pitfalls as parallel forms estimates while avoiding a second administration. To the extent that the method of dividing the items into halves does not create parallel forms, the reliability will be underestimated – because the lack of parallelism suppresses their correlation and also because the 'stepping up' method assumes essentially parallel forms. Also, different splitting methods will produce differing results. Common splitting methods include random assignment; odd versus even items; first-half versus second-half; and methods based on content or statistical considerations.

The *coefficient alpha* and *Kuder–Richardson* methods – commonly referred to generically as *internal consistency* methods – are very common estimates computed from single administrations. These methods are based on the assumption that the test measures a single trait and each item is essentially parallel to all other items. These methods compare the common covariance among items to the overall variability. Greater covariance among the items leads to higher estimates of reliability. Coefficient alpha is also commonly interpreted as the theoretical mean of all possible split-half estimates (based on all possible ways that the halves could be split).

Internal consistency methods make relatively stronger assumptions about the underlying test items and suffer from inaccuracies when these assumptions are not met. When the test items are markedly nonparallel, these reliability estimates are commonly taken to be lower-bound estimates of the true reliability. However, because these methods do not involve repeated administrations, they may be quite different from test-retest or parallel forms estimates and are unsuited to estimating the stability of test scores over time.

Other Reliability Topics

Standard Error of Measurement. The *standard error of measurement* (SEM) is the standard deviation of the error score component. The SEM is very important because it can be used to characterize the degree of error in individual or group scores. The SEM is

$$SEM = \sigma_E = \sigma_X\sqrt{1-\rho_{XX'}}. \qquad (2)$$

For users of test information, the SEM may be far more important and useful than the reliability. By assuming that the error component of a test score is approximately normally distributed, the SEM can be used to construct a true score **confidence interval** around an observed score. For example, if an individual's test score is 10 and the test has an SEM of 1.0, then the individual's true score is about 95% likely to be between 8 and 12 (or more correctly, 95% of the candidate confidence bands formed in this way will contain the unknown true scores of candidates).

The interpretation of the SEM should incorporate the kind of reliability estimate used. For example, a common question raised by test-takers is how they might score if they retest. Confidence intervals using a test-retest reliability are probably best suited to answering such questions.

Reliability is *not* Invariant Across Subpopulations. Although reliability is commonly discussed as an attribute of the test, it is influenced by the variability of observed scores (σ_X^2) and the variability of true scores (σ_T^2). The reliability of a test when used with some subpopulations will be diminished if the subpopulation has a lower variability than the general population. For example, an intelligence test might have a high reliability when calculated from samples drawn from the general population but when administered to samples from narrow subpopulations (e.g., geniuses), the reduced score variability causes the reliability to be attenuated; that is, the intelligence test is less precise in making fine distinctions between geniuses as compared to ranking members representative of the breadth of the general population.

As a consequence, test developers should carefully describe the sample used to compute reliability estimates and test users should consider the comparability of the reliability sample to their intended population of test-takers.

The Spearman-Brown Formula. The Spearman–Brown formula provides a means to estimate the effect of lengthening or shortening a test:

$$\rho^*_{XX'} = \frac{n\rho_{XX'}}{1+(n-1)\rho_{XX'}}, \qquad (3)$$

where $\rho^*_{XX'}$ is the new (estimated) reliability, n is the fraction representing the change in test length ($n = 0.5$ implies halving test length, $n = 2$ implies doubling) and $\rho_{XX'}$ is the current reliability of the test. A critical assumption is that the final test is essentially parallel to the current test (technically, the old and new forms must be 'tau-equivalent'). For example, if a 10-item exam were increased to 30 items, then the Spearman–Brown reliability estimate for the 30-item exam would only be accurate to the extent that the additional 20 items have the same characteristics as the existing 10.

True Score Correlation. The correlation between the observed score and the true score is equal to the square root of the reliability: $\rho_{XT} = \sqrt{(\rho_{XX'})}$. If the classical test theory assumptions regarding the independence of error score components are true, then the observed test score, X, cannot correlate with any variable more highly than it does with its true score, T. Thus, the square root of the reliability is considered an upper bound on correlations between test scores and other variables. This is primarily a concern when the test score reliability is low. A correlation of 0.60 between a test score and a criterion may be quite high if the reliability of the test score is only about 0.40. In this case, the observed correlation of 0.60 is about the maximum possible ($\sqrt{0.40}$).

These considerations give rise to the standard formula for estimating the correlation between two variables as if they were both perfectly reliable:

$$\rho_{T_X T_Y} = \frac{\rho_{XY}}{\sqrt{\rho_{XX'}\rho_{YY'}}}, \tag{4}$$

where $\rho_{T_X T_Y}$ is the estimated correlation between the true scores of variables X and Y, ρ_{XY} is the observed correlation between variables X and Y, $\rho_{XX'}$ is the reliability of score X, and $\rho_{YY'}$ is the reliability of score Y. This hypothetical relationship is primarily of interest when comparing different sets of variables without the obscuring effect of different reliabilities and when assessing the correlation of two constructs without the obscuring effect of the unreliability of the measures.

Reliability for Speeded Tests

Some estimates of reliability are not applicable to highly speeded tests. For example, split-half and internal consistency estimates of reliability are inappropriate for highly speeded tests. Reliability methods that involve retesting are probably best, although they may be subject to practice effects. A generalizability study may be particularly helpful in identifying factors that influence the reliability of speeded tests.

Reliability and IRT

Reliability is a concept defined in terms of classical test theory. Modern **item response theory (IRT)** provides much stronger and more flexible results (see, for example, [3]). Such results reveal that reliability and SEM are simplifications of the actual characteristics of tests. In place of reliability and SEM, IRT provides *information functions* for tests and items. The inverse of the test information function provides the standard error of measurement *at a particular point on the latent trait* (*see* **Latent Variable**). IRT analysis generally reveals that scores near the middle of the score distribution are considerably more reliable (i.e., have lower SEM) than scores in the tails of the distribution. This often means that ordering test-takers in the upper and lower percentiles is unreliable, perhaps to the point of being arbitrary.

References

[1] Brennan, R.L. (2001). *Generalizability Theory*, Springer-Verlag, New York.

[2] Gulliksen, H. (1950). *Theory of Mental Test Scores*, Wiley, New York.

[3] Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*, Sage Publications, Newbury Park.

[4] Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading.

Alan D. Mead

Repeated Measures Analysis of Variance

Brian S. Everitt

Volume 4, pp. 1735–1741

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Repeated Measures Analysis of Variance

In many studies carried out in the behavioral sciences and related disciplines. The response variable is observed on each subject under a number of different conditions. For example, in an experiment reported in [6], the performance of field-independent and field-dependent subjects (twelve of each type) in a reverse Stroop task was investigated. The task required reading of words of two types, color and form names, under three cue conditions – normal, congruent, and incongruent. For instance, the word 'yellow' displayed in yellow would be congruent, whereas 'yellow' displayed in blue would be incongruent. The dependent measure was the time (msec) taken by a subject to read the words. The data are given in Table 1. Note that each subject in the experiment has six time measurements recorded. The response variable is time in milliseconds.

The data in Table 1 contain repeated measurements of a response variable on each subject. Researchers, typically, adopt the repeated measures paradigm as a means of reducing error variability and/or as a natural way of assessing certain phenomena (e.g., developmental changes over time, and learning and memory tasks). In this type of design, effects of experimental factors giving rise to the repeated measurements (the so-called *within subject factors*; word type and cue condition in Table 1) are assessed relative to the average response made by a subject on all conditions or occasions. In essence, each subject serves as his or her own control, and, accordingly, variability caused by differences in average response time of the subjects is eliminated from the extraneous error variance. A consequence of this

Table 1 Field independence and a reverse Stroop task

	Form names			Color names		
Subject	Normal condition	Congruent condition	Incongruent condition	Normal condition	Congruent condition	Incongruent condition
			Field-independent			
1	191	206	219	176	182	196
2	175	183	186	148	156	161
3	166	165	161	138	146	150
4	206	190	212	174	178	184
5	179	187	171	182	185	210
6	183	175	171	182	185	210
7	174	168	187	167	160	178
8	185	186	185	153	159	169
9	182	189	201	173	177	183
10	191	192	208	168	169	187
11	162	163	168	135	141	145
12	162	162	170	142	147	151
			Field-dependent			
13	277	267	322	205	231	255
14	235	216	271	161	183	187
15	150	150	165	140	140	156
16	400	404	379	214	223	216
17	183	165	187	140	146	163
18	162	215	184	144	156	165
19	163	179	172	170	189	192
20	163	159	159	143	150	148
21	237	233	238	207	225	228
22	205	177	217	205	208	230
23	178	190	211	144	155	177
24	164	186	187	139	151	163

is that the **power** to detect the effects of within-subject experimental factors is increased compared to testing in a between-subject design. But this advantage of a repeated measures design comes at the cost of an increase in the complexity of the analysis and the need to make an extra assumption about the data than when only a single measure of the response variable is made on each subject (see later). This possible downside of the repeated measures approach arises because the repeated observations made on a subject are very unlikely to be independent of one another. In the data in Table 1, for example, an individual who reads faster than average under say one cue condition, is likely to read faster than average under the other two cue conditions. Consequently, the repeated measurements are likely to be correlated, possibly substantially correlated. Note that the data in Table 1 also contains a *between-subject factor*, cognitive style with two levels, field-independent and field-dependent.

Analysis of Variance for Repeated Measure Designs

The variation in the observations in Table 1 can be partitioned into a part due to between-subject variation, and a part due to within-subject variation. The former can then be split further to give a between-cognitive style mean square and an error term that can be used to calculate a mean square ratio and assessed against the appropriate F distribution to test the hypothesis that the mean reading time in the population of field-dependent and field-independent subjects is the same (see later for the assumptions under which the test is valid).

The within-subject variation can also be separated into parts corresponding to variation between the levels of the within-subject factors, their **interaction**, and their interaction with the between-subject factor along with a number of error terms. In detail, the partition leads to sums of squares and so on for each of the following:

- Cue condition, Cognitive style × Cue condition, error;
- Word type, Cognitive style × Word type, error;
- Cue condition × Word type, Cognitive style × Cue condition × Word type, error.

Again, mean square ratios can be formed using the appropriate error term and then tested against the relevant F distribution to provide tests of the following hypotheses:

- No difference in mean reading times for different cue conditions, and no cognitive style × cue condition interaction.
- No difference in mean reading time for the two types of word, and no interaction of word type with cognitive style.
- No interaction between cue condition and word type, and no second order interaction of cognitive style, cue condition, and word type.

Full details of both the model behind the partition of within-subject variation and the formulae for the relevant sums of squares and so on are given in [2, 4]. The resulting **analysis of variance** table for the cognitive style data is given in Table 2.

Before interpreting the results in Table 2, we need to consider under what assumptions is such an analysis of variance approach to repeated measure designs valid? First, for the test of the between-subject factor, we need only normality of the response and homogeneity of variance, both familiar from the analysis of variance of data not involving repeated measures. But, for the tests involving the within-subject factors, there is an extra assumption needed to make the F tests valid, and it is this extra assumption that is particularly critical in the analysis of variance ANOVA of repeated measures data. The third assumption is known as **sphericity**, and requires that the variances of the differences between all pairs of repeated measurements are equal to each other and the same in all levels of the between-subject factors. The sphericity condition also implies a constraint on the covariance matrix (*see* **Correlation and Covariance Matrices**) of the repeated measurements, namely, that this has a *compound symmetry structure*, i.e., equal values on the main diagonal (the variances of the response under the different within-subject experimental conditions), and equal off-diagonal elements (the covariances of each pair of repeated measures). And this covariance matrix has to be equal in all levels of the between-group factors.

If, for the moment, we simply assume that the cognitive style data meet the three assumptions of normality, homogeneity, and sphericity, then the

Table 2 Repeated measures analysis of variance of reverse Stroop data

Source	Sum of squares	DF	Mean square	Mean square ratio	P value
Cognitive style	17578.34	1	17578.34	2.38	0.137
Error	162581.49	22	7390.07	2.38	
Word type	22876.56	1	22876.56	11.18	0.003
Word type × cognitive style	4301.17	1	4301.17	2.10	0.161
Error	45019.10	22	2046.32		
Condition	5515.39	2	2757.69	21.81	<0.001
Condition × Cognitive style	324.06	2	162.03	1.28	0.288
Error	5562.56	44	126.42		
Word type × Condition	450.17	2	225.08	3.14	0.053
Word type × Condition × Cognitive style	111.05	2	55.53	0.78	0.467
Error	3153.44	44	71.67		

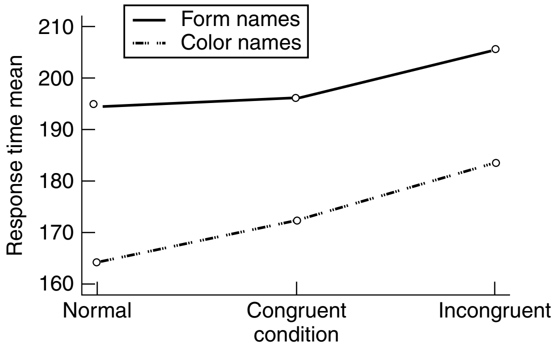


Figure 1 Interaction plot for cue condition × word type means

results in Table 2 lead to the conclusion that mean reading time differs in the three cue conditions and for the two types of word. There is also a suggestion of a cue condition × word type interaction. The mean reaction times for the three cue conditions are; normal −179.6 msec, congruent −184.3 msec, incongruent −194.5 msec. Under the incongruent condition, reaction times are considerably longer than under the other two conditions. The mean reaction times for the two word types are: form names −198.8 msec, color names −173.5 msec. Reaction times are quicker for the color names. To examine the cue condition × word type interaction, a graph of the six relevant mean reaction times is useful. This is shown in Figure 1 (*see* **Interaction Plot**). There is some slight evidence that the difference in reaction time means for the two word conditions is greater in the normal condition than in the other two conditions.

What happens if the assumptions are not valid? Concentrating on the one specific to repeated measures data, namely, sphericity, if this is not valid then the F tests in the repeated measures analysis of variance are positively biased, leading to an increase in the probability of rejecting the null hypothesis when it is true, that is, increasing the size of the Type I error over the nominal value set by the researcher. This will lead to an investigator claiming a greater number of 'significant' results than are actually justified by the data.

How likely are repeated measures data in behavioral science experiments to depart from the sphericity assumption? This is a difficult question to answer, but if the within-subjects conditions are given in a random order to each subject in the study, then the assumption is probably easier to justify than when they are presented in the same order, particularly if there is a substantial time difference between the first condition and the last. The problem is that, in the latter case, observations closer together in time are likely to be more highly correlated than those taken further apart, thus violating the required compound symmetry structure for the repeated measures covariance matrix.

Where departures from sphericity *are* suspected or, perhaps, indicated by the formal test for the condition (see **Sphericity Test**), there are two approaches that might be used:

1. Correction factors,
2. **Multivariate analysis of variance (MANOVA)**.

Correction Factors in the Analysis of Variance of Repeated Measures Data

The effects of departures from the sphericity assumption in a repeated measures analysis of variance have been considered in [3, 5], and it has been shown that the extent to which a set of repeated measures data deviates from the sphericity assumption can be summarized in terms of a quantity that is a function of the covariances and variances of the measures (see [2] for an explicit definition). Furthermore, an estimate of this quantity based upon sample variances and covariances can be used to decrease the degrees of freedom of the *F* tests for within-subject effects, to account for the departure from sphericity. (In fact, two different estimates of the correction factor have been suggested, one by Greenhouse and Geisser [3], and one by Huynh and Feldt [5].) The result is that larger mean square ratio values will be needed to claim statistical significance than when the correction is not used; consequently, the increased risk of falsely rejecting the null hypothesis is confronted. For the cognitive style data, the adjusted *P* values obtained using each estimate of the correction factor are shown in Table 3. Note that since word type has only two levels, the 'corrected' values are identical to those in Table 2, and for these data, the *P* values that do change do not differ greatly from the uncorrected values in Table 2. (This is, perhaps, not surprising here since the test for sphericity indicates that the assumption is valid.) More detailed examples of the use of this correction factor approach are given in [2, 4].

Table 3 Adjusted *P* values for reverse Stroop data

Factor	Greenhouse/ Geisser	Huynh/ Feldt
Word type	0.003	0.003
Word type × Cognitive style	0.161	0.161
Condition	<0.001	<0.001
Condition × Cognitive style	0.286	0.287
Word type × Condition	0.063	0.057
Word type × Condition × Cognitive style	0.447	0.460

Multivariate Analysis of Variance for Repeated Measure Designs

An alternative to the use of the correction factors in the analysis of repeated measures data when the sphericity assumption is judged to be inappropriate is to use a multivariate analysis of variance. Details are given in [2], but the essential feature of this approach is to transform the repeated measures so that they characterize different aspects of the within-subject factors, that is, main effects and interactions, and then use multivariate test criteria on the resulting sets of variables in the usual way. For example, in the cognitive style data, differences between the three cue conditions could be represented by differences (averaged over word types) between normal and congruent, and congruent and incongruent. The cue condition × word type interaction effect would involve differences between color and form names for these same differences in cue conditions. In the multivariate situation, there is no single test statistic that is always the most powerful for detecting all types of departures from the null hypothesis of the equality of mean vectors, and a number of different test statistics are in use. For details of these test statistics and the differences between them, see the multivariate analysis of variance entry. (Here, since there are only two groups involved, all the test criteria are equivalent.)

The results from the multivariate procedure are given in Table 4. (Note that since word type has only two levels, the multivariate result is equivalent to the univariate result given in Table 2.) The results again indicate highly significant condition and word type main effects, but, here, the condition × word type interaction is, unlike in the univariate analysis, also highly significant.

The main advantage of using MANOVA for the analysis of repeated measures designs is that no assumptions now have to be made about the pattern of covariances between the repeated measures. In particular, these covariances need not satisfy the compound symmetry condition. A disadvantage of using MANOVA for repeated measures is often stated to be the technique's relatively low power when the assumption of compound symmetry is actually valid. However, the power of the univariate and multivariate analysis of variance approaches when compound symmetry holds is compared in [1] with the conclusion that the latter is nearly as powerful as

Table 4 Multivariate analysis of variance of reverse Stroop data

Effect	Value	*F*	Hypothesis df	Error *df*	*P* value
Word type:					
Pillai's trace	0.34	11.18	1	22	0.003
Wilk's lambda	0.66	11.18	1	22	0.003
Hotelling's trace	0.51	11.18	1	22	0.003
Roy's largest root	0.51	11.18	1	22	0.003
Word type × Cognitive style:					
Pillai's trace	0.09	2.10	1	22	0.161
Wilk's lambda	0.91	2.10	1	22	0.161
Hotelling's trace	0.10	2.10	1	22	0.161
Roy's largest root	0.10	2.10	1	22	0.161
Condition:					
Pillai's trace	0.66	20.73	2	21	<0.001
Wilk's lambda	0.34	20.73	2	21	<0.001
Hotelling's trace	1.97	20.73	2	21	<0.001
Roy's largest root	1.97	20.73	2	21	<0.001
Condition × Cognitive style:					
Pillai's trace	0.13	1.63	2	21	0.220
Wilk's lambda	0.87	1.63	2	21	0.220
Hotelling's trace	0.16	1.63	2	21	0.220
Roy's largest root	0.16	1.63	2	21	0.220
Word type × Condition:					
Pillai's trace	0.37	5.32	2	21	0.014
Wilk's lambda	0.66	5.32	2	21	0.014
Hotelling's trace	0.51	5.32	2	21	0.014
Roy's largest root	0.51	5.32	2	21	0.014
Condition × Word type × Cognitive style:					
Pillai's trace	0.48	0.53	2	21	0.595
Wilk's lambda	0.95	0.53	2	21	0.595
Hotelling's trace	0.51	0.53	2	21	0.595
Roy's largest root	0.51	0.53	2	21	0.595

the former when the number of observations exceeds the number of repeated measures by more than 20.

Summary

An analysis of variance can often be safely applied to repeated measures data arising from psychological experiments when the within-subject conditions are presented in a random order to each subject, since with such designs, the sphericity assumption is likely to be justified. There is, however, one type of repeated measures data where sphericity is very unlikely to hold, namely, **longitudinal data**. For such data, the only within-subject factor is "time", and so randomization is no longer an option. Consequently, it is very likely that observations taken closer together in the study will be more similar than those separated by a longer time interval; consequently, assuming compound symmetry will not, in general, be sensible. For this reason, longitudinal data requires more sophisticated approaches, for example, **linear multilevel models**. Such models can also deal with the frequently occurring practical problem of **missing data** in longitudinal data sets, in particular when such observations are missing because subjects drop out of the study (*see* **Dropouts in Longitudinal Studies: Methods of Analysis**).

References

[1] Davidson, M.L. (1972). Univariate versus multivariate tests in repeated measures experiments, *Psychological Bulletin* **77**, 446–452.

[2] Everitt, B.S. (2000). *Statistics for Psychologists*, Lawrence Erlbaum, Mahwah.

[3] Greenhouse, S.W. & Geisser, S. (1959). On the methods in the analysis of profile data, *Psychometrika* **24**, 95–112.

[4] Howell, D.C. (2002). *Statistical Methods for Psychology*, Duxbury, Pacific Grove.

[5] Huynh, H. & Feldt, L.S. (1976). Estimated of the correction for degrees of freedom for sample data in randomized block and split-plot designs, *Journal of Educational Statistics* **1**, 69–82.

[6] Pahwa, A. & Broota, K.D. (1981). Field-independence, field dependence as a determinant of colour-word interference, *Journal of Psychological Research* **30**, 55–61.

Further Reading

Crowder, M.J. & Hand, D.J. (1990). *Analysis of Repeated Measures*, Chapman & Hall, London.

(*See also* **Generalized Estimating Equations (GEE)**; **Generalized Linear Mixed Models**; **Marginal Models for Clustered Data**)

BRIAN S. EVERITT

Replicability of Results

Bruce Thompson

Volume 4, pp. 1741–1743

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Replicability of Results

Quantitative researchers seek to identify relationships that recur under stated conditions. Scholars in the physical sciences have the luxury of the idiosyncrasies of human personality not confounding their results. A physicist observing a quark and a neutrino running away from each other may make an inference that these two atomic particles have like charges. The physicist need not qualify generalizations with statements about the nutrition of the quark during gestation, or the quality of schooling received by different quarks during their early years.

The business of behavioral science is much more difficult. Behavioral scientists attempt to formulate generalizations about people, which hold up reasonably well, but recognize that few statements apply equally well to all people. Behavioral scientists attempt to overcome the vagaries of individual differences in various ways, including conducting studies with large numbers of people, so that the idiosyncrasies may 'wash out' within the large sample.

Some behavioral scientists erroneously believe that statistical significance testing evaluates the replicability of results. In fact, statistical significance does not evaluate result replicability, as **Cohen** [1] and others [10] have shown. Because statistical tests do not evaluate result replicability, and replicability is very important, other methods must be used to test result replicability.

Thompson [9] suggested that there are two kinds of replicability evidence: external and internal. External replicability evidence requires the researchers to conduct the research study a second time, to determine whether the results are stable.

Another form of external replicability analysis involves 'meta-analytic thinking' (*see* **Meta-Analysis**) in which the researcher focuses on explicitly and directly comparing study effect sizes with the effect sizes in the related prior studies [2, 11, 13]. If all the effects across studies are similar, the researcher has more confidence that the results in a given study are not purely serendipitous.

The most persuasive evidence of result replicability is actual study replication. The problem is that most researchers do not have the time or the resources to replicate every study prior to publishing their results or submitting their theses. In such cases, internal replicability analyses [9] are a partial substitute for true external replication.

The basic idea of internal replicability analyses is to mix up the participants in different ways, to determine whether the results remain stable across different combinations of people. The intent is to approximate modeling the variations in personalities that would occur if an actual new sample had been drawn. These internal replicability analyses are never as good as true replication, but are generally more informative as regards replicability compared to what many researchers do to establish replicability of their results - nothing!

There are three primary principles of logic for conducting internal replicability analyses: **cross-validation**, the **jackknife**, and the **bootstrap**. (Carl Huberty has suggested combining the latter two methods to create another alternative - the jackstrap. The more serious point is that the researcher can do whatever seems reasonable to evaluate result replicability, even if the approach has not been previously employed.)

Traditionally, cross-validation involves randomly splitting the sample into two nonoverlapping subsamples, replicating the analysis in both subsamples, and then conducting additional empirical analyses to determine whether the results are similar in the two subsamples. Thompson [6] provides details in a heuristic example for the multiple regression case. Factor analytic examples are provided by Thompson [12].

The jackknife was popularized by **John Tukey**. In the jackknife, the primary resampling mechanism is to redo the analysis by successively dropping out subsets of participants of a given size, k (e.g., $k = 1$, $k = 2$). For example, the researcher might drop out subsets of participants where each dropped subset is size 1. In a regression involving 300 participants, the analysis would be done with all 300 participants. And the regression would then be done 300 more times, each with a sample size equal to $n - 1$, where a different participant is dropped each time.

The bootstrap was popularized by Bradley Efron; [3] is a readable explanation. Lunneborg [4] provides a technical but comprehensive explanation. Here resamples are drawn each of a size n, but typically are drawn with replacement. For example, in a regression analysis involving 250 participants, the first resample might be drawn such that Wendy's data on all the variables is drawn as a set five times, while Deborah's data is not drawn at all. However, in the second resample, Wendy's data might not be drawn at all, while Deborah's data might be drawn three times.

Usually when the bootstrap is invoked, thousands of resamples are drawn and analyzed. Indeed, both the jackknife and the bootstrap are called 'computationally intensive', because they usually cannot be done unless specialized software is used to execute the numerous analyses required. This software, though not part of SPSS, is widely available, especially as regards the bootstrap.

The cross-validation method is the least sophisticated of the three methods because it involves only one sample split. The problem is that for a given data set numerous splits are possible, and different splits might yield contradictory results as regards the same data.

Because both the jackknife and the bootstrap are computationally intensive, but the bootstrap has the appeal of mixing up the participants in very many ways to see if the results remain stable, researchers considering these two choices quite frequently opt for the bootstrap. Although the bootstrap sounds like a tremendous amount of work, the work is done by a modern microcomputer in seconds, and is thus quite painless.

A special challenge arises when using the bootstrap with multivariate statistics (*see* **Multivariate Analysis: Overview**). Multivariate statistics will usually yield several functions or factors in a given analysis. The orders of the factors may arbitrarily vary across resamples. This is only a logistical issue, because usually the order of a given construct is not that meaningful. But some way must be found to compare a given construct across the thousands of resamples such that apples are compared to apples [12]. Thompson [5] proposed invoking a special statistical rotation method, called Procrustean rotation, to solve this problem (*see* **Procrustes Analysis**). This solution has been generalized to bootstrap factor analysis [5], descriptive **discriminant analysis** [8], and **canonical correlation analysis** [7].

References

[1] Cohen, J. (1994). The earth is round ($p < 0.05$), *American Psychologist* **49**, 997–1003.

[2] Cumming, G. & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions, *Educational and Psychological Measurement* **61**, 532–575.

[3] Diaconis, P. & Efron, B. (1983). Computer-intensive methods in statistics, *Scientific American* **248**(5), 116–130.

[4] Lunneborg, C.E. (2000). *Data Analysis by Resampling: Concepts and Applications*, Duxbury, Pacific Grove.

[5] Thompson, B. (1988). Program FACSTRAP: a program that computes bootstrap estimates of factor structure, *Educational and Psychological Measurement* **48**, 681–686.

[6] Thompson, B. (1989). Statistical significance, result importance, and result generalizability: three noteworthy but somewhat different issues, *Measurement and Evaluation in Counseling and Development* **22**(1), 2–5.

[7] Thompson, B. (1991). Invariance of multivariate results, *Journal of Experimental Education* **59**, 367–382.

[8] Thompson, B. (1992). DISCSTRA: a computer program that computes bootstrap resampling estimates of descriptive discriminant analysis function and structure coefficients and group centroids, *Educational and Psychological Measurement* **52**, 905–911.

[9] Thompson, B. (1994). The pivotal role of replication in psychological research: empirically evaluating the replicability of sample results, *Journal of Personality* **62**, 157–176.

[10] Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: three suggested reforms, *Educational Researcher* **25**(2), 26–30.

[11] Thompson, B. (2002). What future quantitative social science research could look like: confidence intervals for effect sizes, *Educational Researcher* **31**(3), 24–31.

[12] Thompson, B. (2004). *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*, American Psychological Association, Washington.

[13] Thompson, B. (in press). Research synthesis: effect sizes, in *Complementary Methods for Research in Education*, J. Green, G. Camilli & P.B. Elmore, eds, American Educational Research Association, Washington.

BRUCE THOMPSON

Reproduced Matrix

Charles E. Lance

Volume 4, pp. 1743–1744

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Reproduced Matrix

A reproduced matrix (symbolized as $\hat{\boldsymbol{\Sigma}}$) is a matrix of correlations or covariances that is calculated from parameter estimates obtained for a path analytic (*see* **Path Analysis and Path Diagrams**), factor analytic (*see* **Factor Analysis: Confirmatory**), or latent variable **structural equation model**. Take, for example, the hypothetical path model shown in Figure 1. If we were to gather data on the X and Y variables shown in this figure, we could estimate this model's parameters in the following three equations:

$$Y_1 = \alpha_1 + \gamma_{11}X_1 + \gamma_{12}X_2 + \gamma_{13}X_3 + d_1 \quad \text{(1a)}$$

$$Y_2 = \alpha_2 + \beta_{21}Y_1 + \gamma_{23}X_3 + \gamma_{24}X_4 + d_2 \quad \text{(1b)}$$

$$Y_3 = \alpha_3 + \beta_{31}Y_1 + \beta_{32}Y_2 + d_3 \quad \text{(1c)}$$

and calculate $\hat{\boldsymbol{\Sigma}}$ in part from the estimated model parameters. $\hat{\boldsymbol{\Sigma}}$ actually consists of three conceptually distinct parts, however, correlations (or covariances) among the endogenous variables ($\hat{\boldsymbol{\Sigma}}_{YY'}$), correlations among the exogenous variables ($\hat{\boldsymbol{\Sigma}}_{XX'}$), and correlations between the exogenous and endogenous variables ($\hat{\boldsymbol{\Sigma}}_{XY} = \hat{\boldsymbol{\Sigma}}'_{YX}$) so that overall

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{YY'} & \hat{\boldsymbol{\Sigma}}_{YX} \\ \hat{\boldsymbol{\Sigma}}_{XY} & \hat{\boldsymbol{\Sigma}}_{XX'} \end{bmatrix}. \quad (2)$$

As is shown in Figure 1, the exogenous variables are assumed to be correlated (as is usually the case) so that $\hat{\boldsymbol{\Sigma}}_{XX'}$ is given from the data. However, correlations between the Xs and Ys in $\hat{\boldsymbol{\Sigma}}_{XY}$ result from various functions of model parameters, including direct effects of Xs on Ys (e.g., the effect of X_1 on $Y_1 - \gamma_{11}$), indirect effects of Xs on Ys (e.g., the effect of X_1 on Y_2 through $Y_1 - \gamma_{11}\beta_{21}$), and/or common causal effects (e.g., X_3 affects both Y_1 and $Y_2 - \gamma_{13}\gamma_{23}$). Similarly, correlations among the Ys in $\hat{\boldsymbol{\Sigma}}_{YY'}$ also arise from direct effects (e.g., the effect of Y_1 on $Y_2 - \beta_{21}$), indirect effects (the effect of Y_1 on Y_3 through $Y_2 - \beta_{21}\beta_{32}$), and common causal effects (Y_1 affects both Y_2 and $Y_3 - \beta_{21}\beta_{31}$). So, once the model's parameters in (1a) through (1c) are estimated from data, they can be used to calculate or *reproduce* the correlations among the variables in the model by using the parameter estimates to solve the equations reflecting the decomposition of effects specified by the model.

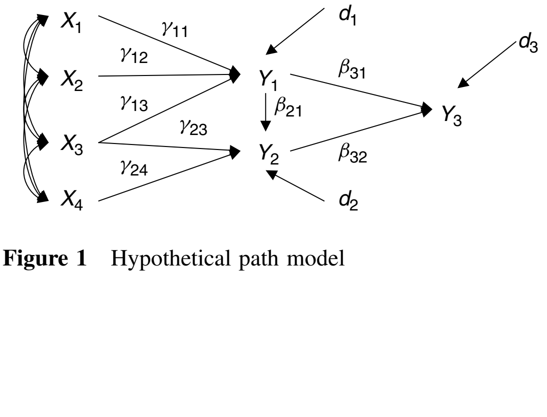


Figure 1 Hypothetical path model

An important question is whether the *reproduced* correlations equal (or approximate) the *observed* correlations calculated directly from data. The reproduced matrix will equal (or approximate) the observed correlations if (a) the model being estimated is correct, or (b) as many model parameters are estimated as there are elements in the observed correlation matrix. However, neither of these will generally be the case. First, it is widely accepted that models that are tested in the behavioral sciences are rarely 'true' (see [2]). Second, most models in the behavioral sciences estimate fewer (and often many fewer) parameters than elements in the matrix of correlations among variables in the model, and this is true of the hypothetical model in Figure 1. This can be seen by rewriting (1a) through (1c) as follows:

$$\begin{aligned} Y_1 &= \alpha_1 + \gamma_{11}X_1 + \gamma_{12}X_2 \\ &\quad + \gamma_{13}X_3 + \mathbf{0X}_4 + d_1 \qquad \text{(3a)} \\ Y_2 &= \alpha_2 + \beta_{21}Y_1 + \mathbf{0X}_1 + \mathbf{0X}_2 \\ &\quad + \gamma_{23}X_3 + \gamma_{24}X_4 + d_2 \qquad \text{(3b)} \\ Y_3 &= \alpha_3 + \beta_{31}Y_1 + \beta_{32}Y_2 \\ &\quad + \mathbf{0X}_1 + \mathbf{0X}_2 + \mathbf{0X}_3 + \mathbf{0X}_4 + d_3 \qquad \text{(3c)} \end{aligned}$$

where the bold elements represent 'zero-effect' hypotheses, that is, hypotheses that certain effects that *could be* estimated within a particular model are actually zero. If a model contains no zero-effect hypotheses, then as many parameters are estimated as there are elements in the observed correlation matrix and the reproduced correlation matrix will equal the observed correlation matrix by tautology.[1] For models in which there are one or more 'zero restrictions' as in (3a) through (3c), the reproduced correlation matrix will not necessarily equal the observed correlation matrix and, in fact, it most likely will not. This is because one or more of the zero restrictions may be incorrect (i.e., the true effect is actually nonzero) or the restriction may hold only approximately. There are a number of ways that the plausibility of zero-effect hypotheses can be tested (see [1]), but the most popular, omnibus statistical test is the χ^2 statistic based on the maximum likelihood fit function ($\chi^2 = (n-1)F_{\mathrm{ML}}$) where $n =$ sample size and

$$F_{\mathrm{ML}} = \log|\hat{\boldsymbol{\Sigma}}| + \mathrm{tr}(\boldsymbol{S}\,\hat{\boldsymbol{\Sigma}}^{-1}) - \log|\boldsymbol{S}| - (p+q), \tag{4}$$

where log refers to the natural logarithm, $\boldsymbol{S}$ refers to the observed or sample correlation (or covariance) matrix, $|\boldsymbol{W}|$ and $\mathrm{tr}(\boldsymbol{W})$ denote the determinant and trace of matrix $\boldsymbol{W}$, respectively, and p and q refer to the number of Y and X variables, respectively. If the model's zero restrictions are (at least approximately) tenable, then the discrepancy $\log|\hat{\boldsymbol{\Sigma}}| - \log|\boldsymbol{S}|$ from (4) will be small as will the discrepancy $\mathrm{tr}(\boldsymbol{S}\,\hat{\boldsymbol{\Sigma}}^{-1}) - (p+q)$ so that the resulting χ^2 will also tend to be 'small.' If one or more zero restrictions do *not* hold, the discrepancy between $\hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{S}$ increases, along with the F_{ML} discrepancy function and the χ^2 statistic. As such, the χ^2 statistic is a badness-of-fit index that is inexorably tied to the discrepancy between the observed and reproduced correlation (or covariance) matrices and whose degrees of freedom is closely related to the number of zero restrictions that are imposed on the model tested. Large and statistically significant χ^2 statistics indicate that one or more zero restrictions imposed in a path, factor analytic, or latent variable structural equation model are implausible.

References

[1] James, L.R., Mulaik, S.A. & Brett, J.M. (1982). *Causal Analysis: Models, Assumptions, and Data*, Sage Publications, Beverly Hills.

[2] MacCallum, R.C. (2003). Working with imperfect models, *Multivariate Behavioral Research* **38**, 113–139.

Note

1. The issue of model identification is actually much more complex an issue than can be treated here. Our humble goal is merely to introduce some notions related to model identification as they help understand the important role of the reproduced matrix in assessing model fit.

(*See also* **Factor Analysis: Confirmatory**)

CHARLES E. LANCE

Resentful Demoralization

Patrick Onghena

Volume 4, pp. 1744–1746

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Resentful Demoralization

Resentful demoralization is a validity threat in intervention studies (*see* **Clinical Trials and Intervention Studies**) in which comparison groups not obtaining a desirable treatment are aware of this inequity or find out about this during the study, become discouraged or retaliatory, and as a result perform worse on the outcome measures [4, 5]. The prototypical setting for this threat to operate is a randomized two-group intervention study with a treatment group and a no-treatment control group in which treatment allocation is obtrusive and participants in the control group get deprived of certain facilities offered to the participants in the treatment group. As a consequence, people in the control group could feel neglected and behave differently than they would have behaved if they had not known about the favorable intervention in the treatment group. The most likely effect of resentful demoralization is that the observed treatment effect gets inflated and that the intervention program looks more effective than it actually is.

As a social interaction threat to validity, this confound has the same basis as **compensatory rivalry** (i.e., perceived inequality between the comparison groups), but the emotional and behavioral reaction that is involved, is quite the opposite. Whereas in case of compensatory rivalry the participants in the deprived group(s) become competitive, in case of resentful demoralization they get dejected or vindictive and bring about inferior performance. For both threats, the treatment effects become confounded by the differential motivation to perform, but in the case of compensatory rivalry, the treatment effect probably gets underestimated, while in the case of resentful demoralization it gets overestimated because of the confound.

Campbell and Stanley [3] already hinted at this problem when they warned us of the possibility of 'reactive arrangements' in experimental research, but the term 'resentful demoralization' made its first appearance in Cook and Campbell's list of **internal validity** threats [4]. According to Shadish, Cook, and Campbell [5], resentful demoralization is even so closely associated with the treatment construct itself that it should be included as part of the treatment construct description. Therefore, in their revised list Shadish et al. [5] classified this confound among the threats to construct validity. According to these authors, internal validity threats are disturbing factors that can occur even without a treatment, while this is obviously not the case with resentful demoralization: 'The problem of whether these threats should count under internal or construct validity hinges on exactly what kinds of confounds they are. Internal validity confounds are forces that could have occurred in the absence of the treatment and could have caused some or all of the outcome observed' (p. 95).

Resentful demoralization is clearly related to compensatory rivalry, but is also related to other construct and internal validity threats as well. If participants have the impression that they are treated unfairly, then they could try to get the beneficial treatment anyhow, resulting in **compensatory equalization**. Or if the demoralization is vast, then the participants of the control group might decide to stop participation altogether, resulting in differential **attrition**. Furthermore, notice that, although the prototypical setting described above is a randomized trial in which treatment allocation is conspicuous, also a **quasi-experiment** working with eligible participants for one of the treatment arms is particularly susceptible to this kind of bias.

Resentful demoralization can be avoided or minimized by isolating the comparison groups in time (e.g., using a waiting list condition) or space (e.g., using geographically remote groups), or making the participants unaware of the intervention being applied (e.g., using blinding). If no such design control is possible, poststudy assessment strategies could be useful, for instance, by asking the participants in a debriefing whether they felt uncomfortable being assigned to the control condition, and by relating these responses to the outcome.

A good example of the latter strategy is given in the **prospective study** by Berglund et al. [1, 2] in which 98 of the 199 cancer patients who wanted to participate in the study were randomly assigned to the structured rehabilitation program 'Starting Again', and the other patients were assigned to the control condition. This program consisted of 11 two-hour sessions focusing on physical training, information, and training of coping skills for cancer patients, and although the results were generally positive, the researchers were suspicious about the possibility of resentful demoralization because the patients were aware of the treatment assignment procedure by the informed consent. However, their poststudy analysis showed that although a few patients may have been resentful, this did not significantly affect the outcome variables. Furthermore, they were able to compare the 101 patients assigned to the control condition with 73 patients who did not wish to participate in the study in the first place, and found no negative effects resulting from being randomized to the control condition.

As a conclusion, if assignment-related motivational effects, such as resentful demoralization, in some of the comparison groups are suspected, then researchers should be cautious about any causal generalization of the treatment. However, in intervention studies that have to deal with these potentially confounding motivational effects, the credibility of the results may be strengthened by design control, post-study assessment, and qualified interpretations.

References

[1] Berglund, G., Bolund, C., Gustafsson, U.-L. & Sjödén, P.-O. (1994). One-year follow-up of the 'Starting Again' group rehabilitation program for cancer patients, *European Journal of Cancer* **30A**, 1744–1751.

[2] Berglund, G., Bolund, C., Gustafsson, U.-L. & Sjödén, P.-O. (1997). Is the wish to participate in a cancer rehabilitation program an indicator of the need? Comparisons of participants and non-participants in a randomized study. *Psycho-Oncology* **6**, 35–46.

[3] Campbell, D.T. & Stanley, J.C. (1963). *Experimental and Quasi-Experimental Designs for Research*, Rand-McNally, Chicago.

[4] Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Rand-McNally, Chicago.

[5] Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston.

Patrick Onghena

Residual Plot

Jeremy Miles

Volume 4, pp. 1746–1750

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Residual Plot

The **residuals** in a regression analysis (*see* **Multiple Linear Regression**) (or related, e.g., **analysis of variance**) are the difference between the scores on the outcome variable predicted from the regression equation and the observed scores. The distributional (and other) assumptions that are made when using regression analysis relate to the residuals, rather than the outcome variable. Residuals can be transformed in different ways to improve their interpretability (*see* **Residuals**). Here, when we refer to residuals, we generally mean studentized-deleted residuals (also known as externally studentized residuals), unless explicitly stated otherwise.

Draper and Smith [2] suggest that residuals should be plotted

- overall, in a **histogram** or **boxplot**;
- against the time in which the data were collected, in a **scatterplot**;
- against the predicted values of the outcome variable;
- against the independent variable(s);
- in any other way that seems to be sensible.

Distribution of Residuals

The distribution of the residuals should always be examined to check for normality and for **outliers**. If outliers are detected, these may be further examined, and tested for statistical significance. If the data are nonnormal, this may suggest that a transformation of the data would be appropriate (although it is possible for a normally distributed outcome variable to give rise to nonnormal residuals).

The following data were taken from a study examining 51 women who were undergoing a biopsy for breast cancer. Anxiety, depression, and self-esteem were measured one week before undergoing the biopsy for breast cancer, and on the day that they underwent the biopsy. The analysis presented looks at anxiety on the day of the biopsy as an outcome, and predictors are anxiety, depression, and self-esteem, measured one week before the biopsy. The predictors had a large and highly significant effect on the outcome variable ($R^2 = 0.82$, $p < 0.001$). The standardized coefficients were -0.17 ($p = 0.039$) for self-esteem, 0.192 ($p = 0.030$) for depression, and 0.644 ($p < 0.001$) for anxiety.

The residuals can be examined using a histogram, **probability plots** or boxplots – each gives the same information, but presents it in a slightly different format. Although any of the types of residual can be used, the most useful is probably the studentized-deleted (also known as the externally studentized, or the jack-knifed) residual, because this is the most easily interpreted.

The histogram (Figure 1) shows that distribution is approximately symmetrical, but the tails of the distribution may be heavier than we would expect from a normal distribution – we might consider them to be outliers.

The probability plot (Figure 2) similarly shows that the distribution is symmetrical and deviates away from the normal distribution – probability plots are poor at detecting outlying cases, and so it is not clear that there are potential outliers here. Finally, Figure 3 shows the box plots for three types of residuals, the standardized, studentized (also known as the internally studentized) and the studentized-deleted. The distribution of the studentized-deleted residuals seems to have wider tails than the other distributions. The boxplot is also useful for identifying outliers – there are clearly three points that might be considered to be outliers. The largest one of these has a studentized-deleted residual of 3.77. The Bonferroni-corrected probability of finding a studentized residual with an absolute value of 3.77 is equal to 0.07 and close to the conventional cutoff of 0.05. However, we should not discard data without good reason, and taking a conservative approach would be better in this

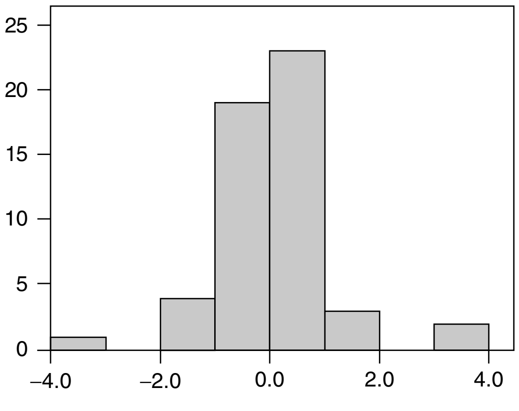


Figure 1 Histogram of studentized-deleted residuals

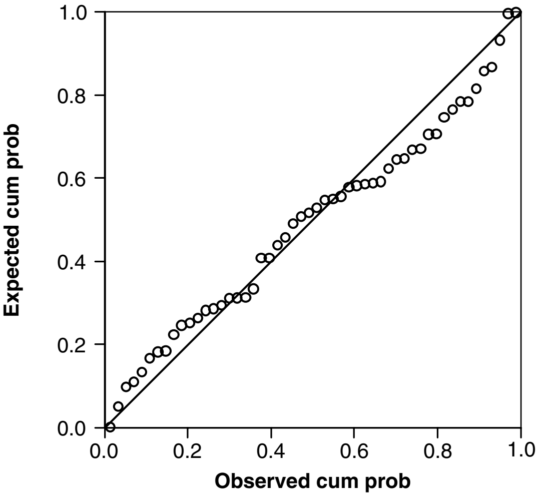


Figure 2 Normal probability plot of studentized-deleted residuals

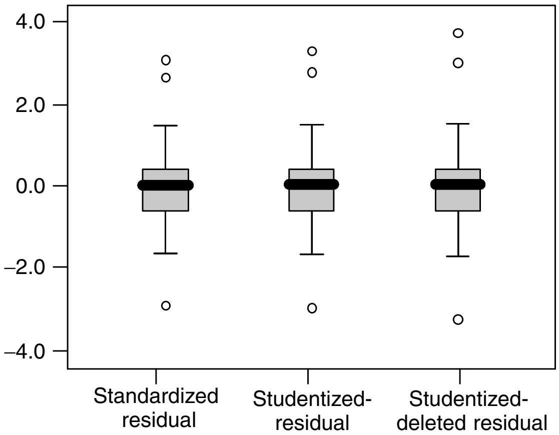


Figure 3 Boxplots of standardized residual, studentized residual, and studentized-deleted residual. Note that the studentized-deleted residual has the larger variance, and the residuals appear to be more extreme

case, so we shall leave the data point in the file. (It is still probably worth treating the case with some suspicion, and ensuring that there is nothing else that leads us to have misgivings about it.)

Residuals Against Time

Plotting residuals against time or participant number in a scatterplot, where participant number reveals the order in which the data were collected, is the next task (*see* **Index Plots**). This serves two purposes: first, it reassures us that there is no linear relation with time, and second, it can be used to detect violations of the independence assumption.

If the data were collected over a period (as they almost certainly were), then it is possible that something has changed over that period which would have had an effect on the data. This may need to be taken into account or may invalidate the research. There are a number of ways in which time might have an effect. If the data were collected over the course of a day, it is possible that the conditions changed over that day – the experimenter may have become more tired, a piece of equipment may have gradually lost its calibration, the room temperature may have gradually increased, for example. In physiological studies, it is possible that the substance under examination may have decayed over time – if the control group is analyzed first, this might lead to a spurious result if the effect of time is not taken into account.

Figure 4 shows the scatterplot of the studentized residual against the participant number. Note that here I have added two reference lines, at ±1.96, indicating that 95% of the points should lie between these lines. Figure 4 also contains a loess regression line (*see* **Scatterplot Smoothers**) which can help to identify any trend in the data.

Examining Figure 4 suggests two possible effects that may have occurred in these data. First, there seems to be a trend in that the residuals are increasing in value over time. In this case, where it appears that

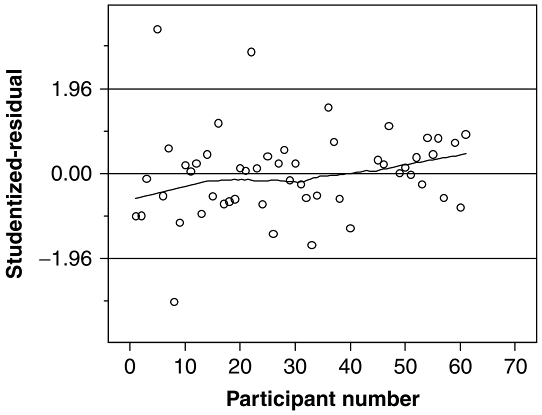


Figure 4 Scatterplot of studentized residuals plotted against participant number with loess regression line

there is a monotonic relationship, we can examine the correlation between the residuals and the participant number. This result is not statistically significant ($r = 0.14$, $p = 0.33$), and therefore does not support the hypothesis.

The second possibility shown by this graph is that the variance is nonconstant at different levels of the predictor variable. This effect is known as heteroscedasticity (*see* **Heteroscedasticity and Complex Variation**). One possible cause of heteroscedasticity of these residuals is the reliability changing over time – in the case that may occur here, it may be that the measurements of the outcome variable are increasing in reliability over time, and that there is more error at the start of the study than at the end. It could be that the increase in measurement reliability is caused by the experimenter becoming more familiar with the procedure or with a piece of equipment. We can test the heteroscedasticity assumption using White's test, and we find that $\chi^2 = 3.16$, $df = 2$, $p = 0.21$; again, there is no statistical evidence to support the hypothesis that the variance is nonconstant.

The second use of this type of residual plot is that it can help us detect violations of the independence assumption in **Generalized Linear Model** analyses.

It is assumed in regression that residuals are independent – that is, knowing the value of one residual should not help us predict the value of the next residual. For example, if we tested our participants in groups or batches, we might find that the groups were in some way similar to each other – if our outcome was alcohol consumption, it is feasible that there might be a conformity effect, and in some groups everyone drank a lot, and in others everyone drank little (it is also possible that the opposite effect might have occurred – one person drinking heavily may have stopped the others from drinking as much). Similarly, this problem may occur if we select participants who are in some way in naturally occurring groups, and where these groupings might affect the outcome variable – common examples of this problem include children in classroom groups, patients treated by the same general practitioner (GP) or hospital and students living in the same hall of residence. Figure 5 shows a scatterplot of participant number against the residual for 35 participants. The chart shows that there are five groups of participants, the first with positive residuals, then a group with negative residuals, and so on. The loess line has been added, to emphasize this effect. This type of effect is (or should be) relatively rare – researchers are increasingly aware of the problems of clustered data, and either try to avoid the clustering or alternatively take it into account (see [3] and [4]).

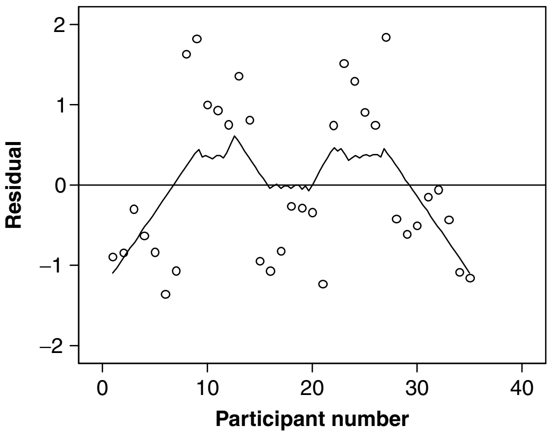


Figure 5 Plot of participant number against residual, where there are clustered data with loess regression line

The second way in which nonindependence may occur is if the previous measurement is in some way predictive of the next measurement (or any future measurement.) The most common example of this effect is in time-series designs, where instead of measuring multiple people on one occasion, one person (or a small number of people) is measured on multiple occasions. If we are measuring reaction time, it is possible that this will slow down towards the end of a long series of trials. If we are measuring mood over the course of a number of days, my mood yesterday is likely to be a better predictor of my mood today than it is of my mood next Tuesday.

If graphical examination of the data reveals that there is a clustering effect that was not anticipated, then either the Durbin–Watson test or a runs test should be carried out to examine whether the effect is statistically significant.

Residuals Plotted Against Predicted Values

Perhaps the most important plot is that of the residuals plotted against the fitted values, shown in

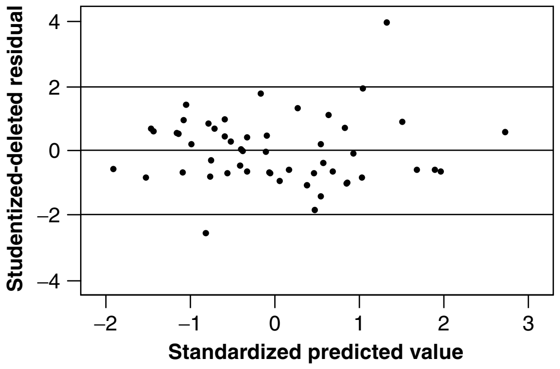


Figure 6 Standardized predicted values plotted against studentized-deleted residuals. Additional horizontal lines have been drawn at ±1.96 on the y-axis

Figure 6. This can be used to reveal several different potential problems: outliers, nonlinearity, heteroscedasticity, and nonnormality.

Outliers are revealed in this graph as cases that are far from the other cases – the graph has additional gridlines extending from the y-axis and ±1.96. We would expect 95% of cases to lie within these lines. The point at the top of the graph, with a residual approaching four is particularly problematic.

The points in the graph should form a straight line across the chart – there should be no tendency of the residuals to be above or below zero at any point. Figure 7 shows the general shape that is indicative of a nonlinear relationship in the data, and hence that the violation of linearity has been violated.

The same plot allows us to examine the assumption of constant variance across the range of predicted values. Here we would be looking for a constant spread of the residuals around the center line. Figure 8 shows the shape of the plot that would be expected if the nonconstant variance assumption were to be violated.

Finally, something that is occasionally disturbing to data analysts is the presence of stripes, which make the residual plot look rather different from the examples generally presented in textbooks. Stripes occur because of the measurement properties of the variables – the measures are treated as continuous, but are actually discrete. In Figure 9, the outcome variable is a test with seven items, which are scored as correct or incorrect. Each individual therefore scores a whole number, from a minimum of zero to a maximum of seven.

Cook and Weisberg [1] provide an excellent guide to graphics in regression; there is a computer program (ARC) to accompany the book.

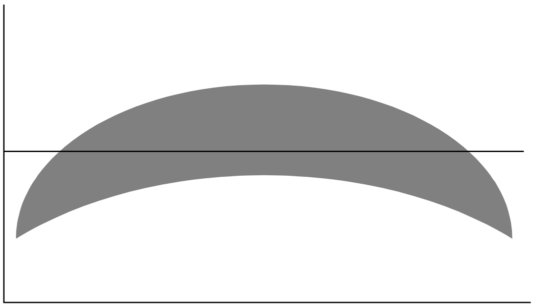


Figure 7 Scatterplot of residuals against predicted values showing nonlinearity

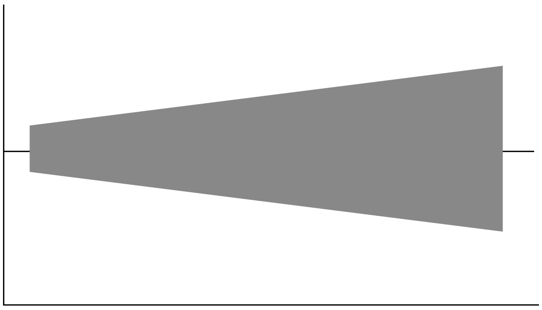


Figure 8 Scatterplot of residuals against predicted values, showing nonconstant variance

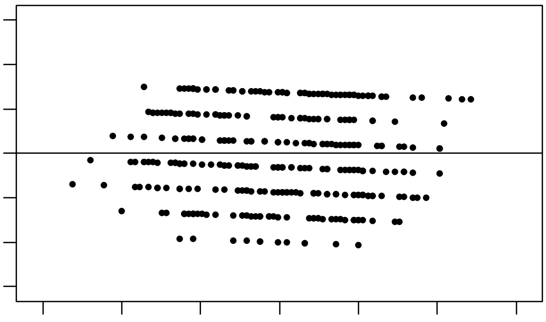


Figure 9 Residual plot showing the presence of stripes

References

[1] Cook, R.D. & Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*, John Wiley & Sons, New York.

[2] Draper, N. & Smith, H. (1981). *Applied Regression Analysis*, 2nd Edition, John Wiley & Sons, New York.

[3] Hox, J. (2002). *Multilevel Analysis: Techniques and Applications*, Lawrence Erlbaum Associates, Mahwah.

[4] Leyland, A. & Goldstein, H. (2001). *Multilevel Modelling of Health Statistics*, Wiley Series in Probability and Statistics, Wiley, Chichester.

JEREMY MILES

Residuals

JEREMY MILES

Volume 4, pp. 1750–1754

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Residuals

It is (too) commonly believed that distributional assumptions for many statistical tests are made on the variables – this is not the case: for most statistical techniques, the assumptions that are made are dependent on the distribution of the residuals.

Consider a group of 10 individuals, of different ages, who have been presented with a quiz that measured their knowledge of a range of areas. The age of each person and the number of items that they recalled are shown in the first two columns of Table 1. If we fit a least squares linear **regression model** to these data, we find that

$$\hat{x} = 0.223 \times a + 12.52, \tag{1}$$

where x is the score and a is the age of the individual. The hat on the x indicates that this is a predicted score, not the actual score. The 95% **confidence interval** for the regression coefficient is 0.063 to 0.383, and the associated significance is 0.012.

In other words, we would expect that people who are one year older will score 0.223 points higher on the quiz.

If we wished to use the actual score for each individual, we need to add a term to take account of the difference between the predicted score and the score that each individual achieved – we refer to this as *error* (e_i), and these are the residuals.

$$x_i = 0.223 \times a + 12.52 + e_i \tag{2}$$

Table 1 Data, predicted values, and residuals

Age (a)	Number of items correct (x)	Predicted score ($\hat{x}$)	Residual ($x - \hat{x} = e$)
20	15	16.98	−1.98
25	21	18.10	2.90
30	19	19.21	−0.21
35	18	20.33	−2.33
40	21	21.44	−0.44
45	25	22.56	2.44
50	22	23.67	−1.67
55	26	24.79	1.21
60	31	25.90	5.10
65	22	27.02	−5.02

(The i subscript has been added to index each individual.) Rearranging this equation gives:

$$e_i = x_i - (0.223 \times a + 12.52) \tag{3}$$

However, note that the part of the equation $0.223 \times a + 12.52$ is equal to $\hat{x}$, therefore we can substitute this into the equation, giving:

$$e_i = x_i - \hat{x}_i \tag{4}$$

These values are shown in Table 1.

In regression and related analyses (**analysis of variance**, *t* Test, **analysis of covariance**, etc.), we assume that the residuals are sampled from a normally distributed population with a mean equal to zero.

To illustrate the difference between the distribution of the variables and the distribution of the residuals, consider the simple example of an independent samples *t* Test. Figure 1 shows that the distribution of the outcome variable appears to be positively skewed. Figure 2 shows the distribution of two different groups – here it can be seen that the distribution in Figure 1 is actually comprised of two different distributions, one for each of two groups, and that the group with the higher mean has a smaller sample size. The predicted values in this case are simply the means for each of the two groups, and so the residuals are the difference between each value and the mean for that group. Figure 3 shows the distribution of the residuals – this is clearly from a normal distribution.

In the case of the two group *t* Test, it was easy to see how the distribution of the variable could make us believe that we may have had a problem with our variable – in a more complex analysis, for example, if we were to carry out an analysis of covariance, in an experiment with a 2 × 2 design and a covariate,

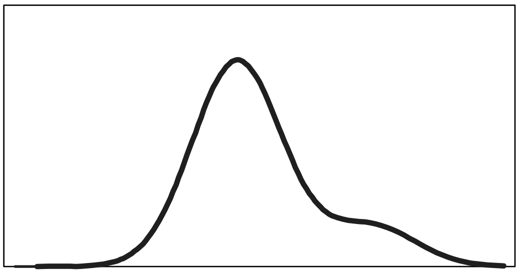


Figure 1 Histogram that shows distribution of outcome variable

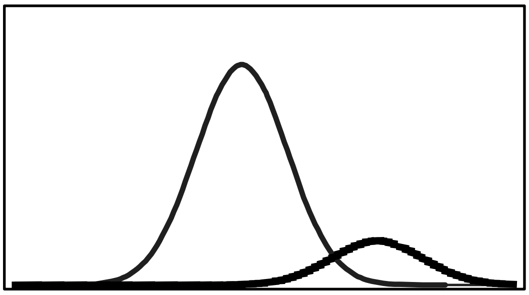


Figure 2 Histogram that shows distribution of two groups

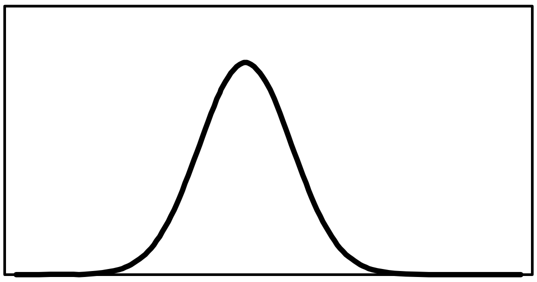


Figure 3 Histogram of residuals

it could be difficult to relate the distribution of the residuals to the distribution of the outcome variable.

One particular reason to examine residuals is to ensure that there are no **outliers** or influential cases. An outlier is a case with an unusual value, or combination of values, in a dataset – we are concerned about outliers because that individual will have undue influence on our parameter estimates.

Univariate outliers can be detected through the usual data cleaning procedures; however, there are other types of outliers, termed **multivariate outliers**, which cannot be detected through such methods (*see* **Outlier Detection**). Consider the dataset shown in Table 1 – but with one additional subject, a 15-year-old, who scores 33 on the test. The 15-year-old is younger than the other participants, but not excessively younger, and 33 is the highest score on the test, but not excessively high (see Table 2). From the **scatterplot** in Figure 4, we see that this individual (indicated on the chart with the solid black spot) does seem to lie outside of the main set of points.

If we reanalyze our regression with this additional case, we find that the regression coefficient has dropped to 0.067 (95% CI −0.173, +0.308; $p = 0.542$). This individual has obviously had considerable influence on our analysis and has caused us to dramatically alter our conclusions. Table 2 shows the new predicted scores and residuals. Simply scanning the residuals by eye shows that the first case does have a residual that is higher than the others. We could also view these data using a **histogram** or a **boxplot**, such as in Figure 5.

Table 2 Data including outlier observation, new predicted values and residuals

Age (a)	Number of items correct (x)	Predicted score ($\hat{x}$)	Residual ($x - \hat{x} = e$)
15	33	21.32	11.68
20	15	21.65	−6.65
25	21	21.99	−0.99
30	19	22.33	−3.33
35	18	22.66	−4.66
40	21	23.00	−2.00
45	25	23.34	1.66
50	22	23.67	−1.67
55	26	24.01	1.99
60	31	24.35	6.65
65	22	24.68	−2.68

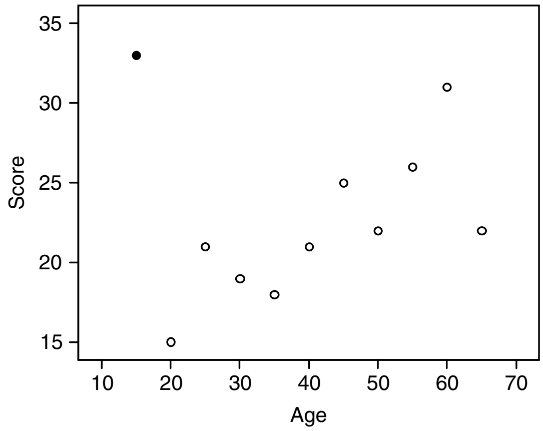


Figure 4 Scatterplot of score versus age. Solid black spot indicates a potential outlier

Transformation of Residuals

Standardized Residuals. Residuals in their raw form are difficult to interpret, as the scale is that of the outcome variable – they are not in a common metric that we can interpret. Residuals can be transformed in a number of different ways to a metric that we

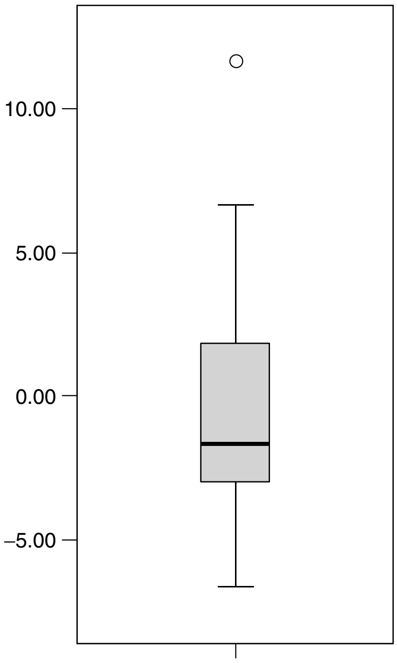


Figure 5 Boxplot of residuals, showing outlier

understand better, for example, to a normal or a t distribution. We should note from the outset that there are differences among the names given for these different transformations by different books and computer programs.

An additional problem is the lack of common names for the different types of residuals, which I shall point out along the way. (A further warning is required here – computer programs are also inconsistent with their naming – different programs will use different names for the same thing, and the same name for different things. It is important to know exactly what the program is doing when using any of these statistics.)

The first type of transformation is to divide the residuals by their standard deviation, to produce residuals with a standard normal distribution (mean = 0, standard deviation = 1) in a large sample, and t distribution (with df equal to $n - k - 2$) in a smaller sample. These are also referred to as *unit normal deviate* residuals by Draper and Smith [2].

The standardized residual (e^s) for each case is given by:

$$e_i^s = \frac{e_i}{s_e} \tag{5}$$

where e_i is the residual for each case, and s_e is the standard deviation of the residuals. (Also be warned that some references, for example, Fox [3], use the term 'standardized residual' to refer to what we shall call 'studentized residuals'.) Note that the standard deviation of the residuals is given by:

$$s_e = \sqrt{\frac{\Sigma e^2}{n - k - 1}} \tag{6}$$

We must incorporate k, the number of predictor variables, into the equation as well as the sample size. (Although in large samples, this will have a negligible effect on the calculation; given that we are almost certainly using a computer, we might as well do it properly.)

Because standardized residuals follow an approximately standard normal distribution, we can make statements about the likelihood of different values arising. We would expect approximately 1 case in 20 to have an absolute value greater than 2, 1 case in 100 to fall outside an absolute value of 2.6, and 1 in 1000 to fall beyond 3.1. If, therefore, we have a case with an absolute standardized residual of 3, in a sample size of 50, we should consider looking at that case.

In the data of people's ages and their scores, the sum of squared residuals is equal to 279.5, and the standard deviation is therefore equal to the square root of $279.5/(11 - 1 - 1)$, which gives 5.57. Dividing each of the raw residuals by 5.57 will therefore give the standardized residuals (see column 5 of Table 3).

Studentized Residuals. There is a problem with the use of standardized residuals; some [4] have argued that they should not be called standardized residuals at all. The problem that we need to address is that the variances of the residuals are not equal, as we have considered them to be. The variance of the residual is dependent on the scores on the predictor variables. In the case of analyses with one predictor variable, the variance of the residual depends on the distance of the predictor variable from its mean – extreme scores on the predictor variable are associated with lower variance of the residuals. In the multiple predictor case, the distance from the centroid of all predictor variables is used to ascertain the variance of the residuals.

The distance from the mean or from the centroid of all predictor variables is called the *leverage*, (*see*

Table 3 Predicted scores, residuals, and hat (leverage) values for original data

Age (a)	Number of items correct (x)	Predicted score ($\hat{x}$)	Residual ($x - \hat{x} = e$)	Standardized residual	Leverage (hat value)	Studentized residual	Studentized deleted residual
20	15	16.98	−1.98	2.10	0.32	2.54	3.00
25	21	18.10	2.90	−1.19	0.25	−1.37	−1.28
30	19	19.21	−0.21	−0.18	0.18	−0.20	−0.19
35	18	20.33	−2.33	−0.60	0.14	−0.64	−0.62
40	21	21.44	−0.44	−0.84	0.11	−0.88	−0.85
45	25	22.56	2.44	−0.36	0.10	−0.38	−0.37
50	22	23.67	−1.67	0.30	0.11	0.31	0.32
55	26	24.79	1.21	−0.30	0.14	−0.32	−0.32
60	31	25.90	5.10	0.36	0.18	0.39	0.40
65	22	27.02	−5.02	1.19	0.25	1.37	1.49

Leverage Plot) or the *hat* value, or h, and is the diagonal element of the hat matrix. The minimum value of h_i is $1/n$, the maximum is 1. As might be expected by now, however, there is not complete agreement about the nomenclature – the leverage has a second form, the centred leverage, h^*, which has a minimum value of zero, and a maximum of $(N - 1)/N$.

Most computer programs use the leverage (h); SPSS, however, uses the centred leverage (h^*) but refers to it as 'leverage' – the only clue is that the help file says that the minimum value is zero and the maximum is $(N - 1)/N$.

When we have the leverage values, these can be used to correct the estimate of the standard deviation of the residuals and calculate the studentized residual (e'), using the following equation.

$$e'_i = \frac{e_i}{s_e\sqrt{1 - h_i}} \tag{7}$$

And once again, we note the different titles that are given to these – Cohen et al. [1] call these *internally studentized residuals*, Fox [3] calls these *standardized residuals*, and Draper and Smith [2], to ensure that there really is no confusion, call them the $e_i/\{(1 - \mathbf{R}_{ii})s^2\}^{1/2}$ form of the residuals (noting of course, that they choose to refer to the hat matrix as **R** rather than **H**).

The studentized residuals for our example are shown in Table 3.

Studentized Deleted Residuals. We have dealt with one problem with the residuals, but we have yet another. The residuals do not quite follow a t distribution – if we wanted to use the residuals to ascertain the probability of such a value arising, we need to calculate the studentized deleted residuals.

The standardized residual and the studentized residual are calculated on the basis of the standard deviation of the residuals. However, the residual has an influence on the standard deviation of the residuals, in two ways. First, where the residual is large, the standard deviation will be larger, because that residual will be used in the calculation of the standard deviation – this will have the effect of shrinking the residual. Second, where the case has had an influence on the regression estimates (as is likely if it is an outlier), the regression line will be drawn toward the case, and so the size of the residual will again shrink.

The solution is to remove the case from the dataset, run the regression analysis again, estimate the regression line and the standard deviation of the residuals, then replace the case, and calculate the standardized and studentized residuals. In fact, this can be found using the equation given earlier for studentized residual e' except that s_e is calculated with the offending case omitted. The deleted studentized residuals for our example can be seen in Table 3.

As would be expected by now, the deleted studentized residual is also known by other names, principally as the *externally studentized residual*, but also as the *jackknifed residual.*

The deleted studentized residual is distributed as a t distribution, with $n - k - 2$ degrees of freedom. The null hypothesis that we can test using this value is that a case with a studentized residual that large should have arisen by chance. We can examine the largest absolute studentized deleted residual and calculate the associated probability (it is unlikely that tables will be of much use here, you will need to use a computer). The probability associated with a value of t of 3.00 is 0.017 – we would therefore consider that there was support for the hypothesis that the case was not sampled from the same population as the other residuals and should consider this case to be an outlier. Using this criterion, and a 5% cut off, we will find that 5% of our cases are outliers – obviously not a useful way to proceed. The alternative is to correct this probability by using Bonferroni correction – that is, we either multiply the probability associated with the case by N, or alternatively (and equivalently) we divide the cutoff that we are using (typically 0.05) by N. Taking the first approach, multiplying the probability by N (which is 11), we find the Bonferroni corrected probability to be $0.017 \times 11 = 0.19$. This is above the cutoff of 0.05, and therefore we do not have sufficient evidence against the null hypothesis that the case is sampled from the same population. (Equivalently, we could divide our cutoff of 0.05 by 11, to give a new value of 0.0045; by this criteria also, we do not have evidence against the null hypothesis.) This analysis has, of course, been affected by the smallness of the size of the sample, which was purely for illustrative purposes.

References

[1] Cohen, J., Cohen, P., West, S.G. & Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd Edition, Lawrence Erlbaum, Mahwah.

[2] Draper, N. & Smith, H. (1981). *Applied Regression Analysis*, 2nd Edition, John Wiley & Sons, New York.

[3] Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*, Sage Publications, Thousand Oaks.

[4] Ryan, T.P. (1997). *Modern Regression Methods*, John Wiley & Sons, New York.

JEREMY MILES

Residuals in Structural Equation, Factor Analysis, and Path Analysis Models

TENKO RAYKOV

Volume 4, pp. 1754–1756

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Residuals in Structural Equation, Factor Analysis, and Path Analysis Models

Several types of residuals have been considered for **structural equation**, **factor analysis**, and **path analysis models**. They represent different aspects of discrepancies between model and data.

The traditional residuals, which can be referred to as *covariance structure residuals* (CSRs), are defined as element-wise differences between the empirical covariance matrix $\mathbf{S}$ for a given set of observed variables, $Z_1, Z_2, \ldots, Z_k$, and the implied (reproduced) covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ by a fitted model, M, where $\boldsymbol{\theta}$ is the vector of model parameters ($k > 1$). Denoting the CSRs by $\mathbf{r}_{ij}$ in an empirical application they are obtained as $\mathbf{r}_{ij} = \mathbf{s}_{ij} - \boldsymbol{\sigma}_{ij}(\hat{\boldsymbol{\theta}})$, where $\mathbf{s}_{ij}$ and $\boldsymbol{\sigma}_{ij}(\hat{\boldsymbol{\theta}})$ symbolize the elements in ith row and jth column in $\mathbf{S}$ and $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$, respectively, and $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ designates the model-implied covariance matrix at the parameter estimator ($i, j = 1, \ldots, k$). The CSRs play an important role in evaluating the overall fit of the model. In particular, the chi-square goodness-of-fit test statistic (*see* **Goodness of Fit for Categorical Variables**) when testing adequacy of M – that is, the null hypothesis that there exists a point $\boldsymbol{\theta}^*$ in the parameter space of M, such that $\boldsymbol{\Sigma}(\boldsymbol{\theta}^*) = \boldsymbol{\Sigma}^*$, where $\boldsymbol{\Sigma}^*$ is the population covariance matrix of $Z_1, Z_2, \ldots, Z_k$ – can be viewed, for practical purposes, as a weighted sum of the CSRs. Specifically, when **maximum likelihood estimation** (ML) method is used, the fit function $F_{\mathrm{ML}} = -\ln|\mathbf{S}(\boldsymbol{\Sigma}(\boldsymbol{\theta}))^{-1}| + tr(\mathbf{S}(\boldsymbol{\Sigma}(\boldsymbol{\theta}))^{-1}) - k$ (e.g., [2]; $|.|$ designates determinant and $tr(.)$ matrix trace) can be approximated by the sum of squared CSRs weighted by corresponding elements of the inverse implied covariance matrix, $[\boldsymbol{\Sigma}_{\mathrm{ML}}(\hat{\boldsymbol{\theta}})]^{-1}$; these residuals are similarly related to fit functions with other methods (e.g., [3]). When the mean structure is analyzed – that is, M is fitted to the covariance matrix $\mathbf{S}$ and means $m_1, m_2, \ldots, m_k$ of $Z_1, Z_2, \ldots, Z_k$, respectively – also *mean structure residuals* (MSRs) can be obtained as $m_i - \mu_i(\hat{\boldsymbol{\theta}})$, where $\mu_i(\hat{\boldsymbol{\theta}})$ is the ith variable mean implied by the model at the parameter estimator ($i = 1, \ldots, k$). The CSRs and MSRs contain information about the location and degree of lack of fit of M, and in this sense may be considered local indices of fit (see also below). Standardized CSRs larger than 2 in absolute value may be viewed as indicative of a considerable lack of fit of M at least with regard to the pair of variables involved in each such residual; for a given model, however, these residuals are not independent of one another, and thus caution is advised when more than a few such residuals are examined in this way. Generally, a large positive (standardized) CSR may suggest that introduction of a parameter(s) further contributing to the relationship between the two variables involved may lead to model fit improvement; similarly, a negative (standardized) CSR with large absolute value may suggest that deleting or modifying the value of a parameter(s) currently involved in the variables' relationship may lead to marked improvement of model fit. Residuals discussed thus far can be routinely obtained for structural equation, factor analysis, and path analysis models with popular structural equation modeling software, such as LISREL, EQS, MPLUS, SEPATH, RAMONA, and SAS PROC CALIS (*see* **Structural Equation Modeling: Software**).

In addition to these residuals, *individual case residuals* (ICRs) can also be obtained that pertain to each studied individual (case) and dependent variable. In a path analysis model, say $\mathbf{Y} = \mathbf{BY} + \boldsymbol{\Gamma}\mathbf{X} + \mathbf{E}$ – where $\mathbf{Y}$ is a vector of dependent observed variables, $\mathbf{X}$ is that of independent observed variables, $\mathbf{E}$ is the vector of error terms, and $\mathbf{B}$ and $\boldsymbol{\Gamma}$ are corresponding coefficient matrices (such that $\mathbf{I} - \mathbf{B}$ is invertible, where $\mathbf{I}$ is the identity matrix of appropriate size) – ICRs result as $\mathbf{r}_p = \mathbf{Y}_p - (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\mathbf{X}_p$, where $\mathbf{X}_p$ and $\mathbf{Y}_p$ are the vectors of the pth individual's values on the independent and dependent variables, respectively, and $\mathbf{r}_p$ is the vector of his/her residuals (that is of the same size as $\mathbf{Y}$; $p = 1, \ldots, N$, with N denoting sample size). Estimates of these ICRs are furnished by substituting $\mathbf{B}$ and $\boldsymbol{\Gamma}$ in the last equation with their estimates obtained when fitting the model (cf. [4]). In a factor analysis model, $\mathbf{Y} = \Lambda\boldsymbol{\eta} + \boldsymbol{\varepsilon}$ – where $\boldsymbol{\eta}$ is the vector of latent factors (fewer in number than observed variables), Λ is the factor loading matrix, and $\boldsymbol{\varepsilon}$ that of error terms – ICRs are obtained as $\mathbf{r}_p = \mathbf{Y}_p - \Lambda\mathbf{f}_p$, where $\mathbf{f}_p$ is the vector of factor scores pertaining to the pth case ($p = 1, \ldots, N$; [1, 5]). Estimates of these residuals are furnished when substituting Λ in the last equation with its estimate (along with the factor score – e.g., Bartlett – estimates) obtained when fitting the model. In a structural equation model (with fewer factors than observed variables), $\mathbf{Y} = \Lambda\boldsymbol{\eta} + \boldsymbol{\varepsilon}$ and $\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta}$ – where $\mathbf{B}$ is the structural regression matrix and $\boldsymbol{\zeta}$ the vector of latent disturbances – ICRs are obtained as $\mathbf{r}_p = \mathbf{Y}_p - \Lambda(\mathbf{I} - \mathbf{B})^{-1}\mathbf{g}_p$, where $\mathbf{g}_p$ is the vector of the pth individual's factor scores for the model $\mathbf{Y} = \Lambda(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\zeta} + e(p = 1, \ldots, N;$ e being model error term). Estimates of these ICRs are furnished when in the equation $\mathbf{r}_p = \mathbf{Y}_p - \Lambda(\mathbf{I} - \mathbf{B})^{-1}\mathbf{g}_p$ estimates of Λ and $\mathbf{B}$ (along with factor score estimates) are substituted ([1, 5]; $p = 1, \ldots, N$). Extended individual case residuals (EICRs) for structural equation and factor analysis models with less latent than observed variables are discussed in Raykov & Penev [5]. The EICRs represent ICRs that are obtained using a nonorthogonal projection with a model error covariance matrix, and have been shown to differ across particular equivalent models [5]. Latent individual residuals (LIRs) are discussed in Raykov & Penev [6], which reflect individual case residuals with regard to a latent relationship, and can be used for purposes of studying latent variable relationships, as exemplified in Raykov & Penev [6].

Acknowledgment

I thank M. W. Browne, R. E. Millsap, and D. Rindskopf for valuable discussions on residual evaluation. This work was supported by the Research Institute on Addictions with the State University of New York.

References

[1] Bollen, K.A. & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models, in *Sociological Methodology*, P.V. Marsden, ed., Jossey-Bass, San Francisco, pp. 235–262.

[2] Bollen, K.A. (1989). *Structural Equations with Latent Variables*, Wiley, New York.

[3] Browne, M.W., MacCallum, R.C., Kim, C.-T., Andersen, B.L. & Glaser, R. (2002). When fit indices and residuals are incompatible, *Psychological Methods* **7**, 403–421.

[4] McDonald, R.P. & Bolt, D.M. (1998). The determinacy of variables in structural equation models, *Multivariate Behavioral Research* **33**, 385–402.

[5] Raykov, T. & Penev, S. (2001). The problem of equivalent structural equation models: an individual residual perspective, in *New Developments and Techniques in Structural Equation Modeling*, G.A. Marcoulides & R.E. Schumacker, eds, Lawrence Erlbaum, Mahwah, pp. 297–321.

[6] Raykov, T. & Penev, S. (2002). Exploring structural equation model misspecifications via latent individual residuals, in *Latent Variable and Latent Structure Models*, G.A. Marcoulides & I. Moustaki, eds, Lawrence Erlbaum, Mahwah, pp. 121–134.

TENKO RAYKOV

Resistant Line Fit

Pat Lovie

Volume 4, pp. 1756–1758

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Resistant Line Fit

A simple way of representing a linear relationship between two variables in a functional form on a **scatterplot** is to fit a *resistant line*. The aim is to find a line that makes the deviations in the y direction (the **residuals**) as small as possible while resisting the influence of any outlying points (**outliers**).

The initial line is found by splitting the n ordered x values (with their associated y values) into three approximately equal sized groups or batches. Any observations with the same x value should be kept together in the same batch and, ideally, the left (small x) and right (large x) batches should each contain the same number of observations. The **medians** for the x and associated y values in the left and right batches (x_L, x_R, y_L, y_R) determine the line through the point cloud. (Note that the pairing of the x and y values is ignored when finding the medians.)

If all we require is a quick visual indication of fit, then the initial resistant line is that which passes through the left and right batch x and y medians on the scatterplot; if the middle batch median deviates markedly from the line, this suggests possible nonlinearity.

Rough values of the slope b and intercept a for the resistant line can be obtained directly from the graph to give an equation for the fitted y values $\hat{y}_i = a + bx_i (\mathrm{i} = 1, \ldots, n)$. Alternatively, the coefficients can be found, using the medians, from the expressions $b = (y_R - y_L)/(x_R - x_L)$ and $a = mean(y_R - bx_R, y_M - bx_M, y_L - bx_L)$.

We can also assess linearity by looking at the half-slopes instead of 'by eye'. The left half-slope is the slope of the line joining the medians of the left and middle batches. Similarly, the right half-slope is that of the line between the medians of the middle and right batches. A minor modification of the slope formula given above is all that is required to obtain these values. If one of the half-slopes is more than twice the other, that is, the half-slope ratio is greater than two, we should not be attempting to fit a straight line.

A polishing routine can then used to adjust the line if the residuals (the differences between the observed and fitted y values) show any distinct pattern on a scatterplot of the residuals against the x values (or indeed on a **stem and leaf plot**). In essence, what we try to do is to balance up the size of the (median) residuals in the left and right batches.

To polish the line, we go through the same initial steps of the fitting process but this time using the residuals as new y values and find the slope b_{res}. Next, we adjust b by adding the value b_{res} to it, then recalculate the intercept and finally obtain the residuals for this new fitted line. The procedure is repeated if necessary until the residuals show no evidence of a relationship with the x values, that is, they have zero slope. Generally, the slope changes by smaller and smaller amounts at each iteration, eventually converging to a stable value. Clearly, these iterations are tedious to do by hand and thus are best left to statistical packages such as Minitab (*see* **Software for Statistical Analyses**).

As an illustration, suppose that our data (see Table 1) are Semester 1 and Semester 2 examination scores (each out of a maximum of 50) achieved by a sample of 11 students on a two-semester elementary statistics course. Naturally, we have already inspected a scatterplot (see Figure 1) and are fairly confident that there is a linear relationship between the variables; we note, however, one observation (10, 39) that is distinctly out of line with the remainder of the data. To fit a resistant line with Semester 1 score on the x axis, we split the data into three batches and find the medians for the x and y values:

The initial resistant line determined by the medians is superimposed on the scatterplot in Figure 1. It seems to pass a little too high of center for small values of x (apart from the 'outlier'), although we have

Table 1 Semester 1 and Semester 2 examination scores on an elementary statistics course

x (Semester 1)		***y*** (Semester 2)	
10		39	
15	$x_L = 17$	17	$y_L = 21$
19		21	
20		21	
25		29	
28	$x_M = 28$	37	$y_M = 29$
30		24	
32		33	
38	$x_R = 39$	36	$y_R = 39$
40		42	
44		43	

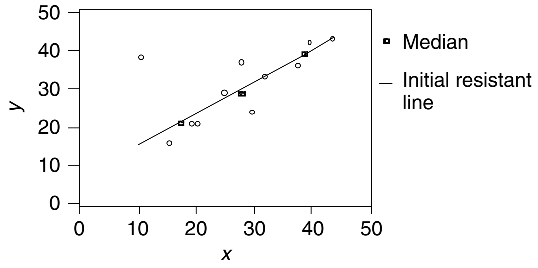


Figure 1 Scatterplot with initial resistant line

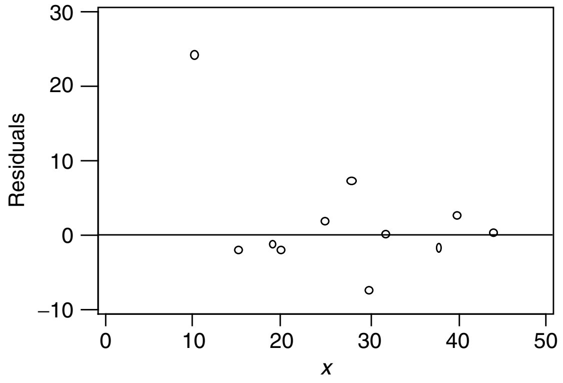


Figure 2 Residual plot for initial resistant line fit

to be careful not to be too dogmatic because our data set is very small. Also, as the line does not miss the middle batch median by much, nonlinearity is not a worry.

The equation of the initial resistant line calculated from the slope and intercept formulae given earlier is $\hat{y}_i = 6.76 + 0.82x_i$ $(i = 1, 2, \ldots, 11)$. Although it is difficult to discern any pattern in the residual plot (Figure 2), except that the residuals seem to be a little larger for middle sized x values, we shall allow Minitab to do some polishing for us. Minitab takes a few further iterations to come up with a final equation $\hat{y}_i = 3.81 + 0.91x_i$ $(i = 1, 2, \ldots, 11)$. In passing, we note that the half slope ratio is 1.25, confirming a fairly linear relationship.

Superimposing this polished line on the scatterplot (Figure 3), we see that new line has been shifted downward on the left side and now passes a little closer to the points with smaller x values and even further away from our outlier. This seems a reasonable fit to the data and would allow us to make some cautious predictions about performance at the end of Semester 2.

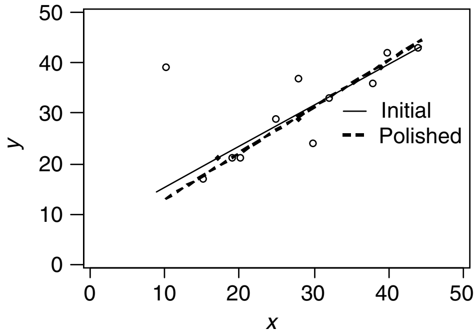


Figure 3 Scatterplot with initial (solid) and polished (dotted) resistant lines

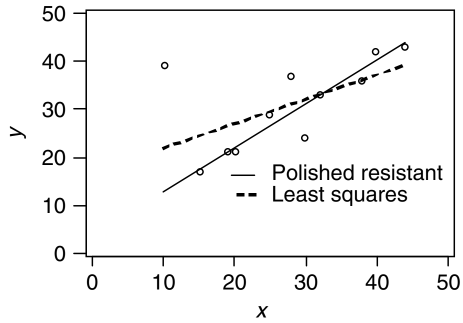


Figure 4 Scatterplot with polished resistant (solid) and least squares (dotted) lines overlaid

Suppose though that we had just pressed ahead with traditional least squares **regression** for these data. The equation of the least squares line is $\hat{y}_i = 17.38 + 0.50x_i$ $(i = 1, 2, \ldots, 11)$, which has a considerably higher intercept and a smaller slope than our polished resistant line (see Figure 4). The least squares line has been pulled toward the observation (10, 39), which we had already branded as an outlier.

To overcome the effect of this anomalous observation, we shall have to omit it from the analysis. The equation of regression line then becomes $\hat{y}_i = 4.60 + 0.88x_i$ $(i = 1, 2, \ldots, 10)$, which is quite close to our resistant line solution.

As we have seen, resistant line fitting can be a useful alternative to the least squares method for describing the relationship between two variables when the presence of outliers makes the latter procedure risky. However, it is essentially an exploratory technique and lacks the inferential framework of traditional regression analysis. Further discussion of resistant line fitting can be found in [1].

Reference

[1] Velleman, P.F. & Hoaglin, D.C. (1981). *Applications, Basics and Computing of Exploratory Data Analysis*, Duxbury Press, Boston.

PAT LOVIE

Retrospective Studies

Dean P. McKenzie

Volume 4, pp. 1758–1759

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Retrospective Studies

A retrospective study is a type of **observational study** (no random assignment to groups) in which the events being studied have already occurred, before the study has begun. Information on such events may already have been collected, perhaps for administrative purposes (e.g., employment records), or it may be obtained by questioning individuals. For example, Hart et al. [6] interviewed the carers of patients diagnosed with Alzheimer's disease. The interviewees were asked about the presence and duration of patients' symptoms such as anxiety, which had occurred in the past three years. **Case-control studies** are generally retrospective, indeed, the term retrospective study originally referred to this type of design [8]. The past experiences, (e.g., smoking behavior) of those identified as having a particular disease (e.g., lung cancer) or outcome (the cases), are compared with those who do not (the controls). A further type of retrospective study is the historical, or retrospective, **cohort study** [5]. A group or cohort of individuals is identified, based on characteristics found in historical databases such as accommodation, employment, medical (including case notes and charts), military or school records. Information on exposures and disease or outcome of interest is obtained from these or other records, and the cohort is followed up over 'historical time'. In other words, the cohort will be studied from a point backward in time, up to the more recent past, or present. For example, Zammit et al. [10] studied a historical cohort of over 50 000 Swedish males, who had originally been conscripted for compulsory military training during 1969–1970. A variety of demographic and other information, including self-reported cannabis use, had been collected and stored. Zammit et al. [10] sought to establish whether cannabis use (the exposure) was a risk factor for schizophrenia (the disease). The original records were accessed and linked to historical psychiatric diagnostic information, obtained from the Swedish hospital discharge registry, for the period 1970–1996. The **odds** of developing schizophrenia over this period for persons reporting cannabis use during 1969–1970 were then compared with the odds for those who did not. In contrast, a prospective cohort study (*see* **Prospective and Retrospective Studies**) is one in which individuals are selected based on their current, rather than their past, characteristics. Information is then collected from the beginning of the study until some point in the future. Prospective studies readily allow the directionality of events to be examined (e.g., is cannabis use a consequence of psychiatric illness rather than a probable cause? [2]), but suffer from the problem of drop-outs (a particular problem in studies involving illicit drug users [3]). Prospective studies can be extremely expensive and time-consuming. This is especially true if a large cohort has to be followed up over many years until the event of interest (e.g., Alzheimer's disease) occurs [1, 9, 4], in which case retrospective studies may be the only practical option. Retrospective studies readily and (comparatively) inexpensively allow the analysis of many thousands of individuals, over several decades, provided that the necessary historical information has been collected, or can be remembered. Memories may be highly inaccurate [1], and database information incomplete. Changes in measurement and diagnostic methods may also have occurred over time. For events such as suicide, however, there may be no alternative to the use of retrospective studies [1]. In practice, the distinction between prospective and retrospective studies can be somewhat blurred. A prospective cohort study might have a retrospective component, for example, individuals may be initially asked about their childhood experiences. Finally, either a prospective, or retrospective, cohort study, may provide data for a nested case-control study [5, 9] in which cohort members identified as having a particular diagnosis (e.g., depression) or outcome are compared with those who do not. Further information about retrospective studies can be found in [1], [4], [5], [8] and [9] with a light-hearted example given in [7].

References

[1] Anstey, K.J. & Hofer, S.M. (2004). Longitudinal designs, methods and analysis in psychiatric research, *Australian and New Zealand Journal of Psychiatry* **38**, 93–104.

[2] Arseneault, L., Cannon, M., Poulton, R., Murray, R., Caspi, A. & Moffitt, T.E. (2002). Cannabis use in adolescence and risk for adult psychosis: longitudinal prospective study, *British Medical Journal* **325**, 1212–1213.

[3] Bartu, A., Freeman, N.C., Gawthorne, G.S., Codde, J.P. & Holman, C.D.J. (2004). Mortality in a cohort of opiate and amphetamine users in Perth, Western Australia, *Addiction* **99**, 53–60.

[4] de Vaus, D.A. (2001). *Research Design in Social Research*, Sage Publications, London.

[5] Doll, R. (2001). Cohort studies: history of the method II. retrospective cohort studies, *Sozial – und Praventivmedizin / Social and Preventive Medicine* **46**, 152–160.

[6] Hart, D.J., Craig, D., Compton, S.A., Critchlow, S., Kerrigan, B.M., McIlroy, S.P. & Passmore, A.P. (2003). A retrospective study of the behavioural and psychological symptoms of mid and late phase Alzheimer's disease, *International Journal of Geriatric Psychiatry* **18**, 1037–1042.

[7] Nelson, M.R. (2002). The mummy's curse: historical cohort study, *British Medical Journal* **325**, 1482–1484.

[8] Paneth, N., Susser, E. & Susser, M. (2002). Origins and early development of the case-control study: part 1, early evolution, *Sozial – und Praventivmedizin / Social and Preventive Medicine* **47**, 282–288.

[9] Tsuang, M.T. & Tohen, M. (2002). *Textbook in Psychiatric Epidemiology*, 2nd Edition, Wiley-Liss, Hoboken.

[10] Zammit, S., Allebeck, P., Andreasson, S., Lundberg, I. & Lewis, G. (2002). Self reported cannabis use as a risk factor for schizophrenia in Swedish conscripts of 1969: historical cohort study, *British Medical Journal* **325**, 1199–1201.

DEAN P. MCKENZIE

Reversal Design

John Ferron

Volume 4, pp. 1759–1760

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Reversal Design

Reversal designs [1] are a type of **single-case design** used to examine the effect of a treatment on the behavior of a single participant. The researcher measures the behavior of the participant repeatedly during what is referred to as the baseline phase. After the baseline has been established, the researcher implements the treatment and continues to repeatedly measure the behavior during the treatment phase. The researcher then removes the treatment and reestablishes the baseline condition. Since the treatment is withdrawn during this third phase, many refer to this type of design as a withdrawal design.

One typically expects the behavior will be stable during the baseline phase, improve during the treatment phase, and then reverse, or move back toward baseline levels during the second baseline phase. When improvement is seen during the treatment phase, some may question whether this improvement was the result of the treatment or whether it resulted from maturation or some event that happened to coincide with treatment implementation. If we see the behavior return to baseline levels during the second baseline phase, these alternative explanations seem less plausible, and we feel more comfortable attributing the behavior changes to the treatment. Put another way, the reversal increases the **internal validity** of the design.

The minimum number of phases in a reversal design is three: a baseline phase (A), followed by a treatment phase (B), followed by the second baseline phase (A). It is possible, however, to extend the basic ABA design to include more phases, creating other phase designs, such as an ABAB design or an ABABAB design.

Reversal designs also vary in how the phase shifts are determined. In some cases, the assignment is systematic, for example, a researcher may decide that there will be eight observations in each phase. This method works well when baseline observations are constant, which allows one to assume temporal stability and use what has been referred to as the scientific solution to causal inference. When baseline observations are not constant, inferences become more difficult and one may alter the method of assigning phase shifts to facilitate drawing treatment–effect inferences.

One option is to use a response-guided strategy in which the data are viewed and conditions are changed after the researcher judges the data within a phase to be stable. If the researcher is able to identify and control the factors leading to variability in the baseline data, the variability can be reduced, hopefully leading to constancy in the later baseline observations and relatively straightforward inferences about treatment effects (*see* **Multiple Baseline Designs**).

When baseline variability cannot be controlled, one may turn to statistical methods for making treatment–effect inferences, which may motivate the use of some restricted form of **randomization** to choose the intervention points. For example, the intervention points could be chosen randomly under the constraint that there are at least five observations in each phase. A randomization test (*see* **Randomization Based Tests**) [3] could then be constructed to allow inference about the presence of a treatment effect. To make inferences about the size of a treatment effect when confronted with variable baseline data, one could turn to statistical modeling options (*see* **Statistical Models**) [2].

References

[1] Baer, D.M., Wolf, M.M. & Risley, T.R. (1968). Some current dimensions of applied behavior analysis, *Journal of Applied Behavior Analysis* **1**, 91–97.

[2] Huitema, B.E. & McKean, J.W. (2000). Design specification issues in time-series intervention models, *Educational and Psychological Measurement* **60**, 38–58.

[3] Onghena, P. (1994). Randomization tests for extensions and variations of ABAB single-case experimental designs: a rejoinder, *Behavioral Assessment* **14**, 153–171.

JOHN FERRON

Risk Perception

Brian S. Everitt

Volume 4, pp. 1760–1764

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Risk Perception

Alistair Cooke in *A Letter from America* during the Gulf War of 1990 told the sad story of an American family of four who cancelled their planned holiday to Europe because of the fears of terrorist attacks on the country's airlines. They decided to drive to San Francisco instead. At the last junction, before leaving their small Midwest town, they collided with a large truck with fatal consequences for them all.

Life is a risky business and deciding which risks are worth taking and which should be avoided has important implications both for an individual's lifestyle, and the way society operates. The benefits gained from taking a risk need to be weighed against the possible disadvantages. An acceptable risk is proportional to the amount of benefits. For the individual, living life to the fullest means achieving a balance between reasonable and unreasonable risk, and this balance is dependent largely on the individual's personality. But, in society as a whole, where the same balancing act is required, it has to be achieved through political action and legislation. If risks could be assessed and compared in a calm and rational manner, it would benefit both individuals and society. There is, however, considerable evidence that such assessment and comparison is not straightforward.

Defining and Quantifying Risk

A dictionary definition of risk is 'the possibility of incurring misfortune or loss'. (The word risk is derived from the Greek word 'rhiza', which refers to such hazards of sailing as: too near the cliffs, contrary winds, turbulent downdraughts, and swirling tides (*see* **Relative Risk**)). Quantifying risk and assessing risk involves the calculation and comparison of probabilities, although most expressions of risk are compound measures that describe both the probability of harm and its severity. Americans, for example, run a risk of about 1 in 7000 of dying in a traffic accident. This probability is derived by noting that in the year 2000 there were about 40 000 traffic accident deaths among a population of about 280 000 in the United States. The figure of 1 in 7000 expresses the overall risk to society. The risk to any particular individual clearly depends on her exposure: how much she is on the road, where she drives and in what weather, whether she is psychologically accident prone, what mechanical condition the vehicle is in, and so on. Because gauging risk is essentially probabilistic, it follows that a risk estimate can assess the overall chance that an untoward event will occur, but it is powerless to predict any specific event. Just knowing that the probability of tossing a head with a fair coin is one-half leaves one unable to predict which particular tosses will result in heads and which in tails.

Risk Perception

Risk perception is one's opinion of the likelihood of the risk that is associated with a certain activity or lifestyle. Risk perception is be influenced by sociological, anthropological, and psychological factors, with the result that people vary considerably in which risks they consider acceptable and which they do not, even when they agree on the degree of risk involved. For example, many people with no fear of traveling large distances by car or train consider the prospect of flying, even with a well-respected commercial airline, to be a nightmare, often requiring several trips to the airport bar before being able to board the airplane. For others, air travel represents the very model of a low-risk form of transport. As Table 1 shows, flying is actually one of the safest forms of transport.

Perhaps one reason for some people's excessive and clearly misguided fear of flying is the general view that being killed in a plane crash must be a particularly nightmarish way to die. Another possibility is that the flying phobic considers the sky an alien environment, and this consideration distorts the perception of the risk involved. A third possible reason is that flying accidents are more prominent in the media than those involving automobiles, although the latter are far more common. Perception of risk is also likely to be influenced by whether we feel in control of a perceived risk.

Table 1 Causes of death and their probability

Cause of death	Probability
Car trip across the United States	1 in 14 000
Train trip across the United States	1 in 1 000 000
Airline accident	1 in 10 000 000

Research shows that people tend to overestimate the probability of unfamiliar, catastrophic, and well-publicized events and to underestimate the probability of unspectacular or familiar events that claim one victim at a time. For example, in one study [2], respondents rated 90 hazards, each with respect to 18 qualitative characteristics such as whether the risk was voluntary or involuntary, personally controllable or not, and known to those exposed or not. A **principal component analysis** of the data identified two major components, and the location of the 90 hazards with respect to those two components. The first factor was labeled as 'dread' risk. This factor related judgments of scales such as uncontrollability, fear, and involuntariness of exposure. Hazards that rate high on this factor include nuclear weapons, nerve gas, and crime. Hazards that rate low on this factor include home appliances and bicycles. A second factor, labeled 'unknown risk' related judgments of the observability of risks, such as whether or not the effects are delayed in time, the familiarity of the risk, and whether or not the risks are known to science. Hazards that rate high on this dimension include solar electric power, DNA research and satellites; those that rate low include motor vehicles, fire fighting, and mountain climbing.

Slovic, Fischhoff, and Lichtenstein [2] conclude that perceptions of risk are clearly related to the position of an activity in the principal component space, particularly in respect of the 'dread' factor. The higher a hazard's score on this factor, the higher its perceived risk, the more people want to see its current risks reduced, and the more they want to see strict regulation employed to achieve the desired reduction in risk. It seems that the risks that kill people and the risks that scare people are different.

The findings in [2] are supported by the results of polls of college students and members of the League of Women Voters in Oregon. Both groups considered nuclear power their number-one 'present risk of death', far riskier than motor vehicle accidents, which kill almost 42 000 Americans each year, or cigarette smoking, which kills 150 000, or handguns, which kill 17 000. Experts in risk assessment in the same poll considered motor vehicle accidents their number-one risk, with nuclear power below the risk of swimming, railroads, and commercial aviation. Here, the experts seem to have the most defendable conclusions. Average annual fatalities expected from nuclear power, according to most scientific estimates, are fewer than 10. Nuclear power does not appear to merit its risk rating of number one. It appears that the two well-educated and influential segments of the American public polled in Oregon seem to have been misinformed. The misinformants are not difficult to identify. Journalists report almost every incident involving radiation. A truck containing radioactive material is involved in an accident, a radioactive source is temporarily lost, a container leaks radioactive materials – all receive nationwide coverage, whereas the 300 Americans killed each day in other types of accidents are largely ignored. Reports in the media concentrate on issues and situations that frighten – and therefore interest – readers and viewers. The media fills its coverage with opinions (usually from interested parties) rather than facts or logical perspectives. In terms of nuclear power, for example, phrases such as *deadly radiation* and *lethal radioactivity* are common, but the corresponding *deadly cars* and *lethal water* would not sell enough newspapers, although thousands of people are killed each year in automobile accidents and by drowning. The problem is highlighted by a two-year study of how frequently different modes of death become front-page stories in the New York Times. It was found that the range was from 0.02 stories per 1000 annual deaths from cancer to 138 stories per 1000 annual deaths from airplane crashes.

Misperception of risk fueled by the media can lead to unreasonable public concern about a hazard, which can cause governments to spend a good deal more to reduce risk in some areas and a good deal less in other areas. Governments may, for example, spend huge amounts of money protecting the public from, say, nuclear radiation, but are unlikely to be so generous in trying to prevent motor vehicle accidents. They react to loudly voiced public concern in the first case and to the lack of it in the second. But, if vast sums of money are spent on inconsequential hazards, little will be available to address those that are really significant.

Examples of disparities that make little sense are not hard to find. In the late 1970s, for example, the United States introduced new standards on emissions from coke ovens in steel mills. The new rules limited emissions to a level of no more than 0.15 mg/m^3 of air. To comply with this regulation, the steel industry spent $240 million a year. Supporters of the change estimated that it would prevent about 100 deaths from cancer and respiratory disease each year, making the average annual cost per life saved $2.4 million. It is difficult to claim that this is money well spent when a large scale program to detect lung cancer in its earliest stages, for example, might be expected to extend the lives of 7000 cancer victims an average of one year for $15 000 each, and when the installation of special cardiac-care units in ambulances could prevent 24 000 premature deaths a year at an average cost of a little over $200 for each year of life.

Politicians often sanction huge expenditures to save an identifiable group of trapped miners, but not to improve mine safety or to reduce deaths among the scores of anonymous miners who die from preventable work related causes each year. Politically, at least, Joseph Stalin might have got it just about right when he mused that a single death is a tragedy, but a million deaths is a statistic.

Risk Presentation

The ability to make a rational assessment of risk is important for the individual, but also for a society that hopes to be governed by sensible, justifiable policies and legislation. It is unfortunate that, in general, people have both a limited ability to interpret probabilistic information and a mistrust of experts (sadly, statisticians in particular). But, rather than dismissing public understanding of technical issues as being insufficient for 'rational' decision making, experts (including statisticians) need to make a greater effort to increase the public's appreciation of how to evaluate and compare risks. Risks can be presented in ways that make them more transparent. For example, risks presented as annual fatality rates per 100 000 persons fail to reflect the fact that some hazards such as pregnancy and motor cycle accidents cause death at a much earlier age than other hazards such as lung cancer caused by smoking. One way to overcome this problem is to calculate the average loss of life expectancy due to exposure to the hazard. Table 2 gives some examples of risks presented in this way.

Table 2 Life expectancy reduction from a number of hazards

Risk	Days lost
Being unmarried – male	3500
Cigarette smoking – male	2250
Heart disease	2100
Being unmarried – female	1600
Cancer	980
Being 20% overweight	900
Low socioeconomic status	700
Stroke	520
Motor vehicle accidents	207
Alcohol	130
Suicide	95
Being murdered	90
Drowning	41
Job with radiation exposure	40
Illicit drugs	18
Natural radiation	8
Medical X rays	7
Coffee	6
Oral contraceptives	5
Diet drinks	2
Reactor accidents	2
Radiation from nuclear industry	0.02

So, for example, the average age at death for unmarried males is 3500 days younger than the corresponding average for men who are married. This does not, of course, imply a cause (marrying) and effect (living 10 years longer) relationship that is applicable to every individual, although it does, in general terms at least, imply that the institution of marriage is 'good' for men. And, the ordering in Table 2 should largely reflect society's and government's ranking of priorities for increasing the general welfare of its citizens. Thus, rather than introducing legislation about the nuclear power industry or diet drinks, a rational government should set up computer dating services that stress the advantages of marriage (particularly for men) and encouraging people to control their eating habits. It is hard to justify spending any money or effort on reducing radiation hazards or dietary hazards such as saccharin.

Perhaps the whole problem of the public's appreciation of risk evaluation and risk perception would diminish if someone could devise a simple scale of risk akin to the *Beaufort scale* for wind speed or the *Richter scale* for earthquakes. Such a 'riskometer' might provide a single number that would allow meaningful comparisons among risks from all types of hazards, whether they be risks due to purely voluntary activities (smoking and hang gliding), risks incurred in pursuing voluntary, but virtually necessary, activities (travel by car or plane, eating meat), risks imposed by society (nuclear waste, overhead power lines, legal possession of guns), or risks due to acts of God (floods, hurricanes, lightning strikes).

Table 3 Risk index values based on Paulos [1]. The lower the number, the greater the risk

Event or activity	Risk index
Playing Russian roulette once a year	0.8
Smoking 10 cigarettes a day	2.3
Being struck by lightning	6.3
Dying from a bee sting	6.8

One such scale is described by Paulos [1], and is based on the number of people who die each year pursuing various activities. If one person in N dies, the associated risk index is set at $\log_{10} N$; 'high' values indicate hazards that are not of great danger, whereas 'low' values suggest activities to be avoided if possible. (A logarithmic scale is used because the risks of different events and activities differ by several orders of magnitudes.) If everybody who took part in a particular pursuit or was subjected to a particular exposure died, then Paulos's risk index value becomes zero, corresponding to certain death. (Life is an example of such a deadly pursuit.) In the United Kingdom, 1 in every 8000 deaths each year results from motor vehicle accidents; consequently, the risk index value for motoring is 3.90. Table 3 shows examples of values for other events.

Paulos's risk index would need refinement to make it widely acceptable and practical. Death, for example, is not the only concern; injury and illness are also important consequences of exposure to hazards and would need to be quantified in any worthwhile 'index'. But, if such an index could be devised, it might help prevent the current, often sensational approach to hazards and risks that is adopted by most journalists. A suitable riskometer rating might help improve both media performance and the general public's perception of risks.

Summary

The general public's perceptions of risk are often highly inaccurate, but by underestimating common risks while exaggerating exotic ones, we may end up protecting ourselves against unlikely perils while failing to take precautions against those that are far more dangerous. For example, people may be persuaded by sensational news stories that chemicals and pesticides considerably increase the risk of certain types of cancer. Perhaps they do, but the three main causes of cancer remain smoking, dietary imbalance, and chronic infections. Statisticians, psychologists, and others should put more effort into finding ways of presenting risks so that they may be more rationally appraised and compared. This is unlikely to be easy because people tend to form opinions rather quickly, usually in the absence of strong supporting evidence. Strong beliefs about risk, once formed, change very slowly and are extraordinarily persistent in the face of contrary evidence. Risk communication research must take up this challenge if the public is ever going to be persuaded to be rational about risk.

References

[1] Paulos, J. (1994). *Innumeracy*, Penguin Books, London.

[2] Slovic, P., Fischhoff, B. & Lichtenstein, S. (1980). Facts and fears: understanding perceived risk, in *Societal Risk Assessment: How Safe is Safe Enough?* R.C. Schwing & W.A. Albers, eds, Plenum Press, New York.

Further Reading

Royal Society Study Group. (1983). *Risk Assessment*, The Royal Society, London.

BMJ. (1990). *The BMA Guide to Living with Risk*, Penguin, London.

Walsh, J. (1996). *True Odds*, Merritt Publishing, Santa Monica.

(*See also* **Odds and Odds Ratios**; **Probability: An Introduction**)

Brian S. Everitt

Robust Statistics for Multivariate Methods

Maria-Pia Victoria-Feser

Volume 4, pp. 1764–1768

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Robust Statistics for Multivariate Methods

Robust statistics, as a concept, probably dates back to the prehistory of statistics. It has, however, been formalized in the sixties by the pioneering work of Huber [9, 10] and Hampel [6, 7]. Robust statistics is an extension of classical statistics, which takes into account the fact that models assumed to have generated the data at hand are only approximate. It provides tools to investigate the robustness properties of a statistic T (such as estimators, test statistics) as well as robust estimators and robust testing procedures (*see* **Robust Testing Procedures**).

Although one would easily agree that models can only describe approximately the reality, what is more difficult to understand is the effect of this fact on the properties of classical statistics T for which it is assumed that models are exact. Suppose that the hypothetical (multivariate) model is denoted by F but that the data at hand have been generated by the general mixture $F_\varepsilon = [1 - \varepsilon]F + \varepsilon H$, with H a contamination distribution. Assuming F_ε means that the data have been generated by the model F with probability $[1 - \varepsilon]$ and by a contamination distribution H with probability ε. Note that a particular case for H is a distribution assigning a probability of one to an arbitrary point, that is, producing so-called outliers. If ε is large, the contamination distribution has an important weight in the mixture distribution and an analysis based on F_ε (assuming F as the central model) is meaningless. On the other hand, if ε is small, an analysis based on F_ε should not be entirely determined by the contamination. It is, therefore, important to find or construct statistics T that are not entirely determined by data contamination, that is, robust under slight model deviations (*see* **Finite Mixture Distributions**).

A well-known tool to assess the effect (on the bias of T) of infinitesimal amounts ε of contamination is the *influence function* (*IF*) introduced by Hampel [6, 7] and further developed in [8]. Another tool is the *breakdown point* (*BDP*), which measures the maximal amount ε of (any type of) contamination that T can withstand before it 'breaks down' or gives unreliable results (see for example [8]). A statistic T with bounded *IF* is said to be robust (in the infinitesimal sense). It should be stressed that most classical procedures are not robust, and, in particular, all classical procedures for models based on the multivariate normal distribution (*see* **Catalogue of Probability Density Functions**) are not robust. This is the case, for example, for regression models (*see* **Multiple Linear Regression**), **factor analysis**, **structural equation models**, **linear multilevel models** (which include **repeated measures analysis of variance**).

In practice, to detect observations from a contamination distribution (i.e., contaminated data) is not an obvious task. For models based on the p-variate normal distribution $F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$, a useful measure is the **Mahalanobis distance** d_i defined on each (multivariate) observation $\mathbf{x}_i$ by

$$d_i^2 = (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \qquad (1)$$

The d_i takes into account the covariance structure of the data, which is very important in multivariate settings (*see* **Multivariate Analysis: Overview**). Indeed, as an example we consider scores on psychological tests collected for the study of age differences in working memory (see [1] for more details), which is presented as a multi **scatterplot** in Figure 1. A close look at the scatterplot between the variables ML1TOT and ML2TOT reveals that there is a minority of subjects not 'fitting' the covariance structure described by the bulk of data (i.e., the majority). On the other hand, on the univariate level, that is, when looking a the scores only on one of each variable, this minority of subjects has not so extreme scores. The point here is that, when dealing with multivariate models, the screening of the data at the univariate level is not sufficient to detect contaminated data (*see* **Multivariate Outliers**).

Unfortunately, scatter plots show only the behavior of the data at the bivariate level, and (exact) bivariate normality does not imply (exact) normality of higher orders. It is, therefore, important to be able to rely on general measures such as (1). However, (1) supposes the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ to be known, which, in practice, is never the case. If nonrobust estimators are used, then they are biased in the presence of data contamination, which means that the d_i will in the best case not reveal the right contaminated data (masking effect), and, in the worst, reveal false contaminated data.

Robust statistics for multivariate models have first been used for the estimation of multivariate mean (location) and covariance (scatter). In this setting, it is desirable for the robust estimators to

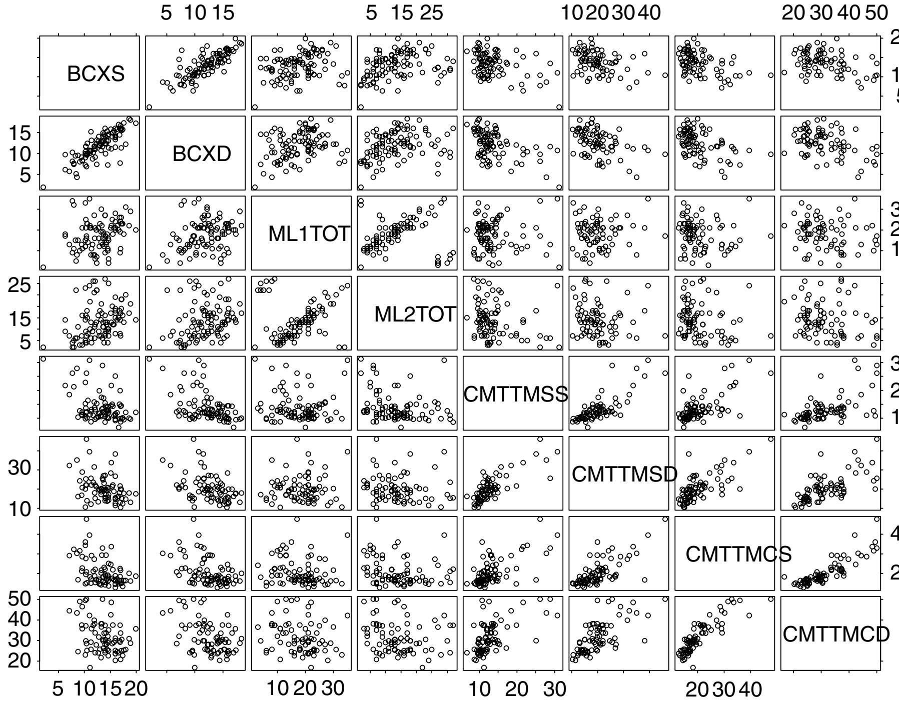


Figure 1 Multiscatter plot of the working memory study data

be affine equivariant (a linear transformation of the data results in a known transformation of the estimates), to have relatively high *BDP* (see [13]) and to be computationally efficient. The first high *BDP* affine equivariant estimator is the *minimum volume ellipsoid* (MVE) proposed by [17]. The ellipsoid (of dimension p) containing at least half of the data with minimum volume is found and the sample mean and covariance of these data define the MVE. The latter is very computationally intensive, and is known to have poor efficiency. However, it is used, for example, to detect contaminated data or as a starting point for more efficient estimators based on weighted means and covariances.

A general class of estimators in which one can find robust ones is the class of M-estimators (see [10]) that generalize **maximum likelihood estimators (MLE)**. M-estimators (*see* **M Estimators of Location**) are defined for general parametric models F_θ as the solution in θ of

$$\frac{1}{n}\sum_{i=1}^{n}\psi(\mathbf{x}_i,\theta)=0 \tag{2}$$

When the ψ-function is the score function $s(\mathbf{x},\theta)=(\partial/\partial\theta)\log f(\mathbf{x},\theta)$, one gets the MLE. Such estimators, under very mild conditions, have known asymptotic properties that can be used for inference (see e.g., [8]). For the multivariate normal model, another popular class of estimators is the class of S-estimators (see [18]), which can be computed iteratively by means of

$$\frac{1}{n}\sum_{i=1}^{n}w_i^{\mu}(\boldsymbol{\mu}-\mathbf{x}_i)=0 \tag{3}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[w_i^{\delta}\boldsymbol{\Sigma}-w_i^{\eta}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T\right]=0 \tag{4}$$

where the weights $w_i^{\mu}, w_i^{\eta}, w_i^{\delta}$ are decreasing functions of the Mahalanobis distances d_i. Note that when the former are equal to 1 for all i, one gets the classical sample means and covariances. The choice for the weights define different estimators (see e.g., [16]). When there are missing data, [1] proposed an adaptation of (3) and (4) as an alternative to the EM algorithm (see [4]). For the working memory data (which include missing data), the correlation between ML1TOT and ML2TOT was found to be 0.84 (robust estimation), whereas it is equal to 0.20 when using the EM algorithm. Other robust estimators for multivariate location and scatter (and their statistical properties) can be found in, for example, [3], [5], [11], [12], [14], [19], [20], [21], [22] and [23].

Although the multivariate normal distribution (*see* **Catalogue of Probability Density Functions**) is the central distribution for several models, the covariance matrix is not always present in a free form. Indeed, like in **structural equations models** or in mixed linear models (*see* **Linear Multilevel Models**), the true covariance matrix is structured. For example, it could be supposed that the variances are all equal, and the covariances are all equal (one-way ANOVA with repeated measures). In these cases, it is important to estimate the covariance matrix by taking into account its structure, and not just estimate it freely, and then 'plug-in' the estimate in the model to estimate the other parameters. [2] proposed a general class of S-estimators for constrained covariance matrices that can be used for example with mixed linear models.

When the models are not based on the multivariate normal distribution, robust statistics become more complex. The Mahalanobis distance does not play anymore a role, and another measure for detecting contaminated data needs to be specified. For M-estimators, [10] proposed a weighting scheme based on the score function itself, that is,

$$\psi(\mathbf{x},\theta) = w_c(\mathbf{x},\theta)s(\mathbf{x},\theta) \tag{5}$$

with

$$w_c(\mathbf{x},\theta) = \min\left\{1; \frac{c}{\|s(\mathbf{x},\theta)\|}\right\} \tag{6}$$

where $\|\mathbf{x}\| = \left(\sum_{j=1}^{p} x_j^2\right)^{1/2}$ denotes the Euclidian norm. Observations corresponding to large (absolute) value of the score function are hence downweighted. The score function in a sense replaces the Mahalanobis distance for multivariate normal models. The parameter c can be chosen for efficiency arguments. With nonsymmetric models, (5) leads to inconsistent estimators, and, therefore, a shift needs to be added to (5) to make the M-estimator consistent (see for example, [8] and [15]). This can make the robust estimator computationally nearly unfeasible. With nonnormal multivariate models, robust statistics, therefore, still need to be further developed.

References

[1] Cheng, T.-C. & Victoria-Feser, M. (2002). High breakdown estimation of multivariate mean and covariance with missing observations, *British Journal of Mathematical and Statistical Psychology* **55**, 317–335.

[2] Copt, S. & Victoria-Feser, M.-P. (2003). *High Breakdown Inference in the Mixed Linear Model.* Cahiers du département d'econométrie no 2003.6, University of Geneva.

[3] Davies, P.L. (1987). Asymptotic behaviour of S-estimators of multivariate location parameters and dispersion matrices, *The Annals of Statistics* **15**, 1269–1292.

[4] Dempster, A.P., Laird, M.N. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* **39**, 1–22.

[5] Donoho, D.L. (1982). Breakdown properties of multivariate location estimators, Ph.D. qualifying paper, Department of Statistics, Harward University.

[6] Hampel, F.R. (1968). Contribution to the theory of robust estimation, Ph.D. thesis, University of California, Berkeley.

[7] Hampel, F.R. (1974). The influence curve and its role in robust estimation, *Journal of the American Statistical Association* **69**, 383–393.

[8] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, John Wiley, New York.

[9] Huber, P.J. (1964). Robust estimation of a location parameter, *Annals of Mathematical Statistics* **35**, 73–101.

[10] Huber, P.J. (1981). *Robust Statistics*, John Wiley, New York.

[11] Kent, J.T. & Tyler, D.E. (1996). Constrained M-estimation for multivariate location and scatter, *The Annals of Statistics* **24**, 1346–1370.

[12] Lopuhaä, H.P. (1991). τ-estimators for location and scatter, *Canadian Journal of Statistics* **19**, 307–321.

[13] Maronna, R.A. (1976). Robust M-estimators of multivariate location and scatter, *The Annals of Statistics* **4**, 51–67.

[14] Maronna, R.A. & Zamar, R.H. (2002). Robust estimates of location and dispersion for high-dimensional datasets, *Technometrics* **44**, 307–317.

[15] Moustaki, I. & Victoria-Feser, M.-P. (2004). *Bounded-Bias Robust Inference for Generalized Linear Latent Variable Models.* Cahiers du département d'econométrie no 2004.02, University of Geneva.

[16] Rocke, D.M. & Woodruff, D.L. (1996). Identification of outliers in multivariate data, *Journal of the American Statistical Association* **91**, 1047–1061.

[17] Rousseeuw, P.J. (1984). Least median of squares regression, *Journal of the American Statistical Association* **79**, 871–880.

[18] Rousseeuw, P.J. & Yohai, V.J. (1984). Robust regression by means of S-estimators, in *Robust and Nonlinear Time Series Analysis*, J.W., Franke & R.D., Martin editors, Hardle W, eds, Springer-Verlag, New York, pp. 256–272.

[19] Stahel, W.A.(1981). Breakdown of covariance estimators, Technical Report 31, Fachgruppe für Statistik, ETH, Zurich.

[20] Tamura, R. & Boos, D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance, *Journal of the American Statistical Association* **81**, 223–229.

[21] Tyler, D.E. (1983). Robustness and efficiency properties of scatter matrices, *Biometrika* **70**, 411–420.

[22] Tyler, D.E. (1994). Finite sample breakdown points of projection based multivariate location and scatter statistics, *Annals of Statistics* **22**, 1024–1044.

[23] Woodruff, D.L. & Rocke, D.M. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators, *Journal of the American Statistical Association* **89**, 888–896.

MARIA-PIA VICTORIA-FESER

Robust Testing Procedures

RAND R. WILCOX

Volume 4, pp. 1768–1769

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Robust Testing Procedures

The term robust testing procedure roughly refers to hypothesis testing methods that are relatively insensitive to violations of the assumptions upon which they are based. This means, in particular, that a robust testing procedure should achieve two fundamental goals. The first is that the actual type I error probability is reasonably close to the nominal level. So if some hypothesis is tested at the .05 level, for example, the actual probability of a type I error should be reasonably close to .05. The second is that small shifts or changes in a distribution should not have an undue influence on **power**, the probability of detecting situations where the null hypothesis is false. In particular, if a hypothesis testing procedure has high power under normality, power should remain reasonably high when data are sampled from a slightly nonnormal distribution.

The mathematical tools for studying and understanding robustness issues have advanced considerably during the last forty years. The mathematical foundations of these methods are summarized by Hample, Ronchetti, Rousseeuw, and Stahel [1] and Huber [2]. A description of these methods, written at a more elementary level, is provided by Staudte and Sheather [3]. Briefly, these tools provide a way of studying and characterizing the effect that small changes in a distribution have on measures of location (such as the mean) and dispersion (such as the usual variance). This has led to an understanding of why conventional hypothesis testing methods, such as the two-sample Student's t Test, and ANOVA F test, (*see* **Catalogue of Parametric Tests**) are not robust, contrary to what was initially thought. Not only do they suffer from problems when trying to control the probability of a type I error but also arbitrarily small departures from normality can substantially lower power relative to other techniques that might be used (*see* **Robustness of Standard Tests**).

Also, these mathematical tools have formed the foundation for new inferential methods that deal with the problems now known to plague conventional techniques. For example, a common way of improving efficiency relative to the sample mean is to downweight **outliers**. But this leads to a technical problem: the usual method for estimating standard errors no longer applies. If, for example, outliers are simply removed and standard methods for means are applied to the remaining data, the wrong standard error is being used, which can result in poor control over the probability of a type I error. But, owing to the new mathematical and statistical tools that have emerged, theoretically sound estimates are now available.

When comparing two or more groups of participants, there are many ways of improving control over the probability of a type I error versus conventional methods based on means. Each approach has advantages and disadvantages, which are discussed in Wilcox [4]. For a more detailed description written at a slightly more advanced level, see Wilcox [5]. Some of these methods deal directly with measures of location, roughly referring to a value intended to reflect what is typical. The mean and median are the best-known examples, but today other measures of location have been found to have practical value such as trimmed means and **M-estimators of location** (*see* **Trimmed Means**).

Conventional methods for means can suffer from low power for three general reasons: unequal variances, skewness, and sampling from distributions where **outliers** are relatively common. (The term outliers refers to values that are unusually small or large.) All three create serious concerns, but perhaps outliers are particularly troublesome. One reason is that modern outlier detection methods suggest that outliers are rather common. Another reason is that outliers can inflate the usual sample variance, which in turn can mean low power (*see* **Robustness of Standard Tests**). But even if no outliers are detected, skewness and unequal variances can result in relatively low power as well.

One general approach when trying to avoid low power, due to nonnormality, is to replace means with a measure of location that provides reasonably high power under normality, but unlike methods based on means, relatively high power is achieved when sampling from nonnormal distributions. A crucial feature of these alternative estimators is that they deal directly with outliers. Methods based on medians can offer improved control over the probability of a type I error, but their power is relatively unsatisfactory when the data are normal. But other measures of location, such as trimmed means and M-estimators, satisfy both criteria. No single method dominates, but there are several inferential techniques that appear to perform well for a broad range of situations (e.g., [5]).

As for regression and correlation, conventional hypothesis testing methods inherit all of the practical problems associated with conventional methods for comparing groups based on means, and new problems are introduced. Again, vastly improved methods have been derived [5]. For example, even under normality, it is known that heteroscedasticity is a serious practical problem when using the ordinary least squares estimator (*see* **Least Squares Estimation**), but methods that deal with this problem are available. Also, both heteroscedasticity and nonnormality can result in relatively poor power when using ordinary least squares, but many modern estimators provide relatively high power under both normality and homoscedasticity as well as nonnormality and heteroscedasticity. Like methods intended to improve on the sample mean, they are based on techniques that are relatively insensitive to outliers.

References

[1] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. (1986). *Robust Statistics*, Wiley, New York.

[2] Huber, P.J. (1981). *Robust Statistics*, Wiley, New York.

[3] Staudte, R.G. & Sheather, S.J. (1990). *Robust Estimation and Testing*, Wiley, New York.

[4] Wilcox, R.R. (2004). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd Edition, Academic Press, San Diego.

[5] Wilcox, R.R. (2003). *Applying Conventional Statistical Techniques*, Academic Press, San Diego.

RAND R. WILCOX

Robustness of Standard Tests

Rand R. Wilcox

Volume 4, pp. 1769–1770

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Robustness of Standard Tests

Standard hypothesis-testing methods make two crucial assumptions: normality and homoscedasticity (equal variances). These methods include student's t, the **analysis of variance (ANOVA)** F test, and basic methods for testing hypotheses about least squares regression parameters and Pearson's correlation (*see* **Multiple Linear Regression**; **Catalogue of Parametric Tests**). If either or both of these assumptions are violated, it is questionable whether these methods provide an adequate control over the probability of type I errors, and whether they have high **power** relative to other methods that might be used. Early investigations into the robustness of the ANOVA F by Box [1] seemed to suggest that it is indeed relatively insensitive to violations of assumptions. But more recent results summarized by Hampel, Ronchetti, Rousseeuw and Stahel [3], Huber [4], Staudte and Sheather [5], and Wilcox [7–9] paint a decidedly different picture. These newer results might appear to contradict results reported by Box, but this is not the case. New theoretical results have helped improve our understanding of where to look for problems, the result being the realization that any method for comparing groups based on means can be expected to have serious practical difficulties for a broad range of situations.

Consider, for example, the usual one-way ANOVA F test for independent groups. Box [1] reported theoretical and empirical results on how unequal variances affect the probability of a type I error, assuming normality. Let R be the largest (population) standard deviation among the groups divided by the smallest standard deviation. Box limited his numerical results to situations where $R \leq \sqrt{3}$. No reason was given for not considering larger ratios, perhaps because at the time there were few, if any, studies looking into how much the standard deviations might differ in practice. But, various empirical studies summarized by Wilcox and Keselman [10] suggest that much larger ratios are common. When comparing two groups only, and when sampling from a normal distribution, reasonably good control over the probability of a type I error is achieved except possibly for very small sample sizes. However, for unequal sample sizes, or when sampling from a nonnormal distribution, this is no longer the case as indicated, for example, in [7–10]. Even under normality with four or more groups, unequal variances can result in poor control over the probability of a type I error, and nonnormality exacerbates this problem.

A basic requirement of any method is that under random sampling, it should converge to the correct answer as the sample sizes get large. For example, when computing a 0.95 confidence interval, the actual probability coverage should approach 0.95. With unequal sample sizes, there are general conditions where student's t does not satisfy this criterion [2].

A rough rule is that, as an experimental design becomes more complicated, standard hypothesis testing methods become more sensitive to violations of assumptions. For example, under normality and with equal sample sizes, student's t is relatively robust in terms of type I errors when the variances are unequal, but this is no longer the case when using the ANOVA F with four or more groups.

There are at least three general reasons why conventional methods for means can have relatively low power: unequal variances, **skewness**, and sampling from distributions where **outliers** are relatively common. Outliers are values that are unusually large or small. Outliers are of particular concern because modern outlier detection methods suggest they are rather common and because they can inflate the usual sample variance, which in turn can mean low power.

As an illustration, consider the following data.

$$\begin{aligned}&\text{Group 1}: \ 6\ 9\ 19\ 9\ 8\ 12\ 14\ 11\ 14\\&\text{Group 2}: \ 4\ 7\ 2\ 10\ 15\ 11\ 1\ 3.\end{aligned}$$

The sample means are 11.3 and 6.6, respectively, and student's t rejects at the 0.05 level. Now suppose the largest value in the first group is increased to 34. Then, the mean for the first group increases from 11.3 to 13, so the difference between the two means has increased from 4.7 to 6.4, yet we no longer reject; the P value is 0.08. Increasing the largest value to 80, the mean for the first group increases to 18, yet the P value increases to 0.19. The reason is that the sample variance has increased as well. The value 80, for example, is an outlier among the data for the first group, and it inflates the sample variance to the point that we no longer reject, even though the difference between the sample means has increased as well. Even small departures from normality can cause problems, a result that became evident with the publication of a seminal paper by Tukey [6]. Theoretical results were developed during the 1960s with the goal of dealing with this problem [3–5], and they form the basis of a wide range of modern inferential techniques [7–10]. Modern **robust testing procedures** not only deal with low power due to nonnormality, they substantially reduce problems associated with skewness and heteroscedasticity (*see* **Heteroscedasticity and Complex Variation**).

References

[1] Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way model, *Annals of Mathematical Statistics* **25**, 290–302.

[2] Cressie, N.A.C. & Whitford, H.J. (1986). How to use the two sample t-test, *Biometrical Journal* **28**, 131–148.

[3] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. (1986). *Robust Statistics*, Wiley, New York.

[4] Huber, P.J. (1981). *Robust Statistics*, Wiley, New York.

[5] Staudte, R.G. & Sheather, S.J. (1990). *Robust Estimation and Testing*, Wiley, New York.

[6] Tukey, J.W. (1960). A survey of sampling from contaminated normal distributions, in *Contributions to Probability and Statistics*, I. Olkin, et al., eds, Stanford University Press, Stanford.

[7] Wilcox, R.R. (2001). *Fundamentals of Modern Statistical Methods: Substantially Increasing Power and Accuracy*, Springer, New York.

[8] Wilcox, R.R. (2004). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd Edition, San Diego, CA: Academic Press.

[9] Wilcox, R.R. (2003). *Applying Conventional Statistical Techniques*, Academic Press, San Diego.

[10] Wilcox, R.R. & Keselman, H.J. (2003). Modern robust data analysis methods: measures of central tendency, *Psychological Methods* **8**, 254–274.

Rand R. Wilcox

Runs Test

CLIFFORD E. LUNNEBORG

Volume 4, pp. 1771–1771

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Runs Test

Two-sample Runs Test

The Wald–Wolfowitz [3] runs test dates from 1940, making it one of the earliest nonparametric tests. It provides a test of a common distribution for two independent random samples. However, the test has low power relative to such alternatives as the Kolmogorov–Smirnov or Cramér–von Mises two-sample tests and has declined in popularity, as attested to by [1] and [2].

Let $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_m)$ be independent random samples of sizes n and m from the two random variables, X and Y. The scale of measurement of the two random variables is at least ordinal and, to avoid problems with ties, ought to be strictly continuous (*see* **Scales of Measurement**).

The hypothesis to be nullified is that the two random variables have a common distribution. The alternative is that the two distributions differ.

Arrange the combined samples from smallest to largest and identify each observation with its source. As an example, Table 1 gives the ranked algebra achievement scores of two samples of students, one taught by the present method (P), and one by a proposed new method (N).

Table 1 Ranked algebra achievement scores

Score	60	64	67	68	69	70	71	72	73	79	80	84
Method	N	N	P	N	P	N	N	P	N	P	P	P

Count the number of runs of the two sources. Here, the number is eight, beginning with a run of two 'N's, followed by a run of one 'P', and finishing with a run of three 'P's. If the distribution of achievement scores is the same under the two methods of instruction, the scores should be well mixed, leading to many short runs. If the distributions differ, the number of runs will be small. Is eight a small enough number of runs as to be unlikely under the null hypothesis?

The runs test is a permutation or randomization test (*see* **Permutation Based Inference**; **Randomization Based Tests**). The null reference distribution consists of the number of runs for all 924 possible permutations of the observations, six attributed to the present method, and six to the new method. This test is implemented, for example, in the XactStat and SC (`www.mole-soft.demon.co.uk`) packages. The exact Wald–Wolfowitz runs test in SC reports a probability of eight or fewer runs under the null hypothesis at 759/924, approximately .82. No evidence is provided by this test against the null hypothesis.

References

[1] Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Edition., Wiley, New York.

[2] Sprent, P. & Smeeton, N.C. (2001). *Applied Nonparametric Statistical Methods*, 3rd Edition, Chapman & Hall/CRC, London.

[3] Wald, A. & Wolfowitz, J. (1940). On a test whether two samples are from the same population, *Annals of Mathematical Statistics* **11**, 147–162.

CLIFFORD E. LUNNEBORG

Sample Size and Power Calculation

Kevin R. Murphy

Volume 4, pp. 1773–1775

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Sample Size and Power Calculation

If a random sample of N observations is drawn from a population, the precision with which sample statistics (e.g., sample means) estimate corresponding population parameters is determined largely by the number of observations. When population distributions are at least approximately normal, the precision of these estimates can be calculated from simple formulas in which N plays a prominent role.

For example, the formula for the standard error of the mean ($\mathrm{SE_M}$) is

$$\mathrm{SE_M} = \sqrt{\frac{\sigma^2}{N}}. \tag{1}$$

The standard error of the difference between two independent sample means ($\mathrm{SE_{M1-M2}}$) is

$$\mathrm{SE_{M1-M2}} = \sqrt{\frac{\sigma^2{}_1}{N_1} + \frac{\sigma^2{}_2}{N_2}}. \tag{2}$$

If you know, or can estimate, both the level of precision you wish to attain and the variability of scores in the population, it is easy to solve for the sample size needed to attain that level of precision. For example, suppose you would like to be 95% certain that the mean response on a survey that uses a five-point response scale is within 0.25 scale points of the population mean, and your best estimate of the population standard deviation is $\sigma = 0.92$. You can easily rearrange the formula for the standard error of the mean to find the N needed to attain this level of accuracy, using

$$N_{\text{needed}} = 1/[(\text{desired precision level}/1.96)^2/\sigma^2]. \tag{3}$$

If $\sigma = 0.92$, a sample with $N = 52$ (i.e., $1/[(0.25/1.96)^2/0.92^2] = 52$) will provide an estimate of the population mean that achieves the desired level of precision. Similarly, if you would like to be 95% certain that the difference between mean responses in two independent samples (with SDs of 0.90 and 0.85, respectively) is within 0.25 scale points of the population difference, you can rearrange the formula for the standard error of the difference between sample means. In this example, you will need samples of $N = 94$ (i.e., $1/[(0.25/1.96)^2/(0.90^2 + 0.85^2)] = 94$) to attain the desired level of precision.

Hypothesis Testing – Comparisons of Means

The **power** of a statistical test of a null hypothesis is defined as the probability that a test statistic will lead you to reject this null hypothesis when it is in fact false. For example, if you use **analysis of variance** to compare the means in k samples drawn from populations with different means, power is the probability that you will correctly conclude that these means differ. Statistical power is determined by three key parameters: (a) the population difference in means, or the **effect size**, (b) the decision criteria used to define results as statistically significant, and (c) the number of observations [1–5]. Power is highest when there are in fact large differences in population means, when less stringent criteria are used to define statistical significance (e.g., $p < .05$ vs. $p < .01$), and/or when large samples are used.

One of the primary applications of power analysis is to determine the sample size needed to have a reasonable chance of rejecting the null hypothesis. Standards and conventions vary somewhat across fields, but power must usually be substantially above 0.50 to be judged adequate [5], and power levels of 0.80 or above are usually sought [1]. A number of approaches have been suggested for estimating power; models based on the noncentral F distribution [6, 7] can be applied to a wide range of data-analytic techniques [5]. For example, the test statistic used to evaluate differences in sample means in the analysis of variance ($F = \text{MSbetween}/\text{MSwithin}$) is distributed as a noncentral F, where the degree of noncentrality reflects the size of the difference in population means. The null hypothesis that population means are identical would produce a noncentrality parameter of zero. Thus, comparing the observed F with the values obtained from a table of the simple (central) F distribution provides a test of the plausibility of the null hypothesis. The greater the difference in population means (i.e., the larger the value of the noncentrality parameter), the more the population distribution of F values will shift upward, and the greater the likelihood that the obtained F will exceed the critical value of F used to define a

Table 1 Sample size required to attain power $= 0.80 (\alpha = 0.05)$

Effect size: % of variance explained	k = number of means	N = sample size	n_j = number of subjects per cell
0.01	2	777	389
	3	956	318
	4	1077	270
	5	1266	253
0.10	2	74	37
	3	92	31
	4	104	26
	5	116	24
0.20	2	34	17
	3	44	15
	4	51	13
	5	57	11

statistically significant result. Effect sizes are often expressed in terms of statistics such as the standardized mean difference (d) or the percentage of variance accounted for by differences in group means (η^2, or its equivalent, R^2).

Power increases as a nonlinear function of N. As the standard error formulas shown earlier suggest, power functions more closely track the square root of N than the absolute value of N. Table 1 lists the sample size needed to obtain a power of 0.80 for rejecting the null hypothesis as a function of the effect size and the number of means compared (k), given the decision criterion $\alpha = 0.05$. In this table, the effect size is indexed by the proportion of variance accounted for by differences between group means in the population [5]; both the total sample size needed and the number of subjects per cell needed are shown in Table 1 (Tables in Cohen [1] are presented in terms of n_j).

For example, if you are comparing the means of three groups, and you expect that differences between groups account for 10% of the variance in scores in the population, you will need at least $N = 92$ observations to attain power of 0.80. If you expect a smaller effect size, for example that group differences account for 1% of the variance in scores, a much larger sample will be needed to achieve power of 0.80 (here, $N = 956$).

Power analyses in analysis of variance designs with multiple factors (*see* **Factorial Designs**) or repeated measures (*see* **Repeated Measures Analysis of Variance**) follow the same general principle, that power is higher when the population effects are large or when the sample sizes are larger. However, in a complex design, you might have substantially different levels of power for different questions that are asked in the study. In multifactor designs, you will generally have more power when asking questions about main effects (i.e., the simple effects of noise and illumination) than about interactions, but the level of power of each in an analysis of variance model is influenced by both the population effect size and the number of observations that go into calculating each of the sample means to be compared.

Power analyses can be extended to complex multivariate procedures (*see* **Multivariate Analysis: Overview**). Stevens [8] discusses applications of power analysis in **multivariate analysis of variance**. Cohen [1] discusses power analyses for a wide range of statistical techniques. Both sources include extensive tables for estimating power, and these can be used to determine the number of cases needed to reach power of 0.80 (or whatever other convention is used to define adequate power) for the great majority of statistical procedures used in the behavioral and social sciences.

References

[1] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edition, Erlbaum, Hillsdale.

[2] Kraemer, H.C. & Thiemann, S. (1987). *How Many Subjects?* Sage Publications, Newbury Park.

[3] Labovitz, S. (1968). Criteria for selecting a significance level: a note on the sacredness of 05, *American Sociologist* **3**, 220–222.

[4] Lipsey, M.W. (1990). *Design Sensitivity*, Sage Publications, Newbury park.

[5] Murphy, K. & Myors, B. (2003). *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*, 2nd Edition, Erlbaum, Mahwah.

[6] Patnaik, P.B. (1949). The non-central χ^2- and F-distributions and their applications, *Biometrika* **36**, 202–232.

[7] Pearson, E.S. & Hartley, H.O. (1951). Charts of a power function for analysis of variance tests, derived from the non-central F-distribution, *Biometrika* **38**, 112–130.

[8] Stevens, J.P. (1980). Power of multivariate analysis of variance tests, *Psychological Bulletin* **86**, 728–737.

KEVIN R. MURPHY

Sampling Distributions

Sabine Landau

Volume 4, pp. 1775–1779

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Sampling Distributions

A *statistic* is a summary measure that can be calculated from a sample (a subset of a population). A sampling distribution is the probability distribution of a statistic. The goal of *inferential statistics* is to use such summary measures to make inferences about parameters that describe the population from which the sample is drawn. Knowledge of the sampling distribution of particular statistics is essential to derive these inferences, and much effort in statistical science has been devoted to achieving this (for an introduction to statistical inference see, for example, [1] and entries on **Classical Statistical Inference: Practice versus Presentation**, **Deductive Reasoning and Statistical Inference**, and **Neyman–Pearson Inference**). To appreciate the concept of the sampling distribution, we need to first consider the frequentist's model (*see* **Probability: Foundations of**) of how data come about.

Population and Sample

The collection of individuals about whom information is desired is usually referred as the *population* or more specifically the *target population*. For example, we might be interested in the prevalence of depression in a clearly defined clinical population (patients who have been given a certain diagnosis, are within a specified age range, are members of a particular ethnic group, etc.). Since it is impractical and often impossible to measure the outcome of interest (here absence/presence of depression) for every member of the target population, a subset of the population (sample) is selected for study. Statistical inferences depend on the assumption of randomness where a *random sample* of size n is a subset of n of the members from the relevant population where the subset is chosen in such a way that every possible subset of size n has the same chance of being selected as any other.

A *parameter* is a characteristic that describes the target population in the same way a statistic describes a sample. For example, in our depression example, the outcome of interest is a binary variable with possible values '1' (depressed) and '0' (not depressed) and its distribution in the population is characterized completely by the proportion of individuals who have depressive symptoms. (When relating to the population, a proportion is also often referred to as a *probability*.) Where a statistic is used to find out about or as a stand in for a population parameter, it is specifically called an *estimator statistic* or, for short, an *estimator*. If it forms the basis of a *statistical test*, it is known as a *test statistic*. Here, we shall look at statistics in the context of estimation. For example, an intuitive (and as it turns out a 'good', see later) estimator of the proportion of individuals with depression in the target population is the proportion of individuals with such symptoms in the sample. Note that in the frequentist's approach to statistical inference, the population (true) parameter is assumed to be a fixed quantity. In contrast, the estimator statistic is a random quantity since it is a function of the data collected.

Sampling Variability

While sampling individuals saves time and money, the value calculated for the sample will almost certainly differ from the population parameter since not all individuals are sampled; the statistic is said be affected by *sampling error*. For example, if the true prevalence of depression was 20%, and we find that in a sample of 50 patients, 12 had depressive symptoms, then our sample estimate of $12/50 = 24\%$ would be affected by a sampling error of 4%. Each sample might have a different sampling error introducing what is known as *sampling variability* of the statistic. The amount of sampling variability will reduce as the sample size increases and more of the population elements are investigated. To visualize this, we can employ a computer to repeatedly draw random samples of sizes $n = 50$ and $n = 100$ from a population with true prevalence 20% (i.e., from a *Bernoulli distribution* with success probability 0.2 (*see* **Catalogue of Probability Density Functions**) and calculate the proportion of individuals with depression (= successes) for each sample. Figure 1 illustrates the observed sample proportions by means of **histograms**. We see that most proportion estimates are near the true parameter value of 0.2 (say within a distance of 0.1) with the average deviation from the true parameter value being larger for the smaller ($n = 50$) samples.

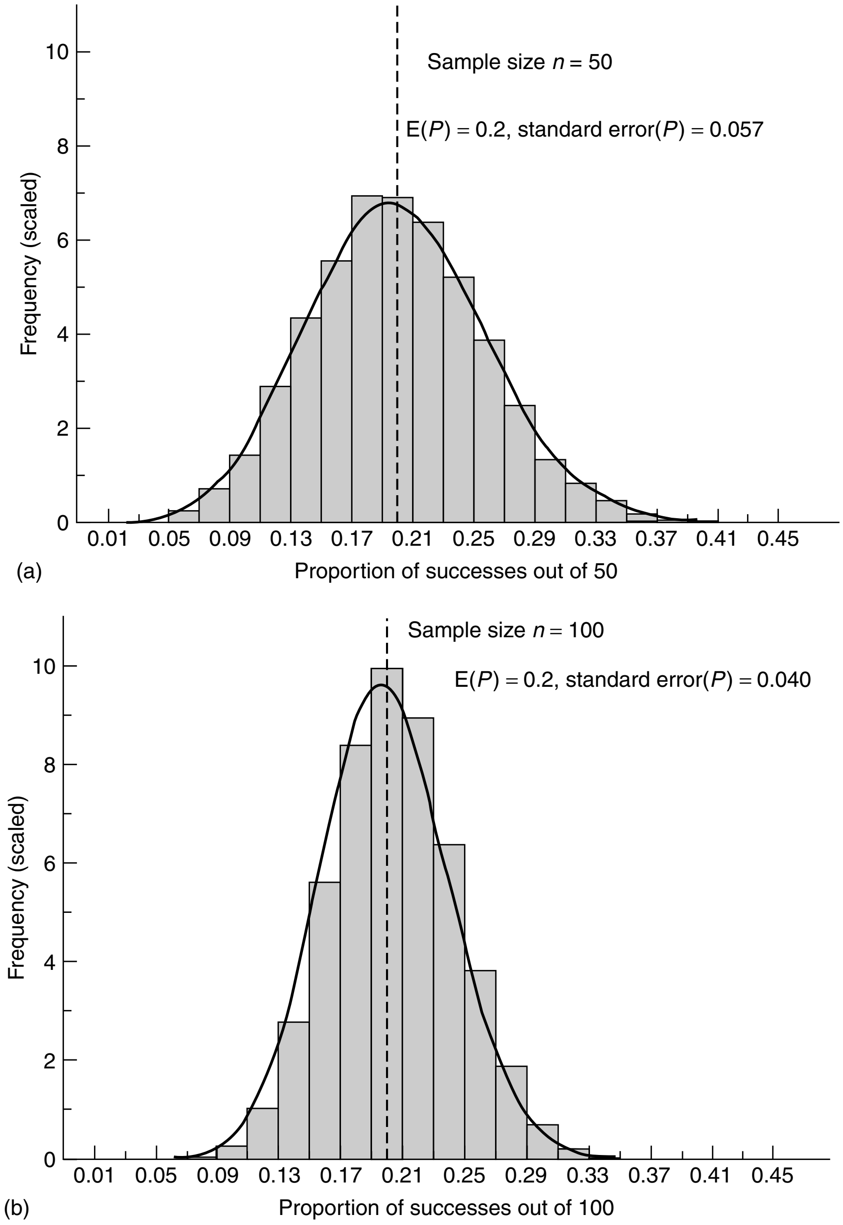


Figure 1 Simulated distribution of sample proportion of successes for two different sample sizes (10 000 simulation runs). The curves show normal density approximations to the histograms

Standard Error and Properties of Estimators

Figure 1 shows the sampling distribution of the proportion of successes based on simulations that mimic the data-generating process. As with any probability distribution, two important characteristics are its mean and variance (or first and second moments (*see* **Expectation**; **Moments**)). The expected value of the sampling distribution is also called the *expected value* or *mean of the statistic*. The standard deviation of the sampling distribution measures the average departure of the estimator from its long-term average and serves to quantify the *precision* of the estimator statistic. To distinguish it from the standard deviation of the population distribution, it is referred to as the *standard error of the statistic* or, for short, the standard error. Note that in contrast to the standard deviation of the population the standard error is affected by sample size. From our simulation, we calculate the expected value of the proportion statistic as $E(P) \approx 0.2$; that is, almost the true parameter 0.2. The value of the standard error of the proportion statistic P_n varies with sample size and we calculate that approximately s.e.$(P_{50}) \approx 0.057$ and s.e.$(P_{100}) \approx 0.040$; that is, if we draw a sample of size $n = 50$ and estimate the proportion, we will on average be 5.7% away from the expected value of the proportion statistic, while if we were to increase the sample size to $n = 100$, our average imprecision would reduce to 4%.

The sampling distribution provides a means by which to compare different estimators. Intuitively good estimators should hit the true population parameter on average. More formally, a desirable property of an estimator is that its expectation should equal the population parameter that it is aiming to estimate, irrespective of what that value is, that is, that it is *unbiased* for the parameter of interest. For example, our simulation suggests that the sample proportion is unbiased for the population proportion. (Unbiasedness can be shown to hold theoretically for any value of the population proportion.) When the estimator is unbiased, its standard error is the square root of the average squared deviation of the sample estimates from the population parameter; or in other words, an average sampling error. The sampling error generated by two alternative unbiased estimators can therefore be compared by their standard errors. Under some circumstances, unbiased estimators can be shown to attain the smallest variance that is possible within a particular probability model (*see* **Information Matrix**; **Estimation**).

The estimators discussed above are more specifically known as *point estimators*. Knowledge of the sampling distribution of relevant statistics also allows the construction of *interval estimators*, which are more commonly referred to as **confidence intervals**. The second strand of classical statistical inference – the testing of hypotheses about the parameters of the population (or populations) at a given *significance level* – also requires knowledge of the sampling distribution of a test statistic under the relevant null hypothesis. Finally, since quality measures of inferential methods such as the standard error of an estimator, the width of a confidence interval, or the *power* of a test, all depend on the sample size, knowledge of the sampling distribution is essential to plan the appropriate size of a study (*see* **Power**).

How to Derive the Sampling Distribution?

Having made the case that the sampling distribution is an essential tool for statistical inference, the question might be asked 'How can the sampling distribution of a statistic be derived from a *single* observed sample?'. After all, the considerations above assumed that we knew the population parameter of interest and could therefore mimic the repeated random sampling from the population. In practice, we do not know the true parameter and we only have one (random) sample. In general, the answer is that we have to specify the probability distribution in the population so that statistical theory or empirical resampling techniques can provide the sampling distribution of a statistic.

Parametric statistical methods assume a parameterized probability distribution in the population. Preferably, such a model assumption should be based on theory or derived empirically from the observed data. However, when theoretical results are not available and the data are sparse, a parameterized shape of the distribution is often simply assumed for convenience. Once a parametric distribution model has been specified, analytic results often facilitate expression of the sampling distribution as a function of the (unknown) population parameters. Alternatively, computing intensive methods such as the *parametric bootstrap* that resample from the estimated population distribution can be used to generate the sampling distribution (*see* **Bootstrap Inference**). In contrast, *nonparametric statistical methods* derive sampling distributions without parameterization of the population distribution. Resampling methods, in particular, the *nonparametric bootstrap*, which resamples from the *empirical* distribution of a sample, are useful in this respect (*see* **Bootstrap Inference**).

The theoretical derivations of sampling distributions can involve considerable amounts of probability theory, and here we just mention two well-known principles for derivation.

1. Let μ denote the population mean of a continuous variable of interest and σ the population standard deviation. If the population has a bell-shaped distribution (a *normal distribution*, *see* **Catalogue of Probability Density Functions**), then the sample mean $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$, where X_i denotes the observation on the ith object in the sample of size n, also has a normal distribution with mean μ and standard error $\sigma/\sqrt{n}$. In other words, when normality can be assumed in the target population, then the sample mean is an unbiased estimator of the population mean and as the sample size n increases, its standard error decreases by the factor $1/\sqrt{n}$. The standard error of the sample mean is commonly estimated from the sample by replacing the population standard deviation in $\sigma/\sqrt{n}$ by a suitable estimator. For a normal population, the sample standard deviation $s = (1/(n-1))\sqrt{\sum_i (X_i - \bar{X})^2}$ is an unbiased estimator for σ.
2. When the sample size is sufficiently large, an important statistical theorem, the **central limit theorem**, states that the sample mean has mean μ and standard error $\sigma/\sqrt{n}$, irrespective of the distributional shape in the population. The larger the sample size, the better the approximation of the sampling distribution of the sample mean by the normal distribution. We can apply the central limit theorem to the depression example, where we were sampling from a population with a binary outcome. Since the proportion of '1s' in a set with binary elements is the same as the arithmetic mean of the '0' and '1' values, the unknown prevalence of depression is the population mean. We can therefore estimate it by the unbiased sample mean (the sample proportion of '1s') and know that the standard error of this estimator is approximately $\sigma/\sqrt{n}$. The population standard deviation is a function of the population distribution and for binary outcomes is known to be $\sigma = \sqrt{(p(1-p))}$, where p denotes the probability of observing a '1'. We can therefore estimate the standard error of the proportion statistic by $\sqrt{(P_n(1-P_n))}/\sqrt{n}$. For example, if 12 out of a sample of 50 subjects showed depressive symptoms, we would estimate the prevalence of depression in the clinical target population as 24% and the standard error of the proportion estimator as $\sqrt{(0.24(1-0.24)/50)} = 0.0604$ or approximately 6%. Comparison of this to the true standard error of the sample proportion of 5.7% shows that the procedure has performed reasonably well.

To conclude, sampling distributions are needed to carry out statistical inferences. They describe the effects of sampling error on statistics as a function of sample size. Frequently encountered sampling distributions are the normal distribution, *Student's t distribution*, the *chi-square distribution* and the *F distribution* (*see* **Catalogue of Probability Density Functions**).

Reference

[1] Howell, D.C. (2002). *Statistical Methods in Psychology*, 5th Edition, Duxbury Press, Belmont.

SABINE LANDAU

Sampling Issues in Categorical Data

Scott L. Hershberger

Volume 4, pp. 1779–1782

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Sampling Issues in Categorical Data

Introduction

A *categorical* variable is one for which the measurement scale consists of a set of categories [5]. Categorical variables may have categories that are naturally ordered (*ordinal* variables), or have no natural order (*nominal* variables). For example, the variable 'health status' with categories 'excellent', 'good', 'satisfactory', and 'poor' is an ordinal variable, as is age with categories 'young', 'middle age', and 'old'. Alternatively, variables such as political party affiliation, with categories 'Democratic', 'Republican', 'Libertarian', and 'Independent', or sex with categories 'male' and 'female' are examples of nominal variables (*see* **Scales of Measurement**).

In most studies with categorical data, the *sampling units* (e.g., people) are classified simultaneously on the levels of the categorical variables. For instance, we might categorize people simultaneously by health status, age, party affiliation, and sex. One particular unit might then be described as a Democratic, young male in good health. The results of cross-classifying the sampling units are frequently arranged as counts in a **contingency table**. The simplest example of a contingency table is the 2 × 2 cross-classification of the sampling units into one of the four cells defined by the two levels of the two variables. When expressed in terms of observed frequencies, a 2 × 2 table might be represented as shown in Table 1.

When expressed in terms of probabilities, a 2 × 2 table might be represented as shown in Table 2.

Table 1 Observed frequencies

f_{11}	f_{12}	|	$f_{1.}$
f_{21}	f_{22}	|	$f_{2.}$
—	—	—	—
$f_{.1}$	$f_{.2}$	|	$f_{..}$

Table 2 Probabilities

p_{11}	p_{12}	|	$p_{1.}$
p_{21}	p_{22}	|	$p_{2.}$
—	—	—	—
$p_{.1}$	$p_{.2}$	|	$p_{..}$

Example of a 2 × 2 Contingency Table

In this example, the results are from an experiment concerned with the association between the true length of a line and the length as perceived by the subjects. Subjects were shown two lines, one line was longer than 12 in. and one line was shorter than 12 in. The subjects were to decide whether the line they were shown was actually longer or shorter than 12 in. The contingency table is shown in Table 3.

The null hypothesis is that a subject's perception of a line's length is not related to its true length. Stated more formally, if correct, this null hypothesis implies that the conditional probability of being in column 1, given that an observation belongs to a known row, is the same for both rows:

$$\frac{p_{11}}{p_{1.}} = \frac{p_{21}}{p_{2.}}. \tag{1}$$

This also implies that

$$\frac{p_{11}}{p_{.1}} = \frac{p_{12}}{p_{.2}}. \tag{2}$$

Taken together, these two equalities result in the **odds ratio ($\boldsymbol{\alpha}$)** (also known as the cross-products ratio) [6]:

$$\alpha = \frac{\left(\dfrac{p_{11}}{p_{12}}\right)}{\left(\dfrac{p_{21}}{p_{22}}\right)} = \frac{p_{11}p_{22}}{p_{12}p_{21}} = 1. \tag{3}$$

When $\alpha > 1$, the two variables are positively associated; when $\alpha < 1$, the two variables are negatively associated. However, odds ratios are not symmetric around one: An odds ratio larger than one by a given amount indicates a smaller effect than an odds ratio smaller than one by the same amount. While the magnitude of an odds ratio is restricted to range between zero and one, it is literally unrestricted above one,

Table 3 Results of perception of line length experiment

	True length		
Perceived length	≤12″	>12″	Total
≤12″	6	1	7
>12″	1	6	7
Total	7	7	14

allowing the ratio to potentially take on any value. If the natural logarithm (ln) of the odds ratio is taken, the odds ratio is symmetric above and below one, with $\ln(1) = 0$.

Under the null hypothesis of independence ($\alpha = 1$), the expected frequency in cell i, j is

$$e_{ij} = \frac{f_{i.}f_{.j}}{f_{..}}. \tag{4}$$

Thus, the χ^2 goodness-of-statistic can be used as a test of independence:

$$\begin{aligned}\chi^2 &= \frac{\sum_i \sum_j (f_{ij} - e_{ij})^2}{e_{ij}} \\ &= \frac{f_{..}(f_{11}f_{22} - f_{12}f_{21})^2}{f_{1.}f_{2.}f_{.1}f_{.2}} \\ &= \frac{14\{(6)(6) - (1)(1)\}^2}{(7)(7)(7)(7)} \\ &= 7.143, \quad df = 1, \quad p = .01. \end{aligned} \tag{5}$$

The odds ratio also indicates a substantial relationship between perceived and true line length:

$$\begin{aligned}\alpha &= \frac{\left(\frac{6}{14}\right)\left(\frac{6}{14}\right)}{\left(\frac{1}{14}\right)\left(\frac{1}{14}\right)} = 35.999, \\ \ln(35.99) &= 3.583. \end{aligned} \tag{6}$$

Subjects were more than three-and-a-half times more likely to correctly judge the line length than not.

Effects of Sampling Method

Statistical inferences concerning a contingency table require knowledge of how the observations were sampled. The theoretical probability distribution that best models how the data were sampled should ideally be identified, although in the case of a contingency table of any dimension, the χ^2 goodness-of-fit test is valid under a wide range of sampling schemes [1, 3]. Nonetheless, it is still useful to explicitly define the method of sampling, if for no other reason than the influence of the sampling model on our interpretation of the data [4]. Therefore, we consider three different sampling methods that have been use to elicit subjects' responses to the line lengths.

Sampling Method 1: The Hypergeometric Distribution

Consider the situation in which there were seven lines longer than 12 in. and seven lines shorter than 12 in. When asked to judge line length, subjects were informed that there would be seven of both types. This constrains $f_{1.}$, $f_{2.}$, $f_{.1}$, and $f_{.2}$ to each equal a specific number, in this case seven. When all of the marginal totals are fixed by design, the underlying distribution of responses is best described by the *hypergeometric* distribution (*see* **Catalogue of Probability Density Functions**).

The hypergeometric distribution is defined as follows for a 2 × 2 contingency table [6]. Given a sample space containing a finite number of elements, suppose that the elements are divided into $K = 4$ mutually exclusive and exhaustive cells, with f_{11} in cell 1, f_{12} in cell 2, f_{21} in cell 3, and f_{22} in cell 4. A sample of $f_{..}$ observations is drawn at random without replacement. Then the probability of a *specific configuration* of a contingency table is given by the hypergeometric probability

$$p(f_{11}, f_{12}, f_{21}, f_{22}; f_{..}) = \frac{f_{1.}!f_{2.}!f_{.1}!f_{.2}!}{f_{..}!f_{11}!f_{12}!f_{21}!f_{22}!}. \tag{7}$$

Using the line length response data, the probability of this contingency table's configuration is

$$p(6, 1, 1, 6; 14) = \frac{7!7!7!7!}{14!6!1!1!6!} = .01. \tag{8}$$

When the sampling model required by the experimental design follows a hypergeometric distribution, the marginal totals are fixed. This type of experimental design is frequently used to test the *homogeneity* of distributions; that is, the distribution of responses is the same across two levels of a variable [7]. In our example, a test of homogeneity would examine whether the probability of correctly identifying the line length category was the same for both line length categories.

Sampling Method 2: The Binomial Distribution

Consider the situation again in which there were seven lines longer than 12 in. and seven lines shorter than 12 in. When asked to judge line length, subjects are *not* informed that there are seven of both types. Although subjects are aware that there will be a total of 14 lines to judge, and that there are two types of lines, no information is given as to the distribution of the lines in the two categories. Thus, no constraints are placed on the specific values of the $f_{1\cdot}$, $f_{2\cdot}$, $f_{\cdot 1}$, and $f_{\cdot 2}$ marginal totals. In this situation, where $f_{\cdot\cdot}$ is fixed, and where each observation can result in one of two choices (i.e., the line is longer or shorter than 12 in.; the probability of selecting either is .5), the probability of the contingency table is given by the *binomial* probability

$$p(f_+; f_{\cdot\cdot}, p) = \frac{f_{\cdot\cdot}!}{f_+!(f_{\cdot\cdot} - f_+)!}(p)^{f_+}(1-p)^{f_{\cdot\cdot}-f_+}, \tag{9}$$

where $f_+ =$ the number of correct classifications [3]. Thus, for our data, the binomial probability of the contingency table is

$$p(12; 14, .5) = \frac{14!}{12!(2)!}(.5)^{12}(.5)^2 = .01. \tag{10}$$

Binomial distributions occur when the number of classes = 2 (the line is longer or shorter than 12 in.), each class having a known probability p of selection. Here, $p = 7/14 = .5$.

Under the same experimental conditions, where subjects are unaware of how many there are of each type of line, but know the types of lines is greater than two, each with a known probability of selection, the contingency table's configuration follows the *multinomial* probability distribution. Thus, the multinomial distribution is a generalization of the binomial distribution.

Sampling Method 3: The Negative Binomial Distribution

Now we consider the situation in which there are still only two types (classes) of lines, but the total number $f_{\cdot\cdot}$ of lines is not fixed. Here, we are interested in the $f_{\cdot\cdot}$ required to successfully judge a certain number (f_+) of lines. Thus, $f_{\cdot\cdot}$ is left free to vary but f_+ is fixed. Let us assume that the line length experiment was stopped when we found that subjects had correctly judged 12 lines. Therefore, 14 judgments were required to produce 12 successful ones. Observations generated from this experimental design follow a *negative binomial* distribution [2, 3] (*see* **Catalogue of Probability Density Functions**):

$$p(f_{\cdot\cdot}; f_+, p) = \frac{(f_{\cdot\cdot}-1)!}{(f_+-1)!(f_{\cdot\cdot}-f_+)!}(p)^{f_+}(1-p)^{f_{\cdot\cdot}-f_+}. \tag{11}$$

The negative binomial probability of our contingency table's specific configuration is

$$p(14; 12, .5) = \frac{(13)!}{(11)!(2)!}(.5)^{12}(.5)^2 = .01. \tag{12}$$

As noted earlier, under all three sampling methods, the hypergeometric, the binomial, and the negative binomial, the probability of our specific contingency table is the same.

References

[1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition, Wiley, New York.

[2] Kudô, A., & Tarumi, T. (1978). 2 × 2 tables emerging out of different chance mechanisms, *Communications in Statistics, Part A – Theory and Methods* **7**, 977–986.

[3] Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables*, Sage Publications, Thousand Oaks.

[4] Pearson, E.S. (1947). The choice of statistical tests illustrated on the interpretation of data classified in a 2 × 2 table. *Biometrika* **34**, 139–167.

[5] Powers, D.A. & Xie, Y. (2000). *Statistical Methods for Categorical Data Analysis*, Academic Press, San Diego.

[6] Simonoff, J.S. (2003). *Analyzing Categorical Data*, Springer, New York.

[7] Wickens, T.D. (1986). *Multiway Contingency Tables Analysis for the Social Sciences*, Erlbaum, Hillsdale.

(*See also* **Measures of Association**)

SCOTT L. HERSHBERGER

Saturated Model

SCOTT L. HERSHBERGER

Volume 4, pp. 1782–1784

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Saturated Model

A *saturated model* contains as many parameters as there are data points, providing a perfect fit to the data [1]. Consider as an example a **log-linear model** fit to a **contingency table**. A loglinear model that specifies all possible main effects and interactions is saturated because the number of parameters equals the number of cells of the table. Saturated models have no residual variance – the *deviance* is zero – and are most useful for comparing the fit of *hierarchically nested models* (*see* **Hierarchical Models**).

A model's *identification status* is an issue especially relevant to the concept of a saturated model. Models can be (a) *under-identified*, (b) *just-identified*, or (c) *over-identified* [4]. Under-identification occurs when not enough relevant data is available to obtain unique parameter estimates. Note that when the degrees of freedom of a model is negative, at least one its parameters is under-identified. Just-identified models are always identified in a trivial way: Just-identification occurs when the number of data elements equals the number of parameter to be estimated. This is the saturated model. If the model is just-identified, a solution can always be found for the parameter estimates that will result in perfect fit – a discrepancy function equal to zero. Over-identification occurs when the number of data points available is greater than that which is needed to obtain a unique solution for all of the parameters. In fact, with an over-identified model, the degrees of freedom are always positive so that model fit can be explicitly tested. An over-identified model also implies that, for at least one of the model parameters, there is more than one model equation that the solution to the parameter must satisfy. The number of additional equations the solution must satisfy is generally referred to as the number of *over-identifying constraints.*

Most of the familiar statistical models within the family of the **Generalized Linear Model** are saturated or just-identified owing to restrictions that are placed on the parameters of the models [3]. Without these restrictions, the models would be under-identified. For example, let us consider the standard **analysis of variance (ANOVA)** model. For $i = 1, \ldots, n$ and $j = 1, \ldots, m$, with m fixed, let

$$y_{ij} \sim \mathrm{N}(\mu_j, \sigma^2),$$

$$\mu_j = \alpha + \theta_j, \qquad (1)$$

where y_{ij} is person i's outcome in treatment j, μ_j is the population mean, θ_j is the difference between the mean of group j and the population mean, and α is the model's intercept. The parameters of interest are α, $\theta_1, \ldots, \theta_m$; however, the model is under-identified. One restriction among several that can be introduced to just-identify the analysis of variance (ANOVA) model is to impose the restriction is that $\sum_{j=1}^{m} \theta_j = 0$. Because $\theta_j = \alpha - \mu_j$, it follows that $\alpha = (1/m)\sum_{j=1}^{m} \mu_j$. Therefore, α represents a constant effect for the population, which is an average of all the cell means, whereas $\theta_j = \mu_j - (1/m)\sum_{j=1}^{m} \mu_j$ represents the deviation of the cell mean μ_j from the average of all of the cell means. It is, therefore, the main effect due to the jth level of the factor. Importantly, the statistical meaning of the parameters α and θ depends on the identification restriction.

A fully specified log-linear model is another example of a saturated, just-identified model. For instance, suppose that we wish to investigate the relationships between two categorical variables, X and Y, where X has I categories and Y has J categories. Then the saturated ('full') loglinear model is

$$\log(m_{ij}) = \lambda + \lambda_i X + \lambda_j Y + \lambda_{ij} XY, \qquad (2)$$

for each combination of the $I \times J$ levels of the m cells, $i = 1, 2, \ldots, I$, and $j = 1, 2, \ldots J$. $\mathrm{Log}(m_{ij})$ is the log of the expected cell frequency of the cases for cell ij in the contingency table; μ is the overall mean of the natural log of the expected frequencies; $\lambda_i X$ is the main effect for variable X, $\lambda_j Y$ is the main effect for variable Y, and $\lambda_{ij} XY$ is the interaction effect for variables X and Y.

In order for the saturated loglinear model be just-identified, constraints must be imposed on the parameters. Several alternative constraint specifications will accomplish this [5]. For example, as in the analysis of variance (ANOVA) model, we may require that the sum of the parameters over all categories of each variable be zero. For a 2×2 table in which two variables, X (two categories) and Y (two categories):

$$\lambda_1 X + \lambda_2 X = 0,$$

$$\lambda_1 Y + \lambda_2 Y = 0,$$

$$\lambda_{11} XY + \lambda_{12} XY = 0,$$

$$\lambda_{21}XY + \lambda_{22}XY = 0,$$
$$\lambda_{11}XY + \lambda_{21}XY = 0, \qquad (3)$$

which implies that

$$\lambda_1 X = -\lambda_2 X,$$
$$\lambda_1 Y = -\lambda_2 Y, \qquad (4)$$

and

$$\lambda_{11}XY = -\lambda_{21}XY = -\lambda_{12}XY = \lambda_{22}XY. \qquad (5)$$

These restrictions result in the estimation of four parameters: One parameter estimated for λ, one parameter estimated for X (corresponding to the first category), one parameter for Y (corresponding to the first category of Y), and one parameter for the interaction of X and Y (corresponding to the first categories of X and Y. The values of the remaining parameters are derived from the four estimated parameters. The model is saturated – just-identified – because the model, which has four cells, has four estimated parameters. The expected cell frequencies will exactly match the observed frequencies. In order to find a more parsimonious model that is explicitly testable, an unsaturated model must be specified that introduces one or more over-identifying constraints on the parameter estimates. Such a model has degrees of freedom greater than zero, and can be achieved by setting some of the effect parameters to zero.

For example, if both categorical variables are mutually independent, then the following *independence model* describes the relationship between X and Y:

$$\log(m_{ij}) = \lambda + \lambda_i X + \lambda_j Y. \qquad (6)$$

Further, we may decide that Y is not a significant predictor of the cell frequencies:

$$\log(m_{ij}) = \lambda + \lambda_i X, \qquad (7)$$

or alternatively, that X is not:

$$\log(m_{ij}) = \lambda + \lambda_j Y. \qquad (8)$$

Conceivably, neither X nor Y may be useful, resulting in the most restricted baseline model:

$$\log(m_{ij}) = \lambda. \qquad (9)$$

The examples of unsaturated loglinear models given above are hierarchically nested models. Hierarchical models include all lower terms composed from variables in the highest terms in the model. Therefore, the model

$$\log(m_{ij}) = \lambda + \lambda_i X + \lambda_{ij} XY \qquad (10)$$

would not be considered a nested model – for the $\lambda_{ij}XY$ term to be present in the model, both of its constituent variables must be as well. The choice of a preferred model is typically based on the formal comparison of goodness-of-fit statistics associated with hierarchically nested models: the likelihood ratios and/or deviances of nested models are compared to determine whether retaining the more parsimonious model of the two results in a significant decrement in the fit of the model to the data [2]. A significant decrement in fit implies that the expected frequencies generated by the more parsimonious model are significantly less close to the observed frequencies.

References

[1] Agresti, A. (1996). *An Introduction to Categorical Data Analysis*, Wiley, New York.

[2] Everitt, B.S. (1992). *The Analysis of Contingency Tables*, 2nd Edition, Chapman & Hall, Boca Raton.

[3] Green, S.B., Marquis, J.G., Hershberger, S.L., Thompson, M.S. & McCollam, K.M. (1999). The overparameterized analysis of variance model, *Psychological Methods* **4**, 214–233.

[4] Hershberger, S.L. (2005). The problem of equivalent structural models, in *A Second Course in Structural Equation Modeling*, G.R. Hancock & R.O. Mueller, eds, Information Age Publishing, Greenwich.

[5] Wickens, T.D. (1986). *Multiway Contingency Tables Analysis for The Social Sciences*, Lawrence Erlbaum, Hillsdale.

Scott L. Hershberger

Savage, Leonard Jimmie

Sandy Lovie

Volume 4, pp. 1784–1785

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Savage, Leonard Jimmie

Born: November 20, 1917, in Michigan, USA.
Died: November 1, 1971, in Connecticut, USA.

Although L.J. Savage is not a name that many psychologists would instantly recognize, his achievement in producing a coherent system for Bayesian thinking and inference lay behind much of the 1960s movement in psychology to replace classical **Neyman Pearson inference** with its Bayesian equivalent (*see* **Bayesian Statistics**) (see [2], for both an instance of the approach and one of its early key documents, and the classic textbook by [4]). This also led to a lively experimental program in behavioral decision making founded in part on Bayesian inferential premises, one which was even taken up by **Tversky** and Kahneman, the inventors of *Biases and Heuristics*, with their long-running investigation into so-called base rate problems using the Green and Blue Taxi Cab story and variants as illustrative material (see [3] for an overview of the early material). Of course, the motivation for much of the experimental work in psychology was to see how far human decisions approximated the normative ones prescribed by Bayes theory, but there was also a commitment to Bayes theory itself as a better way of making inferences in the face of uncertainty, the avowed aim of classical statistical inference, which leads us back to Savage.

He was educated initially at the University of Michigan's Ann Arbor campus and received a B.S. in mathematics in 1938, with a Ph.D. in 1941 on an aspect of pure mathematics, specifically differential geometry. During his year at Princeton's Institute of Advanced Study (1941–1942) he came to the attention of John von Neumann, whose own work on the theory of games became a later inspiration for Savage. Neumann encouraged Savage to turn to statistics, and he joined the Statistical Research Group at Columbia University in 1943. It was during his tenure at the University of Chicago that he wrote his most influential book *The Foundation of Statistics* [5]. Somewhat in the same frame of mind as Freud who, in his *Project for a Scientific Psychology* of 1895, attempted to characterize his emerging ideas about psychoanalysis in terms of current medical and physiological knowledge, Savage first of all laid out his ideas on personal probability and utility and then attempted to reinterpret classical statistical inference using these new concepts. Again, like Freud, he admitted defeat over the project, but only in the preface to the second edition! The book also shows the significance of **Bruno de Finetti**'s work on subjective probability and his key notion of *exchangeability* to the power and novelty of the approach. From a realization of the drawbacks of all nonpersonalistic orientations to probability and inference, Savage gradually developed an alternative, Bayesian form of inference, where the personally assessed *prior* probabilities of events are transformed into probabilities *a posteriori* in the light of new information, also personally assessed. The implications of this new form of inference, and the statistical novelties that flow from it, have not yet been fully realized or worked through and are likely to absorb the energies of statisticians for some time to come.

For Savage himself, the *Foundations* represented something of a high point, although he was to publish an important book with Dubins, which recast gambles as a form of stochastic or probabilistic process [1]. He also moved to the University of Michigan in 1960, and then finally to Yale where he stayed until his comparatively early death at the age of 53. Savage is important not only for his own work but also because he introduced American (and British) statistics to de Finetti and hence to the possibility of an alternative and powerful form of statistical modeling and inference.

References

[1] Dubins, L.E. & Savage, L.J. (1965). *How to Gamble if you Must: Inequalities for Stochastic Processes*, McGraw-Hill, New York.

[2] Edwards, W., Lindman, H. & Savage, L.J. (1963). Bayesian statistical inference for psychological research, *Psychological Review* **70**(3), 193–242.

[3] Kahneman, D., Slovic, P. & Tversky, A., eds (1982). *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge.

[4] Phillips, L.D. (1973). *Bayesian Statistics for Social Scientists*, Nelson, London.

[5] Savage, L.J. (1954, 1972). *The Foundation of Statistics*, Wiley, New York.

SANDY LOVIE

Scales of Measurement

Karl L. Wuensch

Volume 4, pp. 1785–1787

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Scales of Measurement

The behavioral scientist who speaks of 'scales of measurement' almost certainly is thinking of the hierarchy of measurement scales proposed by psychophysicist **S. S. Stevens** [5.] A scale was said to be created by specifying the rules by which numbers are assigned to objects or events in such a way that there will be a one-to-one correspondence between some of the properties of the measured things and some of the properties of the measurements. Stevens defined four different types of scales, nominal, ordinal, interval, and ratio, based on the extent to which empirical relationships among the measured objects or events correspond to numerical relationships among the measurements (*see* **Measurement: Overview**).

If the measurement rules are such that the measurements can be used to establish the equivalence of objects with respect to the measured characteristic, then the scale is said to be nominal. As an example of a nominal scale, consider the assignment of the number 1 to all pets that are cats, 2 for dogs, and 3 for other types of pets.

If, in addition to the ability to establish equivalence, the rules allow one to establish order of the measured characteristic, then the scale is considered to be an ordinal scale. In Table 1, imagine that each 'o' in the first row represents a bead and that each of the beads is identical in mass. The numbers in the rows below represent measurements of the mass of the strings of beads. For Scale O, the order of the assigned numbers is identical to the order of the objects' masses, so the scale is ordinal.

Notice that Scale O does not allow one to establish equivalence of differences. Objects A and B differ in mass by the same amount as objects B and C, but the difference in the numerical measurements for objects A and B is not the same as that for B and C. With Scale I, one can determine equivalence of differences. The difference in mass between objects A and B is equal to the difference in mass between objects B and C, just as the difference between the measurements of objects A and B is equal to the difference in measurements between objects B and C. A scale that allows one to establish equivalence, order, and equivalence of differences is called an *interval scale*.

Table 1 Illustration of Stevens' Measurement Rules

Object	A	B	C	D	E
Beads		o	oo	oooo	oooooooo
Scale O	−1.2	0	2	3	6
Scale I	−2	0	2	6	14
Scale R	0	2	4	8	16

Scale I allows one to establish equivalence of differences but not of ratios. Object D has twice the mass of object C, and object E has twice the mass of the object D, but the ratio of the measurements for objects D/C does not equal that for objects E/D. Equivalence of ratios can, however, be determined with Scale R. A scale that allows one to establish equivalence, order, equivalence of differences, and equivalence of ratios is called a *ratio scale*.

Stevens opined that the four scales are best characterized by the types of transformations that can be applied to them without distorting the structure of the scale. With a nominal scale, one can substitute any new set of numbers (or other symbols) for the old set without destroying the ability to establish equivalence. For example, instead of using '1', '2', and '3' to represent cats, dogs, and other types of pets, we could use 'A', 'B', and 'C'. With an ordinal scale, one can apply any order-preserving transformation (such as ranking) and still have an ordinal scale. With an interval scale, only a positive linear transformation will produce another interval scale. With a ratio scale, only multiplying the measurements by a positive constant will produce another ratio scale.

Stevens's scales of measurement can be thought of in terms of the nature of the relationship between the observed measurements and the true scores (the true amounts of the measured characteristic, that is, scores on the underlying construct or **latent variable**). Winkler and Hays [8] presented the following list of criteria used to determine scale of measurement:

1. Two things will receive different measurements only if they are truly nonequivalent on the measured characteristic.
2. Object A will receive a larger score than does object B only if object A truly has more of the measured characteristic than does object B. This will be the case when the measurements are related to the true scores by a positive monotonic function.
3. Where T_i represents the true amount of the measured characteristic for object i, and M_i

represents the score obtained when measuring object i, $M_i = a + bT_i$, $b > 0$. That is, the measurements are a positive linear function of the true scores.

4. $M_i = a + bT_i$, $b > 0$, $a = 0$. That is, the measurements are a positive linear function of the true scores, and the intercept is zero. The zero intercept is often called a '*true zero point*' – an object that receives a score of zero has absolutely none of the measured attribute.

To be ratio, a scale must satisfy all four of the criteria listed above; to be interval, only the first three; to be ordinal, only the first two; to be nominal, only the first.

In addition to defining four scales of measurement, Stevens argued that the type of statistics that are 'permissible' on a set of scores is determined, in part, by the scale of measurement. Consider Stevens's recommendations regarding measures of location. The mode is permissible for any scale, even the nominal scale. If there are 20 cats, 15 dogs, and 12 pets of other types in the shop, cats are the modal pet whether we represent cats, dogs, and others with the numerals 1, 2, and 3, or 1, 4, and 9, or any other three numerals. The median is permissible only for scales that are at least ordinal. It does not matter whether we measure the mass of the pets in the shop in grams, kilograms, or simple ranks – the computed median will represent the same amount of mass with any positive monotonic transformation of the true masses. The **mean** is permissible only for scales that are at least interval. Imagine that we have five pets, named A, B, C, D, and E. Their true masses are 1, 2, 3, 6, and 18, respectively, for a mean of 6. Pet D has a mass exactly equal to the mean mass. Suppose our interval data for these pets is defined by $M = 10 + 2T$. The observed scores are 12, 14, 16, 22, and 46, respectively, for a mean of 22. Again, pet D has a mass exactly equal to the mean mass. Now suppose we have ordinal data, such as simple ranks. The observed scores are 1, 2, 3, 4, and 5, respectively, for a mean of 3. Now it is pet C that appears to have a mass exactly equal to the mean mass, but that is not true.

Stevens's belief that the scale of measurement should be considered when choosing which statistical analysis to employ (the measurement view) was embraced by some and rejected by others. Some of the former authored statistics texts that taught social scientists to consider scale of measurement of great importance when selecting an appropriate statistical analysis [2]. Most controversial was the suggestion that parametric statistics require at least interval level data but that nonparametric statistics were permissible with ordinal data. Many statisticians attacked the measurement view [2, 7], while others defended it [3, 6]. Those opposed to the measurement view argued that the only assumptions necessary when using parametric statistics are mathematical, such as normality and homogeneity of variance. Those favoring the measurement view argued that behavioral researchers are interested in drawing conclusions about underlying constructs, not just observed variables, and accordingly they must consider the scale of measurement, that is, the nature of the relationship between true scores and observed scores.

Imagine that we are interested in testing the hypothesis that the mean aggressiveness of cats is identical to the mean aggressiveness of dogs and that this hypothesis is absolutely true with respect to the latent variable. If the relationship between our measurements and the true scores is not linear, the population means on our measurement variable may well not be equivalent. Accordingly, testing hypotheses about the means of latent variables seems more risky with noninterval data than with interval data. The real fly in the ointment here is that one never really knows with certainty the nature of the relationship between the latent variable and the measured variable – for my example, what is the nature of the function relating common measures of animal aggressiveness with the 'true' amounts of aggressiveness? How can one ever know with confidence if such measurements represent interval data or not? In some circumstances, one need not worry about whether or not the data are interval. If one assumes normality and homogeneity of variance, then differences in the means of an observed variable do indicate that the means on the latent variable also differ, regardless of whether the measurements are interval or merely ordinal [1].

One's attitude to the relationship between scale of measurement and the choice of an appropriate statistical analysis may be determined by one's more basic ideas about the nature of measurement [4]. Someone who believes that useful measurements are those that capture interesting empirical relationships among the measured objects or events (representational theory) will argue that scale of measurement is an important characteristic to consider when choosing a statistical analysis, at least when conclusions are to be scale-free. The operationist, by contrast, believes that measurements are always scale-specific, and thus choice of statistical analysis is unrelated to scale of measurement.

Ultimately, one's decision about whether a set of data represents interval or merely ordinal measurement is largely a matter of faith. When one counts number of bites, aggressive postures, and submissive postures of fighting mice and combines them into a composite measure of aggressiveness, what is the nature of the relationship between these measurements and the 'true' amounts of aggressiveness displayed by these animals in 'concrete reality'? Is the relationship linear or not? How could one ever answer such a metaphysical question with certainty? One way to evade this dilemma is to treat reality as being constructed or invented rather than discovered and then argue that the results of the parametric statistical analysis apply only to that 'abstract reality' which is a linear function of our measurements. Such a defining of 'reality' in terms of the observed variables is not much different from a similar device often employed by applied statisticians, defining a population from a sample – the population for which these statistical inferences are made is that population for which this sample could be considered to be a random sample.

References

[1] Davison, M.L. & Sharma, A.R. (1988). Parametric statistics and levels of measurement, *Psychological Bulletin* **104**, 137–144.

[2] Gaito, J. (1980). Measurement scales and statistics: resurgence of an old misconception, *Psychological Bulletin* **87**, 564–567.

[3] Maxwell, S.E. & Delaney, H.D. (1985). Measurement and statistics: an examination of construct validity, *Psychological Bulletin* **97**, 85–93.

[4] Michell, J. (1986). Measurement scales and statistics: a clash of paradigms, *Psychological Bulletin* **100**, 398–407.

[5] Stevens, S.S. (1951). Mathematics, measurement, and psychophysics, in *Handbook of Experimental Psychology*, S.S. Stevens, ed., Wiley, New York, pp. 1–49.

[6] Townsend, J.T. & Ashby, F.G. (1984). Measurement scales and statistics: the misconception misconceived, *Psychological Bulletin* **96**, 394–401.

[7] Velleman, P.F. & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading, *The American Statistician* **47**, 65–72.

[8] Winkler, R.L. & Hays, W.L. (1975). *Statistics: Probability, Inference, and Decision*, 2nd Edition, Holt Rinehart & Winston, New York.

KARL L. WUENSCH

Scaling Asymmetric Matrices

Yoshio Takane

Volume 4, pp. 1787–1790

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Scaling Asymmetric Matrices

Tables with the same number of rows and columns are called square tables. In square tables, corresponding rows and columns often represent the same entities (objects, stimuli, variables, and so on). For example, the ith row of the table represents stimulus i, and the ith column also represents the same stimulus i. Let x_{ij} denote the element in the ith row and the jth column of the table. (We often call it ijth element of the table.) We use X (in matrix form) to denote the entire table collectively. Element x_{ij} indicates (the strength of) some kind of relationship between the row entity (stimulus i) and the column entity (stimulus j). Square tables, in which $x_{ij} \neq x_{ji}$ for some combinations of i and j, are called asymmetric tables. In matrix notation, this is written as $X' \neq X$, where X' indicates the transpose of X.

Asymmetric tables arise in a number of different guises. In some cases, the kind of relationship represented in the table is antisymmetric. For example, suppose you have a set of stimuli, and you ask a group of subjects whether they prefer stimulus i or j for each pair of stimuli. Since j cannot be preferred to i if i is preferred to j, the preference choice constitutes an antisymmetric relationship. Let x_{ij} denote the number of times i is preferred to j. Tables representing antisymmetric relationships are usually asymmetric. These types of tables are often skew-symmetric, or can easily be turned into one by a simple transformation (e.g., $y_{ij} = \log(x_{ij}/x_{ji})$). In the skew symmetric table, $y_{ji} = -y_{ij} (Y' = -Y)$. Skew symmetric data, such as the one just described, are often represented by the difference between the preference values of the two stimuli involved. Let u_i represent the preference value of stimulus i. Then, $y_{ij} = u_i - u_j$. Case V of Thurstone's law of comparative judgment [8], and Bradley-Terry-Luce (BTL) model [1, 7] are examples of this class of models. The scaling problem here, is to find estimates of u_i's, given a set of observed values of y_{ij}'s.

Here is an example: The top panel of Table 1 gives observed choice probabilities among four music composers, labeled as B, H, M and S. Numbers in the table indicate the proportions of times row composers are preferred to column composers. Let us apply the BTL model to this data set. The second panel of the table gives the skew symmetric table obtained by applying the transformation, $y_{ij} = \log(x_{ij}/x_{ji})$, to the observed choice probabilities. Least squares estimates of preference values for the four composers are obtained by row means of this skew symmetric table. B is the most preferred, M the second, then S, and H the least. Something similar can also be done with Thurstone's Case V model. The only difference it makes is that normal quantile (deviation) scores are obtained, when the matrix of the observed choice probabilities is converted into a skew symmetric matrix. The rest of the procedure remains essentially the same as in the BTL model.

Table 1 The BTL model applied to preference choice data involving four music composers

	Observed choice probabilities			
	B	H	M	S
B	.500	.895	.726	.895
H	.105	.500	.147	.453
M	.274	.853	.500	.811
S	.105	.547	.189	.500
	Matrix of $y_{ij} = \log(x_{ij}/x_{ji})$			
B	0	2.143	.974	2.143
H	−2.143	0	−1.758	−.189
M	−.974	1.758	0	1.457
S	−2.143	.189	−1.457	0
	Estimated preference values			
	1.315	−1.022	.560	−0.853

Asymmetric tables can also arise from proximity relationships, which are often symmetric. In some cases, they exhibit asymmetry, however. For example, you may ask a group of subjects to identify the stimulus presented out of n possible stimuli, and count the number of times stimulus i is 'confused' with stimulus j. This is called stimulus recognition (or identification) data, and it is usually asymmetric. There are a number of other examples of asymmetric proximity data such as mobility tables, journal citation data, brand loyalty data, discrete panel data on two occasions, and so on. In this case, the challenge is in explaining the asymmetry in the tables.

A variety of models have been proposed for asymmetric proximity data. Perhaps the simplest model is the quasi-symmetry model (*see* **Quasi-symmetry in Contingency Tables**). The quasi-symmetry is characterized by $x_{ij} = a_i b_j c_{ij}$, where a_i and b_j are row and column marginal effects, and $c_{ij} = c_{ji}$ indicates a symmetric similarity between i and j. This model postulates that after removing the marginal effects, the remaining relation is symmetric. (The special case, in which $a_i = b_i$ for all i, leads to a full symmetric model.) The quasi-symmetry also satisfies the cycle condition stated as $x_{ij}x_{jk}x_{ki} = x_{ji}x_{kj}x_{ik}$. In some cases, the symmetric similarity parameter, c_{ij}, may further be represented by a simpler model, $c_{ij} = \exp(-d_{ij})$, or $c_{ij} = \exp(-d_{ij}^2)$, where d_{ij} is the Euclidean distance between stimuli i and j, represented as points in a multidimensional space.

DEDICOM (DEcomposing DIrectional COMponents, [4]) attempts to explain asymmetric relationships between n stimuli by a smaller number of asymmetric relationships. The DEDICOM model is written as $X = ARA'$, where R is a square asymmetric matrix of order r (capturing asymmetric relationships between r components, where r is assumed much smaller than n), and A is an n by r matrix that relates the latent asymmetric relationships among the r components to the observed asymmetric relationships among the n stimuli. Several algorithms have been developed to fit the DEDICOM model. To illustrate, the DEDICOM model is applied to a table of car switching frequencies among 16 types of cars [4]. (This table indicates frequencies with which a purchase of one type of car is followed by a purchase of another type by the same consumer.) Table 2 reports the analysis results [5]. Labels of the 16 car types consist of two components. The first three characters mainly indicate size (SUB = subcompact, SMA = small specialty, COM = compact, MID = midsize, STD = standard, and LUX = luxury), and the fourth character indicates mainly origin or price (D = domestic, C = captive imports, I = imports, L = low price, M = medium price, and S = specialty). The top portion of the table gives the estimated A matrix (normalized so that $A'A = I$), from which we may deduce that the first component (dimension) represents plain large and mid-size cars, the second component represents fancy large cars, and the third represents small/specialty cars. The bottom portion of the table represents the estimated R matrix that captures asymmetry relationships among the three components. There are more switches from 1 to 3, 1 to 2, and 2 to 3 than the other way round. This three-component DEDICOM model captures 86.4% of the total SS (sum of squares) in the original data.

Table 2 DEDICOM applied to car switching data

Matrix A			
	Dimension		
Car Class	1	2	3
SUBD	.13	−.02	.36
SUBC	.02	.00	.03
SUBI	.03	.01	.30
SMAD	.01	.03	.53
SMAC	.00	.00	.00
SMAI	.00	.01	.09
COML	.24	−.11	.17
COMM	.10	−.01	.06
COMI	.02	.00	.03
MIDD	.54	.00	.12
MIDI	.02	.00	.02
MIDS	.09	.24	.58
STDL	.68	−.08	−.18
STDM	.32	.67	−.27
LUXD	−.23	.69	.05
LUXI	.00	.02	.01
Matrix R (divided by 1000)			
dim. 1	127	57	78
dim. 2	26	92	23
dim. 3	17	12	75

Any asymmetric table can be decomposed into the sum of a symmetric matrix (X_s), and a skew symmetric matrix (X_{sk}). That is, $X = X_s + X_{sk}$, where $X_s = (X + X')/2$, and $X_{sk} = (X - X')/2$. The two parts are often analyzed separately. X_s is often analyzed by a symmetric model (such as the inner product model or a distance model like those for c_{ij} described above). X_{sk}, on the other hand, is either treated like a skew symmetric matrix arising from an antisymmetric relationship, or by CASK (Canonical Analysis of SKew symmetric data, [3]). The latter decomposes X_{sk} in the form of AKA', where K consists of 2 by 2 diagonal blocks of the form $\begin{pmatrix} 0 & k_l \\ -k_l & 0 \end{pmatrix}$ for the lth block. This representation can be analytically derived from the singular value decomposition of X_{sk}.

Generalized GIPSCAL [6] and HCM (Hermitian Canonical Model, [2]) analyze both parts (X_s and X_{sk}) simultaneously. The former represents X by $B(I_r + K)B'$ (where the BB' part represents X_s and the BKB' part represents X_{sk}), under the assumption that the skew symmetric part of R (that is,, $(R - R')/2$) in DEDICOM is positive definite. The HCM first forms an hermitian matrix, H, by $H = X_s + iX_{sk}$ (where i is a symbol for an imaginary number, $i = \sqrt{-1}$), and obtains the eigenvalue-vector decomposition of H.

References

[1] Bradley, R.A. & Terry, M.E. (1952). The rank analysis of incomplete block designs. I. The method of paired comparisons, *Biometrika* **39**, 324–345.

[2] Escoufier, Y. & Grorud, A. (1980). in *Data Analysis and Informatics* E. Diday, L. Lebart, J.P. Pages & Tomassone, eds, North Holland, Amsterdam, pp. 263–276.

[3] Gower, J.C. (1977). The analysis of asymmetry and orthogonality, in J.R. Barra, F. Brodeau, G. Romier & B. van Cuten, eds, *Recent Developments in Statistics*, North Holland. Amsterdam, pp. 109–123.

[4] Harshman, R.A. Green, P.E. Wind, Y. & Lundy, M.E. (1982). A model for the analysis of asymmetric data in marketing research, *Marketing Science* **1**, 205–242.

[5] Kiers, H.A.L. & Takane, Y. (1993). Constrained DEDICOM, *Psychometrika* **58**, 339–355.

[6] Kiers, H.A.L. & Takane, Y. (1994). A generalization of GIPSCAL for the analysis of nonsymmetric data, *Journal of Classification* **11**, 79–99.

[7] Luce, R.D. (1959). *Individual Choice Behavior: a Theoretical Analysis*, Wiley, New York.

[8] Thurstone, L.L. (1927). A law of comparative judgment, *Psychological Review* **34**, 273–286.

(*See also* **Multidimensional Scaling**)

YOSHIO TAKANE

Scaling of Preferential Choice

Ulf Böckenholt

Volume 4, pp. 1790–1794

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Scaling of Preferential Choice

Choice data can be collected by either observing choices in the daily context of the decision makers or by asking a person directly to state their preferences for a single or multiple sets of options. Both data types, which are referred to as revealed and stated preference data (Louviere et al. [5]), may yield similar outcomes. For instance, in an election votes for political candidates represent revealed choice data. Rankings of the same candidates in a survey shortly before the election are an example for stated choice data. The stated preference data may prove useful in predicting the election outcome and in providing more information about the preference differences among the candidates than would be available from the election results alone.

Scaling models serve the dual purpose to summarize stated and revealed choice data and to facilitate the forecasting of choices made by decision makers facing possibly new or different variants of the choice options (Marshall & Bradlow [9]). The representation determined by the scaling methods can provide useful information for identifying both option characteristics that influence the choices and systematic sources of individual differences in the evaluation of these option characteristics. For example, in the election study, voters may base their decision on a set of attributes (e.g., integrity, leadership) but differ in the weights they assign to the attribute values for the different candidates. An application of scaling models to the voting data may both reveal these attributes and provide insights about how voters differ in their attribute assessments of the candidates.

The relationship between the attribute values and the corresponding preference judgments may be monotonic or nonmonotonic. Thus, decision makers may assess 'higher' attribute values as more favorable than 'lower' ones (e.g., quality of a product), or they may prefer a certain quantity of an attribute and dislike deviations in either directions from it (e.g., sweetness of a drink). In the latter case, individuals choose the option that is closest to their 'ideal' option where closeness is a function of the distance between the choice option and the person – specific ideal (De Leuuw [2]). Distance between choice options may be defined in various ways. Applications in the literature include approaches based on the Euclidean measure in a continuous attribute space and tree structures in which both choice and ideal options are represented by nodes (Carroll & DeSoete [1]). In either case, the interpretation of individual preference differences is much simplified provided decision makers use the same set of attributes in assessing the choice options.

Thurstone's [14] random-utility approach has been highly influential in the development of many scaling models. Noting that choices by the same person may vary even under seemingly identical conditions, Thurstone argued that choices could be described as realizations of random variables that represent the options' effects on a person's sensory apparatus. According to this framework, a choice of an option i by decision maker j is determined by an unobserved utility assessment, υ_{ij}, that can be decomposed into a systematic and a random part: $\upsilon_{ij} = \mu_{ij} + \epsilon_{ij}$. The person-specific item mean μ_{ij} is assumed to stay the same in repeated evaluations of the item but the random contribution ϵ_{ij} varies from evaluation to evaluation according to some distribution. Thurstone (1927) postulated that the ϵ_{ij}'s follow a normal distribution. The assumption that the ϵ_{ij}'s are independently Gumbel or Gamma distributed leads to scaling models proposed by Luce [6] and Stern [12]), respectively. Marden [8] and Takane [13] provide general discussions of these different specifications.

According to the latent utility framework, choosing the most preferred option is equivalent to selecting the option with the largest utility. Thus, an important feature of this choice process is that it is comparative in nature. A selected option may be the best one out of a set of available options but it may be rejected in favor of other options that are added subsequently to the set of options. Because choices are inherently comparative, the origin of the utility scale cannot be identified on the basis of the choices alone. One item may be preferred to another one but this result does not allow any conclusions about whether either of the items are attractive or unattractive.

One may question the use of randomness as a device to represent factors that determine the formation of preferences but are unknown to the observer. However, because, in general, it is not feasible to identify or measure all relevant choice determinants (such as all attributes of the choice options of the person choosing, or of environmental factors), it is not possible to answer conclusively the question of whether the choice process is inherently random or is determined by a multitude of different factors. Fortunately, for the development of scaling models this issue is not critical because either position arrives at the same conclusion that choices are described best in terms of their probabilities of occurrence (Manski & McFadden [7]).

In recent years, a number of scaling models have been developed that by building on Thurstone's random-utility approach take into account systematic time and individual-difference effects (Keane [4]). Concurrently, with these developments experimental research in judgment and decision making demonstrated that choice processes are subject to many influences that go beyond the simple evaluations of items. For example, different framings of the same choice options may trigger different associations and evaluations with the results that seemingly minor changes in the phrasing of a question or in the presentation format can lead to dramatic changes in the response behavior of a person (Kahneman [3]). One major conclusion of this research is that the traditional assumption of respondents having well-defined preferences should be viewed as a hypothesis that needs to be tested as part of any modeling efforts.

Preference Data

Typically, stated choice data are collected in the form of incomplete and/or partial rankings. Consider a set of J choice alternatives $(j = 1, \ldots, J)$ and n decision makers $(i = 1, \ldots, n)$. For each decision maker i and choice alternative j, a vector $\mathbf{x}_{ij}$ of observed variables is available that describe partially the pair (i, j). Incomplete ranking data are obtained when a decision maker considers only a subset of the choice options. For example, in the method of paired comparison, two choice options are presented at a time, and the decision maker is asked to select the more preferred one. In contrast, in a partial ranking task, a decision maker is confronted with all choice options and asked to provide a ranking for a subset of the J options. For instance, in the best–worst method, a decision maker is instructed to select the best and worst options out of the offered set of choice options. Both partial and incomplete approaches can be combined by offering multiple distinct subsets of the choice options and obtain partial or complete rankings for each of them. For instance, a judge may be presented with all $\binom{J}{2}$ option pairs sequentially and asked to select the more preferred item in each case.

Presenting choice options in multiple blocks has several advantages. First, the judgmental task is simplified since only a few options need to be considered at a time. Second, it is possible to investigate whether judges are consistent in their evaluations of the choice options. For example, if options are presented in pairs, one can investigate whether respondents are transitive in their comparisons, that is, whether they prefer j to l when they choose j over k and k over l. Third, obtaining multiple judgments from each decision maker simplifies analyses of how individuals differ in their preferences for the choice options. Individual-difference analyses are discussed in more detail in the next section. These advantages need to be balanced with possible boredom and learning effects that may affect a person's evaluation of the choice options when the number of blocks is large.

Revealed choice data differ from stated choice data in a number of ways. Perhaps, most importantly the set of choice alternatives may be unknown and may vary among decision makers in systematic ways. The lack of knowledge of the considered choice set complicates any inferences about the relative advantages of the selected options. Moreover, only top choices are observed typically which provide little information about the nonchosen options. Finally, the timing and context of the revealed choices may vary from person to person which reduces the interindividual comparability of the results. For these reasons, it is useful frequently to combine revealed with stated choice data to obtain a richer and more informative understanding of the underlying preferences of the decision makers.

Thurstonian Models for Preference Data

The choices made by person i for a single choice set can be summarized by an ordering vector $\mathbf{r}_i$. For instance, $\mathbf{r}_i = (h, j, \ldots, l, k)$ indicates that choice option h is judged superior to option j which in turn is judged superior to the remaining options, with the least preferred option being k. The probability of observing this ordering vector can be written as

$$\begin{aligned}\Pr(\mathbf{r}_i = (h, j, \ldots, l, k)|\boldsymbol{\xi}) \\ = \Pr[(\upsilon_{ih} - \upsilon_{ij} > 0) \cap \ldots \cap (\upsilon_{il} - \upsilon_{ik} > 0)],\end{aligned} \qquad (1)$$

where $\boldsymbol{\xi}$ contains the parameters of the postulated distribution function for υ_{ij}. Let $\mathbf{C}_i$ be a $(J-1) \times J$ contrast matrix that indicates the sign of the differences among the ranked items for a given ranking $\mathbf{r}_i$ of J items. For example, for $J = 3$, and the ordering vectors $\mathbf{r}_i = (j, l, k)$ and $\mathbf{r}_{i'} = (k, j, l)$, the corresponding contrast matrices take on the form

$$\mathbf{C}_i = \begin{bmatrix} 1 & 0 & -1 \\ 0 & -1 & 1 \end{bmatrix} \text{ and } \mathbf{C}_{i'} = \begin{bmatrix} -1 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix}, \tag{2}$$

where the three columns of the contrast matrices correspond to the items j, k, and l, respectively. (1) can then be written as

$$\Pr(\mathbf{r}_i|\boldsymbol{\xi}) = \Pr(\mathbf{C}_i\boldsymbol{\upsilon}_i > \mathbf{0}), \tag{3}$$

where $\boldsymbol{\upsilon}_i = (\upsilon_{i1}, \ldots, \upsilon_{iJ})$ contains the option utility assessments of person i. If, as proposed by Thurstone [14], the rankers' judgments of the J items are multivariate normal (*see* **Catalogue of Probability Density Functions**) with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ (*see* **Correlation and Covariance Matrices**), the distribution of the pairwise differences of the $\boldsymbol{\upsilon}$'s is also multivariate normal. Consequently, the probability of observing the rank order vector $\mathbf{r}_i$ can be determined by evaluating an $(J-1)$–variate normal distribution,

$$\Pr(\mathbf{r}_i|\boldsymbol{\xi}) = \frac{|\boldsymbol{\Gamma}_i|^{-\frac{1}{2}}}{(2\pi)^{\frac{(J-1)}{2}}} \int_0^\infty \cdots \int_0^\infty \exp\{-\tfrac{1}{2}(\boldsymbol{\delta}_i - \mathbf{x})'\boldsymbol{\Gamma}_i^{-1}(\boldsymbol{\delta}_i - \mathbf{x})\}\,\mathrm{d}\mathbf{x}, \tag{4}$$

where $\boldsymbol{\delta}_i = \mathbf{C}_i\,\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}_i = \mathbf{C}_i\boldsymbol{\Sigma}\mathbf{C}_i'$. Both the mean utilities and their covariances may be related to observed covariates $\mathbf{x}_{ij}$ to identify systematic sources of individual differences in the evaluation of the choice options.

When a decision maker chooses among the options for T choice sets, we obtain a $(J \times T)$ ordering matrix $\mathbf{R}_i = (\mathbf{r}_{i1}, \ldots, \mathbf{r}_{iT})$ containing person's i rankings for each of the choice set. With multiple-choice sets, it becomes possible to distinguish explicitly between within- and between-choice set variability in the evaluation of the choice options. Both sources of variation are confounded when preferences for only a single choice set are elicited. For example, when respondents compare sequentially all possible pairs of choice options, $T = \binom{J}{2}$, the probability of observing the ordering matrix $\mathbf{R}_i$ is obtained by evaluating a $\binom{J}{2}$-dimensional normal distribution function with mean vector $\mathbf{A}_i\boldsymbol{\mu}$ and covariance matrix $\mathbf{A}_i\boldsymbol{\Sigma}\mathbf{A}_i + \boldsymbol{\Psi}$. The rows of $\mathbf{A}_i$ contain the contrast vectors $\mathbf{c}_{ik}$ corresponding to the choice outcome for the k-th choice set and $\boldsymbol{\Psi}$ is a diagonal matrix containing the within-choice-set variances.

For large J and/or T, the evaluation of the normal distribution function by numerical integration is not feasible with current techniques. Fortunately, alternative methods are available based on Monte Carlo Markov chain methods (*see* **Markov Chain Monte Carlo and Bayesian Statistics**) (Yao et al. [16], Tsai et al. [15]) or limited-information approximations (Maydeu-Olivares [10]) that can be used for estimating the mean and covariance structure of the choice options. Especially, limited-information methods are sufficiently convenient from a computationally perspective to facilitate the application of Thurstonian scaling models in routine work.

An Application: Modeling the Similarity Structure of Values

An important issue in value research is to identify the basic value dimensions and their underlying structure. According to a prominent theory by Schwartz [11] the primary content aspect of a value is the type of goal it expresses. Based on extensive cross-cultural research, this author identified 10 distinct values: [1] power, [2] achievement, [3] hedonism, [4] stimulation, [5] self-direction, [6] universalism, [7] benevolence, [8] tradition, [9] conformity, and [10] security. The circular pattern in Figure 1 displays the similarity and polarity relationships among these values. The closer any two values in either direction around the circle, the more similar their underlying motivations. More distant values are more antagonistic in their underlying motivations. As a result of these hypothesized relationships, one may expect that associations among the value items exhibit systematic increases and decreases depending on their closeness and their degree of polarity.

To test this hypothesized value representation, binary paired comparison data were collected from a random sample of 338 students at a North-American university. The students were asked to indicate for

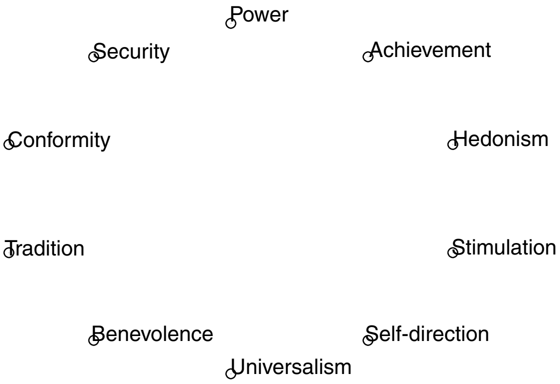


Figure 1 Hypothesized circumplex structure of ten motivationally distinct types of values

each of the 45 pairs formed on the basis of the 10 values, which one was more important as a guiding principle in their life. The respondents were highly consistent in their importance evaluations with less than 5% of the pairwise judgments being intransitive.

Figure 2 displays the first two **principal components** of the estimated covariance matrix $\hat{\mathbf{\Sigma}}$ of the ten values. Because the origin of the value scale cannot be identified on the basis of the pairwise judgments, the estimated coordinates may be rotated or shifted in arbitrary ways, only the distances between the estimated coordinates should be interpreted. Item positions that are closer to each other have a higher covariance than points that are further apart. Consistent with the circular value representation, the first component contrasts self-enhancement values with values describing self-transcendence and the second component contrasts openness-to-change with conservation values. However, the agreement with the circumplex structure is far from perfect. Several values, most notably 'self-direction' and 'security', deviate systematically from their hypothesized positions.

Concluding Remarks

Scaling models are useful in providing parsimonious descriptions of how individuals perceive and evaluate choice options. However, preferences may not always be well-defined and may depend on seemingly irrelevant contextual conditions (Kahneman [3]). Diverse factors such as the framing of the choice task and the set of offered options have been shown to influence strongly choice outcomes. As a result, generalizations of scaling results to different choice situations and

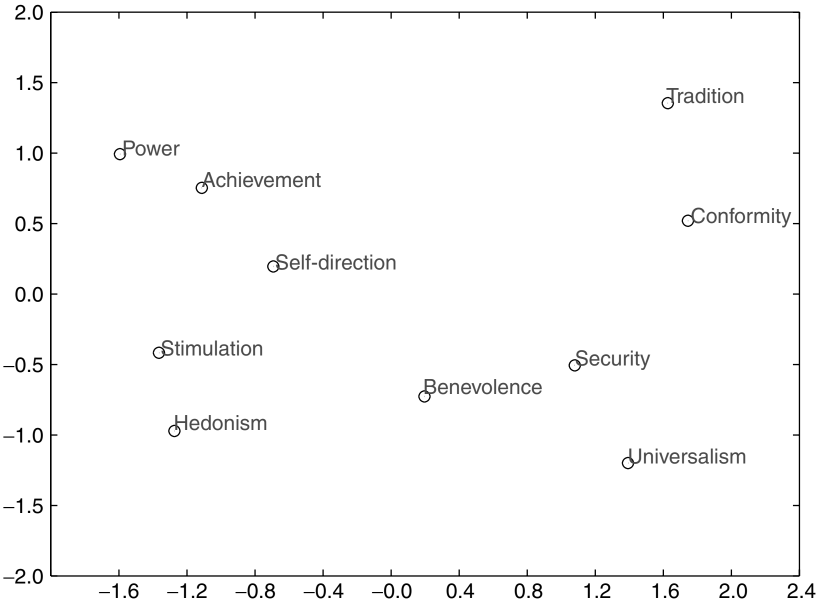


Figure 2 The two major dimensions of the covariance matrix $\hat{\mathbf{\Sigma}}$ estimated from pairwise value judgments

options require much care and frequently need to be based on additional validation studies.

References

[1] Carroll, J.D. & DeSoete, G. (1991). Toward a new paradigm for the study of multiattribute choice behavior, *American Psychologist* **46**, 342–352.

[2] DeLeeuw, J. (2005). Multidimensional unfolding, *Encyclopedia of Behavioral Statistics*, Wiley, New York.

[3] Kahneman, D. (2003). Maps of bounded rationality: psychology for behavioral economics, *American Economic Review* **93**, 1449–1475.

[4] Keane, M.P. (1997). Modeling heterogeneity and state dependence in consumer choice behavior, *Journal of Business and Economic Statistics* **15**, 310–327.

[5] Louviere, J.J., Hensher, D.A. & Swait, J.D. (2000). *Stated Choice Methods*, Cambridge University Press, New York.

[6] Luce, R.D. (1959). *Individual Choice Behavior*, Wiley, New York.

[7] Manski, C. & McFadden, D., eds (1981). *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge.

[8] Marden, J.I. (2005). The Bradley-Terry model, *Encyclopedia of Behavioral Statistics*, Wiley, New York.

[9] Marshall, P. & Bradlow, E.T. (2002). A unified approach to conjoint analysis models, *Journal of the American Statistical Association* **97**, 674–682.

[10] Maydeu-Olivares, A. (2002). Limited information estimation and testing of Thurstonian models for preference data, *Mathematical Social Sciences* **43**, 467–483.

[11] Schwartz, S.H. (1992). Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries, *Advances in Experimental Social Psychology* **25**, 1–65.

[12] Stern, H. (1990). A continuum of paired comparison models, *Biometrika* **77**, 265–273.

[13] Takane, Y. (1987). Analysis of covariance structures and probabilistic binary choice data, *Cognition and Communication* **20**, 45–62.

[14] Thurstone, L.L. (1927). A law of comparative judgment, *Psychological Review* **34**, 273–286.

[15] Tsai, R.-C. & Böckenholt, U. (2002). Two-level linear paired comparison models: estimation and identifiability issues, *Mathematical Social Sciences* **43**, 429–449.

[16] Yao, G. & Böckenholt, U. (1999). Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler, *British Journal of Mathematical and Statistical Psychology* **52**, 79–92.

(*See also* **Attitude Scaling**; **Multidimensional Scaling**; **Multidimensional Unfolding**)

ULF BÖCKENHOLT

Scatterplot Matrices

Daniel B. Wright and Siân E. Williams

Volume 4, pp. 1794–1795

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Scatterplot Matrices

When researchers are interested in the relationships between pairs of several continuous variables, they often produce a series of **scatterplots** for each of the pairs. It can be convenient to view these together on a single screen or page using what is usually called a *scatterplot matrix* (though sometimes referred to as a *draftman's plot*). Many statistical packages have this facility. With k variables, there are $k(k-1)/2$ pairs, and therefore for even small numbers of variables the number of scatterplots can be large. This means each individual scatterplot on the display is small. An example is shown in Figure 1.

Scatterplot matrices are useful for quickly ascertaining all the bivariate relationships, but because of the size of the individual scatterplot it may be difficult to fully understand the relationship. Some of the extra facilities common for two variable scatterplots, such as adding symbols and including confidence limits on regression lines, would create too much clutter in a scatterplot matrix. Here, we have included a line for the linear regression and the univariate **histograms**. Any more information would be difficult to decipher.

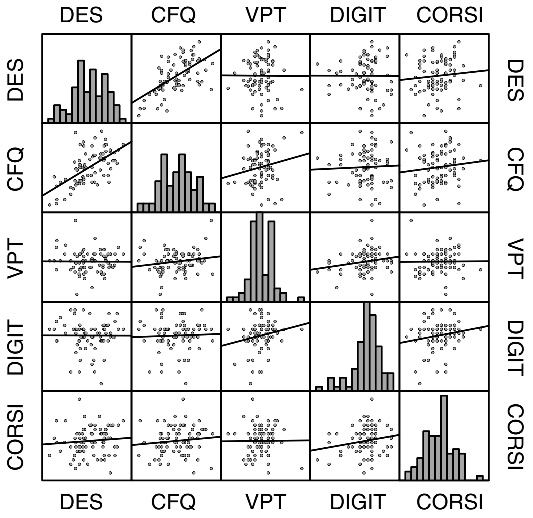


Figure 1 A scatterplot matrix that shows the bivariate relationships between two personality measures (DES – dissociation, CFQ – cognitive failures questionnaire) and impairment from secondary tasks on three working memory tasks (VPT – visual patterns, DIGIT – digit span, CORSI – Corsi block test). Data from [2]

Figure 1 just shows bivariate relationships. Sometimes, it is useful to look at the bivariate relationship between two variables at different values or levels of a third variable. In this case, we produce a **trellis display** or casement display. Consider the following study [3] in which participants heard lists of semantically associated words and were played a piece of music. Later, they were asked to recall the words, and how many times the participant recalled a semantically related word that was not originally presented (a lure) was recorded. Figure 2 shows the relationship between the number of lures recalled and how much the participant liked the music. There were two experimental conditions. In the first, participants were told to recall as many as words as

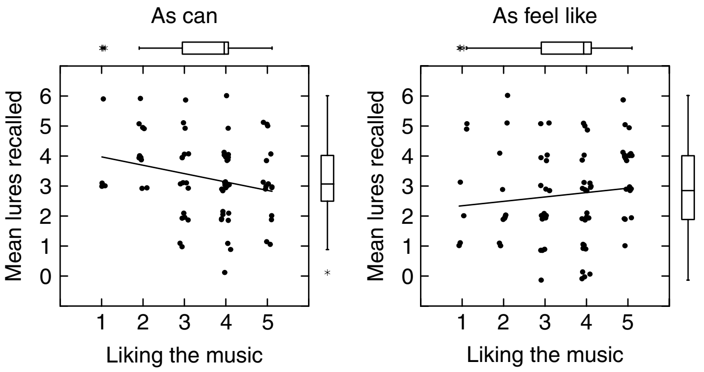


Figure 2 The relationship between liking the music and the number of critical lures recalled depends on the recall task. A random jitter has been added to the points that all are visible. Data from [3]

they could. The more the participant liked the music, the fewer lures were recalled. The argument is that the music put these people in a good mood so they felt satisfied with their recall so did not try as hard. In the second condition, participants were told to recall as many as words as they felt like. Here, the more people liked the music, the more lures they recalled, that is, if they were happy because of the music, they continued to feel like recalling words.

Trellis scatterplots can be used with more than one conditioning variable. However, with more than two conditioning variables, they can be difficult to interpret. If multivariate relationships are of interest, other techniques, such as **three-dimensional scatterplots** and **bubble plots**, are more appropriate.

A useful source for further information on scatterplot matrices is [1].

References

[1] Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. (1983). *Graphical Methods for Data Analysis*, Duxbury, Boston.

[2] Wright, D.B. & Osborne, J. (2005 in press). Dissociation, cognitive failures, and working memory, *American Journal of Psychology*.

[3] Wright, D.B., Startup, H.M. & Mathews, S. (under review). Dissociation, mood, and false memories using the Deese-Roediger-McDermott procedure, *British Journal of Psychology*.

Daniel B. Wright and Siân E. Williams

Scatterplot Smoothers

Brian S. Everitt

Volume 4, pp. 1796–1798

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Scatterplot Smoothers

The **scatterplot** is an excellent first exploratory graph with which to study the dependence of two variables. Often, understanding of the relationship between the two variables is aided by adding the result of a simple linear fit to the plot. Figure 1 shows such a plot for the average oral vocabulary size of children at various ages.

Here, the linear fit does not seem adequate to describe the growth in vocabulary size with increasing age, and some form of polynomial curve might be more appropriate. Since the plotted observations show a tendency to an 'S' shape, a cubic might be a possibility. If such a curve is fitted to the data, it appears to fit the available data well but between the observations, it rises and then drops again. Consequently, as a model of language acquisition, it leads to the absurd implication that newborns have large vocabularies, which they lose by the age on one, then their vocabulary increases until the age of six, when children start to forget words rather rapidly! Not a very sensible model.

An alternative to using parametric curves to fit bivariate data is to use a nonparametric approach in which we allow the data themselves to suggest the appropriate functional form. The simplest of these alternatives is to use a locally weighted regression or loess fit, first suggested by Cleveland [1]. In essence, this approach assumes that the variables x and y are related by the equation

$$y_i = g(x_i) + \in_i, \tag{1}$$

where g is a 'smooth' function and the $\in_i$ are random variables with mean zero and constant scale. Values $\hat{y}_i$ used to 'estimate' the y_i at each x_i are found by fitting polynomials using weighted least squares with large weights for points near to x_i and small weights otherwise. So smoothing takes place essentially by local averaging of the y-values of observations having predictor values close to a target value. Adding such a plot to the data is often a useful alternative to the more familiar parametric curves such as simple linear or polynomial regression fits (*see* **Multiple Linear Regression**; **Polynomial Model**) when the bivariate data plotted is too complex to be described by a simple parametric family. Figure 2 shows the result of fitting a locally weighted regression curve to the vocabulary data. The locally weighted regression fit is able to follow the nonlinearity in the data although the difference in the two curves is not great.

An alternative smoother that can often usefully be applied to bivariate data is some form of spline function. (A spline is a term for a flexible strip of metal or rubber used by a draftsman to draw curves.) Spline functions are polynomials within intervals of the x-variable that are connected across different values of x. Figure 3, for example, shows a linear spline function, that is a piecewise linear function, of the form

$$f(x) = \beta_0 + \beta_1 X + \beta_2 (X - a)_+ \\ + \beta_3 (X - b)_+ + \beta_4 (X - c)_+$$

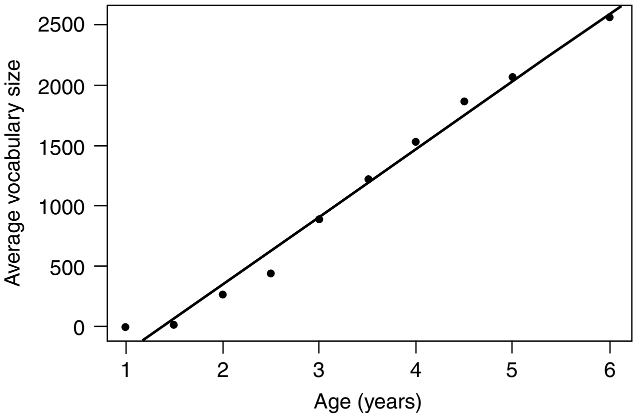


Figure 1 Scatterplot of average vocabulary score against age showing linear regression fit

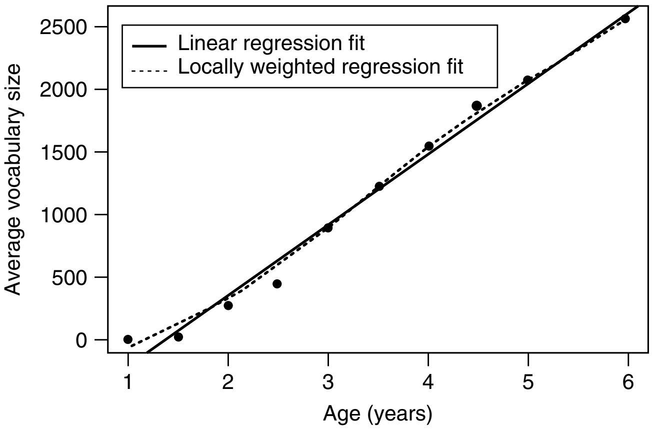


Figure 2 Scatterplot of vocabulary score against age showing linear regression fit and locally weighted regression fit

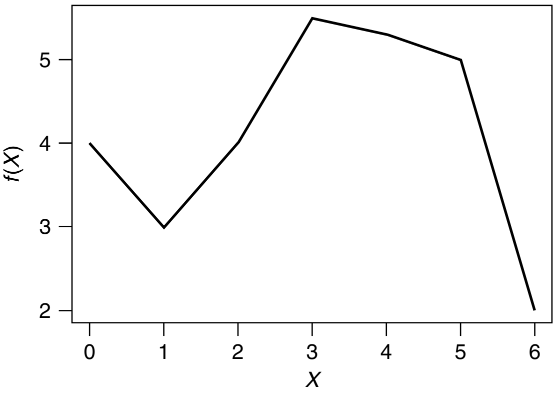


Figure 3 A linear spline function with knots at $a = 1$, $b = 3$, $c = 5$. Taken with permission from [3]

$$\text{where } (u)_+ = u \quad u > 0$$
$$= 0 \quad u \le 0, \tag{2}$$

The interval endpoints, a, b, and c are called *knots*. The number of knots can vary according to the amount of data available for fitting the function.

The linear spline is simple and can approximate some relationships, but it is not smooth and so will not fit highly curved functions well. The problem is overcome by using piecewise polynomials, in particular, cubics, which have been found to have nice properties with good ability to fit a variety of complex relationships. The result is a cubic spline that arises formally by seeking a smooth curve $g(x)$ to summarize the dependence of y on x, which minimizes the rather daunting expression:

$$\sum [y_i - g(x_i)]^2 + \lambda \int g''(x)^2 \mathrm{d}x, \tag{3}$$

where $g''(x)$ represents the second derivative of $g(x)$ with respect to x. Although when written formally this criterion looks a little formidable, it is really nothing more than an effort to govern the trade-off between the goodness-of-fit of the data (as measured by $\sum [y_i - g(x_i)]^2$) and the 'wiggliness' or departure of linearity of g measured by $\int g''(x)^2 \mathrm{d}x$; for a linear function, this part of (3) would be zero. The parameter λ governs the smoothness of g, with larger values resulting in a smoother curve.

The solution to (3) is a cubic spline, that is, a series of cubic polynomials joined at the unique observed values of the explanatory variable, x_i. (For more details, see [2]). Figure 4 shows a further scatterplot of the vocabulary data now containing linear regression, locally weighted regression, and spline smoother fits. When interpolating a number of points, a spline can be a much better solution than a polynomial interpolant, since the polynomial can oscillate wildly to hit all the points; polynomial fit the data globally, while splines fit the data locally.

Locally weighted regressions and spline smoothers are the basis of **generalized additive models**.

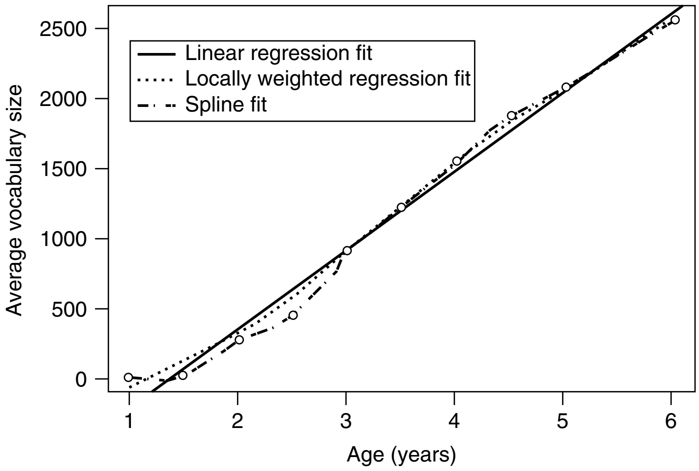


Figure 4 Scatterplot of vocabulary score against age showing linear regression, locally weighted regression, and spline fits

References

[1] Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**, 829–836.

[2] Friedman, J.H. (1991). Multiple adaptive regression splines, *Annals of Statistics* **19**, 1–67.

[3] Harrell, F.E. (2001). *Regression Modelling Strategies with Applications to Linear Models, Logistic Regression and Survival Analysis*, Springer, New York.

BRIAN S. EVERITT

Scatterplots

Siân E. Williams and Daniel B. Wright

Volume 4, pp. 1798–1799

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Scatterplots

Scatterplots are typically used to display the relationship, or association, between two variables. Examples include the relationship between *age* and *salary* and that between inches of *rainfall* in a month and the number of *car accidents*. Both variables need to be measured on some continuum or scale. If there is a natural response variable or a predicted variable, then it should be placed on the y-axis. For example, age would be placed on the x-axis and salary on the y-axis because it is likely that you would hypothesize that salary is, in part, dependant on age rather than the other way round.

Consider the following example. The estimated number of days in which students expect to take to complete an essay is compared with the actual number of days taken to complete the essay. The scatterplot in Figure 1 shows the relationship between the estimated and actual number of days.

Most statistical packages allow various options to increase the amount of information presented. In Figure 1, a diagonal line is drawn, which corresponds to positions where estimated number of days equals actual number of days. Overestimators fall below the diagonal, and underestimators fall above the diagonal. You can see from this scatterplot that most students underestimated the time it took them to complete the essay. **Box plots** are also included here to provide univariate information.

Other possible options include adding different regression lines to the graph, having the size of the points represent their impact on the regression line, using 'sunflowers', and 'jittering'. The use of the sunflowers option and jittering option allow multiple observations falling on the same location of the plot to be counted. Consider Figures 2(a), (b), and (c). Participants were presented with a cue event from their own autobiography and were asked whether that event prompted any other memory [2]. Because participants did this for several events, there were 1865 date estimates in total. If a standard scatterplot is produced comparing the year of the event with the year of the cueing event, the result is Figure 2(a). Because of the large number of events, and the fact that many

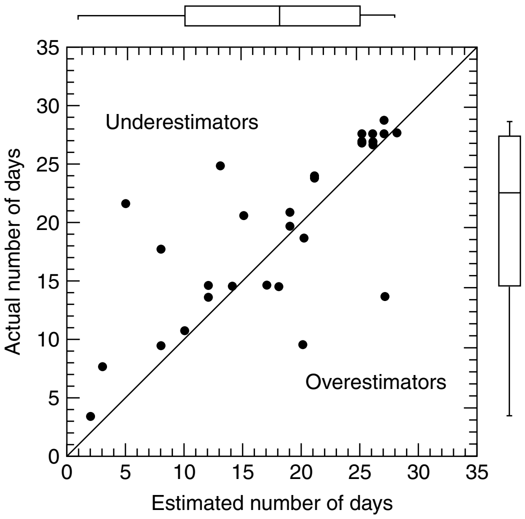


Figure 1 A scatterplot of estimated and actual essay completion times with overlapping case points

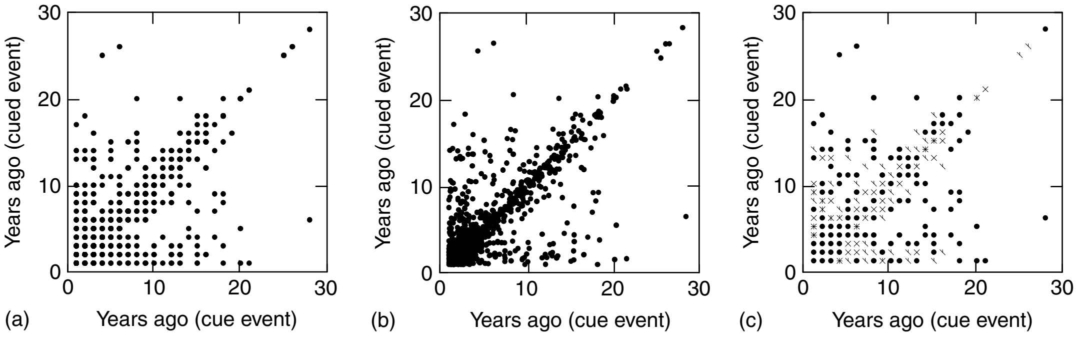


Figure 2 Scatterplots comparing the year of a remembered event with the year of the event that was used to cue the memory. Adapted from Wright, D.B. & Nunn, J.A. (2000). Similarities within event clusters in autobiographical memory, *Applied Cognitive Psychology* **14**, 479–489, with permission from John Wiley & Sons Ltd

overlap, this graph does not allow the reader to determine how many events are represented by each point.

In Figure 2(b), each data point has had a random number (uniformly distributed between -0.45 and +0.45) added to both its horizontal and vertical component. The result is that coordinates with more data points have more dots around them. Jittering is particularly useful with large data sets, like this one. Figure 2(c) shows the sunflower option. Here, individual coordinates are represented with sunflowers. The number of petals represents the number of data points. This option is more useful with smaller data sets. More information on jittering and sunflowers can be found in, for example, [1].

It is important to realize that a scatterplot does not summarize the data; it shows each case in terms of its value on variable x and its value on variable y. Therefore, the choice of regression line and the addition of other summary information can be vital for communicating the main features in your data.

References

[1] Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. (1983). *Graphical Methods for Data Analysis*, Duxbury, Boston.

[2] Wright, D.B. & Nunn, J.A. (2000). Similarities within event clusters in autobiographical memory, *Applied Cognitive Psychology* **14**, 479–489.

Siân E. Williams and Daniel B. Wright

Scheffé, Henry

Brian S. Everitt

Volume 4, pp. 1799–1800

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Scheffé, Henry

Born: April 11, 1907, in New York, USA.
Died: July 5, 1977, in California, USA.

Born in New York City, Scheffé attended elementary school in New York and graduated from high school in Islip, Long Island, in 1924. In 1928, he went to study mathematics at the University of Wisconsin receiving his B.A. in 1931. Four years later, he was awarded a Ph.D. for his thesis entitled 'Asymptotic solutions of certain linear differential equations in which the coefficient of the parameter may have a zero'. Immediately after completing his doctorate, Scheffé began a career as a university teacher in pure mathematics. It was not until 1941 that Scheffé's interests moved to statistics and he joined Samuel Wilks and his statistics team at Princeton. From here he moved to the University of California, and then to Columbia University where he became chair of the statistics department. In 1953, he left Columbia for Berkeley where he remained until his retirement in 1974.

Much of Scheffé's research was concerned with particular aspects of the **linear model** (e.g., [2]), particularly, of course, the **analysis of variance**, resulting in 1959 in the publication of his classic work, *The Analysis of Variance*, described in [1] in the following glowing terms:

> Its careful exposition of the different principal models, their analyses, and the performance of the procedures when the model assumptions do not hold is exemplary, and the book continues to be a standard list and reference.

Scheffé's book has had a major impact on generations of statisticians and, indirectly, on many psychologists. In 1999, the book was reprinted in the Wiley Classics series [3].

During his career, Scheffé became vice president of the American Statistical Association and president of the International Statistical Institute and was elected to many other statistical societies.

References

[1] Lehmann, E.L. (1990). Biography of Scheffé, *Dictionary of Scientific Biography*, Supplement II Vol. 17–18, Scribner, New York.

[2] Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance, *Biometrika* **40**, 87–104.

[3] Scheffé, H. (1999). *The Analysis of Variance*, Wiley, New York.

Brian S. Everitt

Second Order Factor Analysis: Confirmatory

John Tisak and Marie S. Tisak

Volume 4, pp. 1800–1803

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Second Order Factor Analysis: Confirmatory

To motivate the discussion of confirmatory second-order factor analysis, a basic illustration will be provided to highlight the salient features of the topic. For pedagogical reason, the example will be used when all the variables are measured and then it will be repeated when some of the variables are not measured or are latent. This contrast emphasizes the functional similarities between these two approaches and associates the commonly known regression analysis (*see* **Multiple Linear Regression**) with the more complex and less familiar **factor analysis**.

MEASURED VARIABLES. Assume that a random sample of individuals from a specified population has been assessed on Thurstone's Primary Mental Abilities (PMA) scales. Concretely, these abilities are as follows:

Verbal Meaning (VRBM) – a vocabulary recall ability test;
Word Grouping (WORG) – a test of vocabulary recall ability;
Number Facility (NUMF) – a measure of arithmetic reasoning ability;
Letter Series (LETS) – tests reasoning ability by letter series; and
Number Series (NUMS) – a reasoning ability test that uses number series.

Second, assume that the same individuals are measured on Horn and Cattell's Crystallized and Fluid Intelligences. Specifically, these two intelligences are as follows:

Crystallized Intelligence (GC) – measured learned or stored knowledge; and
Fluid Intelligence (GF) – evaluates abstract reasoning capabilities.

Third, consider that once again these individuals are tested on Spearman's General Intelligence (G), which is a global construct of general ability or intelligence. Notice that in moving from the Primary Mental Abilities to Crystallized and Fluid Intelligences to General Intelligence, there is a movement from more specific to more general constructs, which could be considered nested, that is, the more specific variables are a subset of the more general variables. Finally, it is implicit that these variables are infallible or are assessed without measurement error.

First-order regression analysis. If these eight variables were all assessed, then one could evaluate how well the more general Crystallized (GC) and Fluid (GF) Intelligences predict the Primary Mental Abilities using multivariate multiple regression analysis. Precisely, the regression of the PMA scales onto Crystallized and Fluid Intelligences become

$$\begin{aligned} \mathrm{VRBM} &= \beta_0^1 + \beta_1^1\mathrm{GC} + \beta_2^1\mathrm{GF} + \mathrm{e}^1, \\ \mathrm{WORG} &= \beta_0^2 + \beta_1^2\mathrm{GC} + \beta_2^2\mathrm{GF} + \mathrm{e}^2, \\ \mathrm{NUMF} &= \beta_0^3 + \beta_1^3\mathrm{GC} + \beta_2^3\mathrm{GF} + \mathrm{e}^3, \\ \mathrm{LETS} &= \beta_0^4 + \beta_1^4\mathrm{GC} + \beta_2^4\mathrm{GF} + \mathrm{e}^4, \\ \text{and} \qquad \mathrm{NUMS} &= \beta_0^5 + \beta_1^5\mathrm{GC} + \beta_2^5\mathrm{GF} + \mathrm{e}^5, \qquad (1) \end{aligned}$$

where β_0^j are the intercepts for predicting the jth outcome Primary Mental Abilities variable, and where β_1^j and β_2^j are the partial regression coefficients or slopes for predicting the jth outcome variable from Crystallized and Fluid Intelligences, respectively. Lastly, e^1 to e^5 are errors of prediction, that is, what is not explained by the prediction equation for each outcome variable.

Given knowledge of these variables, one could speculate that Crystallized Intelligence would be related to the Verbal Meaning and Word Grouping abilities, whereas Fluid Intelligence would predict the Number Facility, Letter Series, and Word Grouping abilities. Hence, the regression coefficients, $\beta_1^1, \beta_1^2, \beta_2^3, \beta_2^4$, and β_2^5, would be substantial, while, $\beta_2^1, \beta_2^2, \beta_1^3, \beta_1^4$, and β_1^5, would be relatively much smaller.

Second-order regression analysis. Along this line of development, Crystallized (GC) and Fluid (GF) Intelligences could be predicted by General Intelligence (G) by multivariate simple regression analysis. Concretely, the regression equations are

$$\begin{aligned} \mathrm{GC} &= \beta_0^6 + \beta_1^6\mathrm{G} + \mathrm{e}^6 \\ \text{and} \qquad \mathrm{GF} &= \beta_0^7 + \beta_1^7\mathrm{G} + \mathrm{e}^7, \qquad (2) \end{aligned}$$

where β_0^6 and β_0^7 are the intercepts for predicting each of the Crystallized and Fluid Intelligence outcome variables, and where β_1^6 and β_1^7 are the partial regression coefficients or slopes for predicting the two outcome variables from General Intelligence respectively. Additionally, e^6 and e^7 again are errors of prediction. Substantively, one might conjecture that General Intelligence predicts Fluid Intelligence more than it does Crystallized Intelligence, hence β_1^6 would be less than β_1^7.

LATENT VARIABLES. Suppose that Fluid (gf) and Crystallized (gc) Intelligences and General Intelligence (g) are not observed or measured, but instead they are **latent variables**. (Note that the labels on these three variables have been changed from upper to lower case to indicate that they are unobserved.) Thus, the impact of these variables must be determined by latent variable or factor analysis. As before, this analysis might be viewed in a two-step process: A first-order and a second-order factor analysis.

First-order factor analysis. If the five PMA variables were assessed, then one could evaluate how well the latent Crystallized (gc) and Fluid (gf) Intelligences predict the Primary Mental Abilities using a first-order factor analysis. (For details *see* **Factor Analysis: Confirmatory**) Precisely, the regression equations are

$$\begin{aligned}
\text{VRBM} &= \tau_0^1 + \lambda_1^1\text{gc} + \lambda_2^1\text{gf} + \varepsilon^1,\\
\text{WORG} &= \tau_0^2 + \lambda_1^2\text{gc} + \lambda_2^2\text{gf} + \varepsilon^2,\\
\text{NUMF} &= \tau_0^3 + \lambda_1^3\text{gc} + \lambda_2^3\text{gf} + \varepsilon^3,\\
\text{LETS} &= \tau_0^4 + \lambda_1^4\text{gc} + \lambda_2^4\text{gf} + \varepsilon^4,\\
\text{and}\qquad \text{NUMS} &= \tau_0^5 + \lambda_1^5\text{gc} + \lambda_2^5\text{gf} + \varepsilon^5, \qquad (3)
\end{aligned}$$

where τ_0^j are the intercepts for predicting the jth observed outcome Primary Mental Abilities variable, and where λ_1^j and λ_2^j are the partial regression coefficients or slopes for predicting the jth outcome variable from unobserved Crystallized and Fluid Intelligences, respectively. Unlike multiple regression analysis, in factor analysis, the errors, ε^1 to ε^5, are now called *unique factors* and each contain two entities, a specific factor and a measurement error. For example, the unique factor, ε^1, consists of a specific factor that contains what is not predicted in Verbal Meaning by Crystallized and Fluid Intelligences and a measurement error induced by imprecision in the assessment of Verbal Meaning. Notice that these regression coefficients and errors have been relabeled to emphasize that the predictors are now latent variables and to be consistent with the nomenclature used in LISREL, a commonly used computer package for these analyses (*see* **Structural Equation Modeling: Software**), but that their interpretation is analogous to those in the measured variable section.

As before, one could speculate that Crystallized Intelligence would be related to the Verbal Meaning and Word Grouping abilities, whereas Fluid Intelligence would predict the Number Facility, Letter Series, and Word Grouping abilities. Unfortunately, unlike before, the (latent) variables or factors, gc and gf, are unknown, which creates an indeterminacy in the previous equations, that is, the regression coefficients cannot be uniquely determined. A standard solution to this problem is the use of marker or reference variables. Specifically for each factor, an observed variable is selected that embodies the factor. For example, since Crystallized Intelligence could be considered learned knowledge, one might select Verbal Meaning, accumulated knowledge, to represent it. For this case, the intercept is equated to zero ($\tau_0^1 \equiv 0$); the slope associated with gc equated to one ($\lambda_1^1 \equiv 1$) and the slope associated with gf equated to zero ($\lambda_2^1 \equiv 0$). Hence for Verbal Meaning, the regression equation becomes VRBM = gc +ε^1. By similar reasoning, Number Facility could serve as a reference variable for Fluid Intelligence, because they are linked by reasoning ability. Hence, for Number Facility, the regression equation is NUMF = gf +ε^3. Implicit is that the associated intercept and slope for gc are zero ($\tau_0^3 \equiv 0$ and $\lambda_1^3 \equiv 0$) and that the slope for gf is one ($\lambda_2^3 \equiv 1$).

Again, in keeping with our theoretical speculation (i.e., that VRBM and WORG are highly predicted from gc, but not gf, and vice versa for NUMF, LETS, and NUMS), the regression coefficients or loadings, λ_1^2, λ_2^4, and λ_2^5, would be substantial, while λ_2^2, λ_1^4, and λ_1^5, would be relatively much smaller.

Second-order factor analysis. Similar to second-order regression analysis, the latent variables, Crystallized (gc) and Fluid (gf) Intelligences, could be predicted by General Intelligence (g) by a second-order factor analysis. Concretely,

$$\begin{aligned}
\text{gc} &= \alpha_0^1 + \gamma_1^1\text{g} + \zeta^1\\
\text{and}\qquad \text{gf} &= \alpha_0^2 + \gamma_1^2\text{g} + \zeta^2, \qquad (4)
\end{aligned}$$

where α_0^1 and α_0^2 are the intercepts for predicting each of the Crystallized and Fluid Intelligence outcome variables, and where γ_1^1 and γ_1^2 are the regression coefficients or slopes for predicting the two outcome variables from General Intelligence, respectively. Further, ζ^1 and ζ^2 are second-order specific factors or what has not been predicted by General Intelligence in Crystallized and Fluid Intelligences, respectively. Lastly, recognize that measurement error is not present in these second-order specific factors, because it was removed in the first order equations.

As before, one might conjecture that General Intelligence relates more to Fluid Intelligence than it predicts Crystallized Intelligence. Furthermore, as with the first-order factor analysis, there is indeterminacy in that the regression coefficients are not unique. Hence, again a reference variable is required for each latent variable or second-order factor. For example, Fluid Intelligence could be selected as a reference variable. Thus, the intercept and slope for it will be set to zero and one respectively ($\alpha_0^2 \equiv 0$ and $\gamma_1^2 \equiv 1$). So, the regression equation for Fluid Intelligence becomes $\mathrm{gf} = \mathrm{g} + \zeta^2$.

TECHNICAL DETAILS. In the foregoing discussion, similarities between multiple regression and factor analysis were developed by noting the linearity of the function form or prediction equation. Additionally, first- and second-order models were developed separately. Parenthetically, when originally developed, a first-order factor analysis would be performed initially and then a second-order factor would be undertaken using these initial results (or, more accurately, the variances and covariances between the first-order factors). It is the current practice to perform both the first- and second-order factor analysis in the same model or at the same time. Moreover, remember that for the confirmatory approach, in either the first- or second-order factor analytic model, *a priori* specifications are required to ensure the establishment of reference variables.

In the previous developments, the regression coefficients or parameters were emphasized because they denote the linear relationships among the variables. It is important to note that there are also means, variances, and covariances associated with the second-order factor analytic model. Specifically, the second-order factor – General Intelligence in the example – has a mean and a variance parameter. Note that if there was more than one second-order factor, there would be additional mean, variance, and covariance parameters associated with it. Also, the second-order specific factors typically have variance and covariance parameters (the mean of these specific factors are assumed to be zero). Finally, each of the unique factors from the first-order model has a variance parameter (by assumption, their means and covariances are zero).

If normality of the observed variables is tenable, then all of these parameters may be determined by a statistical technique called **maximum likelihood estimation**. Further, a variety of theoretical hypotheses may be evaluated or tested by chi-square statistics.

For example, one could test if Verbal Meaning and Word Grouping are predicted only by Crystallized Intelligence by hypothesizing that the regression coefficients for Fluid Intelligence are zero, that is, $\lambda_2^1 = 0$ and $\lambda_2^2 = 0$.

John Tisak and Marie S. Tisak

Selection Study (Mouse Genetics)

Norman Henderson

Volume 4, pp. 1803–1804

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Selection Study (Mouse Genetics)

Some of the earliest systematic studies on the inheritance of behavioral traits in animals involved artificial selection. These include Tolman's selection for 'bright' and 'dull' maze learners in 1924, Rundquist's selection for 'active' and 'inactive' rats in a running wheel in 1933, and Hall's 1938 selection for 'emotional' and 'nonemotional' rats in an open field. These pioneering studies triggered an interest in selective breeding for a large variety of behavioral and neurobiological traits that has persisted into the twenty-first century. For a description of early studies and some of the subsequent large-scale behavioral selection studies that followed, see [2] and [5].

Selection experiments appear simple to carry out and most are successful in altering the levels of expression of the selected behavior in only a few generations. Indeed, for centuries preceding the first genetic experiments of Mendel, animal breeders successfully bred for a variety of behavioral and related characters in many species. The simplicity of artificial selection experiments is deceptive, however, and the consummate study requires a considerable effort to avoid problems that can undermine the reliability of the genetic information sought. Frequently, this genetic information includes the realized narrow **heritability** (h^2), the proportion of the observed phenotypic variance that is explained by additive genetic variance, but more often the genotypic correlations (*see* **Gene-Environment Correlation**) between the selected trait and other biobehavioral measures. A lucid account of the quantitative genetic theory related to selection and related issues involving small populations can be found in [1]. Some key issues are summarized here.

Realized heritability of the selected trait is estimated from the ratio of the response to selection to the selection differential:

$$\begin{aligned} h^2 &= \frac{\text{Response}}{\text{Selection differential}} \\ &= \frac{\text{Offspring mean} - \text{Base population mean}}{\text{Selected parent mean} - \text{Base population mean}} \end{aligned} \quad (1)$$

Normally, selection is bidirectional, with extreme scoring animals chosen to be parents for high and low lines, and the heritability estimates averaged. In the rare case where the number of animals tested in the base population is so large that the effective N of each of the selected parent groups exceeds 100, (1) will work well, even in a single generation, although h^2 is usually estimated from several generations by regressing the generation means on the cumulative selection differential. Also, when parent Ns are very large and the offspring of high and low lines differ significantly on any other nonselected trait, one can conclude that the new trait is both heritable and genetically correlated with the selected trait, although the correlation cannot be estimated unless h^2 of the new trait is known [3].

Since few studies involve such large parent groups in each breeding generation, most selection experiments are complicated by the related effects of inbreeding, random genetic drift, and genetic differentiation among subpopulations at all genetic loci. Over succeeding generations, high-trait and low-trait lines begin to differ on many genetically influenced characters that are unrelated to the trait being selected for. The magnitude of these effects is a function of the effective breeding size (N_e), which influences the increment in the coefficient of inbreeding (ΔF) that occurs from one generation to the next. When selected parents are randomly mated, $\Delta F = 1/(2N_e)$. When sib matings are excluded, $\Delta F = 1/(2N_e + 4)$. N_e is a function of the number of male and female parents in a selected line that successfully breed in a generation:

$$N_e = \frac{4N_m N_f}{N_m + N_f} \quad \text{(approx.)} \quad (2)$$

N_e is maximized when the number of male and female parents is equal. Thus, when the parents of a selected line consist of 10 males and 10 females, $N_e = 20$ for that generation, whereas $N_e = 15$ when the breeding parents consist of 5 males and 15 females.

When selection is carried out over many generations, alleles that increase and decrease expression of the selected trait continue to segregate into high- and low-scoring lines, but inbreeding and consequently random drift are also progressing at all loci. If we define the base population as having an inbreeding coefficient of zero, then the inbreeding coefficient in any subsequent selected generation, t, is approximately:

$$F_t = 1 - [(1 - \Delta F_1) \times (1 - \Delta F_2) \times (1 - \Delta F_3) \times \cdots \times (1 - \Delta F_t)] \quad (3)$$

If N_e, and thus ΔF, are constant over generations, (3) simplifies to $F_t = 1 - (1 - \Delta F)^t$. Typically, however, over many generations of selection, N_e fluctuates and may even include some 'genetic bottleneck' generations, where N_e is quite small and ΔF large. It can be seen from (3) that bottlenecks can substantially increase cumulative inbreeding. For example, maintaining 10 male and 10 female parents in each generation of a line for 12 generations will result in an F_t of approximately $1 - (1 - 0.025)^{12} = 0.26$, but, if in just one of those 12 generations only a single male successfully breeds with the 10 selected females, N_e drops from 20 to 3.6 for that generation, causing F_t to increase from 0.26 to 0.35.

As inbreeding continues over successive generations, within-line genetic variance decreases by $1 - F$ and between-line genetic variance increases by $2F$ at all loci. Selected lines continue to diverge on many genetically influenced traits due to random drift, which is unrelated to the trait being selected for. Genetic variance contributed by drift can also exaggerate or suppress the response to selection and is partly responsible for the variability in generation means as selection progresses. Without genotyping subjects, the effects of random drift can only be assessed by having replicated selected lines. One cannot obtain empirical estimates of sampling variation of realized heritabilities in experiments that do not involve replicates. Since the lines can diverge on unrelated traits by drift alone, the lack of replicate lines also poses problems for the common practice of comparing high and low lines on new traits thought to be genetically related to the selected trait. Unless inbreeding is extreme or heritability low, the size of a high-line versus low-line mean difference in phenotypic SD units can help determine if the difference is too large to be reasonably due to genetic drift [4].

References

[1] Falconer, D.S. & Mackay, T.F.C. (1996). *Introduction to Quantitative Genetics*, 4th Edition, Longman, New York.

[2] Fuller, J.L. & Thompson, W.R. (1978). *Foundations of Behavior Genetics*, Mosby, New York.

[3] Henderson, N.D. (1989). Interpreting studies that compare high- and low-selected lines on new characters, *Behavior Genetics* **19**, 473–502.

[4] Henderson, N.D. (1997). Spurious associations in unreplicated selected lines, *Behavior Genetics* **27**, 145–154.

[5] Plomin, R., DeFries, J.C., McClearn, G.E. & McGuffin, P. (2000). *Behavioral Genetics: A Primer*, 4th Edition, W. H. Freeman, New York.

(*See also* **Inbred Strain Study**)

NORMAN HENDERSON

Sensitivity Analysis

Jon Wakefield

Volume 4, pp. 1805–1809

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Sensitivity Analysis

Introduction

Consider an experiment in which varying dosage levels of a drug are randomly assigned to groups of individuals. If the **randomization** is successfully implemented, the groups of subjects will be balanced with respect to all variables other than the dosage levels of interest, at least on average (and if the groups are sufficiently large, effectively in practice also). The beauty of randomization is that the groups are balanced not only with respect to measured variables, but also with respect to unmeasured variables. In a nonrandomized situation, although one may control for *observed* explanatory variables, one can never guarantee that observed associations are not due to unmeasured variables. Selection bias, in which the chances of observing a particular individual depend on the values of their responses and explanatory variables, is another potential source of bias. A further source of bias is due to measurement error in the explanatory variable(s) (in the randomized example this could correspond to inaccurate measurement of the dosage received, though it could also correspond to other explanatory variable). This problem is sometimes referred to as *errors-in-variables* and is discussed in detail in [1]. Many other types of sensitivity analysis are possible (for example, with respect to prior distributions in a Bayesian analysis) (*see* **Bayesian Statistics**), but we consider confounding, measurement error, and selection bias only. For more discussion of these topics in an epidemiological context, see [4, Chapter 19].

A general approach to sensitivity analyses is to first write down a plausible model for the response in terms of accurately measured explanatory variables (some of which may be unobserved), and with respect to a particular selection model. One may then derive the induced form of the model, in terms of observed variables and the selection mechanism assumed in the analysis. The parameters of the derived model can then be compared with the parameters of interest in the 'true' model, to reveal the extent of bias. We follow this approach, but note that it should be pursued only when the sample size in the original study is large, so that sampling variability is negligible; references in the discussion consider more general situations.

In the following, we assume that data are not available to control for bias. So in the next section, we consider the potential effects of *unmeasured* confounding. In the errors-in-variables context, we assume that we do observe 'gold standard' data in which a subset of individuals provides an accurate measure of the explanatory variable, along with the inaccurate measure. Similarly, with respect to selection bias, we assume that the sampling probabilities for study individuals are unknown and cannot be controlled for (as can be done in matched case-control studies, see [4, Chapter 16] for example), or that supplementary data on the selection probabilities of individuals are not available, as in two-phase methods (e.g., [6]); in both of these examples, the selection mechanism is known from the design (and would lead to bias if ignored, since the analysis must respect the sampling scheme).

Sensitivity to Unmeasured Confounding

Let Y denote a univariate response and X a univariate explanatory variable, and suppose that we are interested in the association between Y and X, but Y also potentially depends on U, an unmeasured variable. The discussion in [2] provided an early and clear account of the sensitivity of an observed association to unmeasured confounding, in the context of lung cancer and smoking. For simplicity, we assume that the 'true' model is linear and given by

$$E[Y|X, U] = \alpha^* + X\beta^* + U\gamma^*. \qquad (1)$$

Further assume that the linear association between U and X is $E[X|U] = a + bX$. Roughly speaking, a variable U is a *confounder* if it is associated with both the response, Y, and the explanatory variable, X, but is not be caused by Y or on the causal pathway between X and Y. For a more precise definition of confounding, and an extended discussion, see [4, Chapter 8]. We wish to derive the implied linear association between Y and X, since these are the variables that are observed. We use iterated expectation to average over the unmeasured variable, given X:

$$\begin{aligned} E[Y|X] &= E_{U|X}\{E[Y|X, U]\} \\ &= E_{U|X}\left\{\alpha^* + X\beta^* + U\gamma^*\right\} \end{aligned}$$

$$= \alpha^* + X\beta^* + E[U|X]\gamma^*$$
$$= \alpha^* + X\beta^* + (a + bX)\gamma^* = \alpha + X\beta,$$

where $\alpha = \alpha^* + \gamma^* a$ and, of more interest,

$$\beta = \beta^* + \gamma^* b. \quad (2)$$

Here the 'true' association parameter, β^*, that we would like to estimate, is represented with a $*$ superscript, while the association parameter that we can estimate, β, does not have a superscript. Equation (2) shows that the bias $\beta - \beta^*$ is a function of the level of association between X and U (via the parameter b), and the association between Y and U (via γ^*). Equation (2) can be used to assess the effects of an unmeasured confounding using plausible values of b and γ^*, as we now demonstrate though a simple example.

Example Consider a study in which we wish to estimate the association between the rate of oral cancer and alcohol intake in men over 60 years of age. Let Y represent the natural logarithm of the rate of oral cancer, and let us suppose that we have a two-level alcohol variable X with $X = 0$ corresponding to zero intake and $X = 1$ to nonzero intake. A regression of Y on X gives an estimate $\hat{\beta} = 1.20$, so that the rate of oral cancer is $e^{1.20} = 3.32$ higher in the $X = 1$ population when compared to the $X = 0$ population.

The rate of oral cancer also increases with tobacco consumption (which we suppose is unmeasured in our study), however, and the latter is also positively associated with alcohol intake. We let $U = 0/1$ represent no tobacco/tobacco consumption. Since U is a binary variable, $E[U|X] = P(U = 1|X)$. Suppose that the probability of tobacco consumption is 0.05 and 0.45 in those with zero and nonzero alcohol consumption respectively; that is $P(U = 1|X = 1) = 0.05$ and $P(U = 1|X = 0) = 0.45$, so that $a = 0.05$ and $a + b = 0.45$, to give $b = 0.40$. Suppose further, that the log rate of oral cancer increases by $\gamma^* = \log 2.0 = 0.693$ for those who use tobacco (in both alcohol groups). Under these circumstances, from (2), the true association is

$$\hat{\beta}^* = \hat{\beta} - \gamma^* b = 1.20 - 0.693 \times 0.40 = 0.92, \quad (3)$$

so that the increase in the rate associated with alcohol intake is reduced from 3.32 to $\exp(0.92) = 2.51$.

In a real application, the sensitivity of the association would be explored with respect to a range of values of b and γ^*.

Sensitivity to Measurement Errors

In a similar way, we may examine the potential effects of measurement errors in the regressor X. As an example, consider a simple linear regression and suppose the true model is

$$Y = E[Y|X] + \epsilon^* = \alpha^* + \beta^* X + \epsilon^* \quad (4)$$

where $E[\epsilon^*] = 0$, $\text{var}(\epsilon^*) = \sigma_\epsilon^{*2}$. Rather than measure X, we measure a surrogate W where

$$W = X + \delta \quad (5)$$

with $E[\delta] = 0$, $\text{var}(\delta) = \sigma_\delta^2$, and $\text{cov}(\delta, \epsilon^*) = 0$. The least squares estimator of β^* in model (4), from a sample (X_i, Y_i), $i = 1, \ldots, n$, has the form

$$\hat{\beta}^* = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\text{cov}(X, Y)}{\text{var}(X)}. \quad (6)$$

In the measurement error situation we fit the model

$$Y = E[Y|W] + \epsilon = \alpha + \beta W + \epsilon \quad (7)$$

where $E[\epsilon] = 0$, $\text{var}(\epsilon) = \sigma_\epsilon^2$. The least squares estimator of β in model (7), from a sample (W_i, Y_i), $i = 1, \ldots, n$, has the form

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n}(W_i - \bar{W})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(W_i - \bar{W})^2} = \frac{\text{cov}(W, Y)}{\text{var}(W)}, \quad (8)$$

and to assess the extent of bias, we need to compare $E[\hat{\beta}]$ with β^*. From (8) we have

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i + \delta_i - \bar{X} - \bar{\delta})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i + \delta_i - \bar{X} - \bar{\delta})^2}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) + \frac{1}{n}\sum_{i=1}^{n}(\delta_i - \bar{\delta})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 + \frac{2}{n}\sum_{i=1}^{n}(X_i - \bar{X})(\delta_i - \bar{\delta}) + \frac{1}{n}\sum_{i=1}^{n}(\delta_i - \bar{\delta})^2}$$

$$= \frac{\text{cov}(X, Y) + \text{cov}(\delta, Y)}{\text{var}(X) + 2\text{cov}(X, \delta) + \text{var}(\delta)}. \tag{9}$$

Under the assumptions of our model, $\text{cov}(\delta, Y) = 0$ and $\text{cov}(X, \delta) = 0$ and so, from (6)

$$E[\hat{\beta}] \approx \frac{\text{cov}(X, Y)}{\text{var}(X) + \text{var}(\delta)} = \frac{\text{cov}(X, Y)/\text{var}(X)}{1 + \text{var}(\delta)/\text{var}(X)} = \beta^* r$$

where the *attenuation factor*

$$r = \frac{\text{var}(X)}{\text{var}(X) + \sigma_\delta^2} \tag{10}$$

describes the amount of bias by which the estimate is attenuated toward zero. Note that with no measurement error ($\sigma_\delta^2 = 0$), $r = 1$, and no bias results, and also that the attenuation will be smaller in a well-designed study in which a large range of X is available. Hence, to carry out a sensitivity analysis, we can examine, for an observed estimate $\hat{\beta}$, the increase in the true coefficient for different values of σ_δ^2 via

$$\hat{\beta}^* = \frac{\hat{\beta}}{r}. \tag{11}$$

It is important to emphasize that the above derivation was based on a number of strong assumptions such as independence between errors in Y and in W, and constant variance for errors in both Y and W. Care is required in more complex situations, including those in which we have more than one explanatory variable. For example, if we regress Y on both X (which is measured without error), and a second explanatory variable is measured with error, then we will see bias in our estimator of the coefficient associated with X, if there is a nonzero correlation between X and the second variable (see [1] for more details).

Example Let Y represent systolic blood pressure (in mmHg) and X sodium intake (in mmol/day), and suppose that a linear regression of Y on W produces an estimate of $\hat{\beta} = 0.1$ mmHg, so that an increase in daily sodium of 100 mmol/day is associated with an increase in blood pressure of 10 mmHg. Suppose also that $\text{var}(X) = 4$ mmHg. Table 1 shows the sensitivity of the coefficient associated with X, β^*, to different levels of measurement error; as expected, the estimate increases with increasing measurement error.

Sensitivity to Selection Bias

This section concerns the assessment of the bias that is induced when the probability of observing the data of a particular individual depends on the data of that individual. We consider a slightly different scenario to those considered in the last two sections, and assume we have a binary outcome variable, Y, and a

Table 1 The effect of measurement error when $\text{var}(X) = 4$

Measurement error σ_δ^2	Attenuation factor r	True estimate $\hat{\beta}^*$
0	0	0.1
1	0.8	0.125
2	0.67	0.15
4	0.5	0.2

binary exposure, X, and let $p_x^* = P(Y = 1|X = x)$, $x = 0, 1$, be the 'true' probability of a $Y = 1$ outcome given exposure x, $x = 0, 1$. We take as parameter of interest the **odds ratio**:

$$\begin{aligned} \text{OR}^* &= \frac{P(Y = 1|X = 1)/P(Y = 0|X = 1)}{P(Y = 1|X = 0)/P(Y = 0|X = 0)} \\ &= \frac{p_1^*/(1 - p_1^*)}{p_0^*/(1 - p_0^*)}, \end{aligned} \tag{12}$$

which is the ratio of the odds of a $Y = 1$ outcome given exposed ($X = 1$), to the odds of such an outcome given unexposed ($X = 0$).

We now consider the situation in which we do not have constant probabilities of responding (being observed) across the population of individuals under study, and let $R = 0/1$ correspond to the event nonresponse/response, with response probabilities:

$$P(R = 1|X = x,\ Y = y) = q_{xy}, \tag{13}$$

for $x = 0, 1$, $y = 0, 1$; and we assume that we do not know these response rates. We do observe estimates of $p_x = P(Y = 1|X = x, R = 1)$, the probability of a $Y = 1$ outcome given both values of x and *given* response. The estimate of the odds ratio for the *observed* responders is then given by:

$$\text{OR} = \frac{P(Y = 1|X = 1, R = 1)/P(Y = 0|X = 1, R = 1)}{P(Y = 1|X = 0, R = 1)/P(Y = 0|X = 0, R = 1)} = \frac{p_1/(1 - p_1)}{p_0/(1 - p_0)}. \tag{14}$$

To link the two odds ratios we use Bayes theorem on each of the terms in (14) to give:

$$\begin{aligned} \text{OR} &= \frac{\dfrac{P(R = 1|X = 1, Y = 1)P(Y = 1|X = 1)}{P(R = 1|X = 1)} \bigg/ \dfrac{P(R = 1|X = 1, Y = 0)P(Y = 0|X = 1)}{P(R = 1|X = 1)}}{\dfrac{P(R = 1|X = 0, Y = 1)P(Y = 1|X = 0)}{P(R = 1|X = 0)} \bigg/ \dfrac{P(R = 1|X = 0, Y = 0)P(Y = 0|X = 0)}{P(R = 1|X = 0)}} \\ &= \frac{p_1^*/(1 - p_1^*)}{p_0^*/(1 - p_0^*)} \times \frac{q_{11}q_{00}}{q_{10}q_{01}} = \text{OR}^* \times s, \end{aligned} \tag{15}$$

where the selection factor s is determined by the probabilities of response in each of the exposure-outcome groups. It is of interest to examine situations in which $s = 1$ and there is no bias. One such situation is when $q_{xy} = u_x \times v_y$, $x = 0, 1; y = 0, 1$, so that there is 'no multiplicative interaction' between exposure and outcome in the response model. Note that u_x and v_y are *not* the marginal response probabilities for, respectively, exposure, and outcome.

Example Consider a study carried out to examine the association between childhood asthma and maternal smoking. Let $Y = 0/1$ represent absence/presence of asthma in a child and $X = 0/1$ represent nonexposure/exposure to maternal smoking. Suppose a questionnaire is sent to parents to determine whether their child has asthma and whether the mother smokes. An odds ratio of $\widehat{\text{OR}} = 2$ is observed from the data of the responders, indicating that the odds of asthma is doubled if the mother smokes.

To carry out a sensitivity analysis, there are a number of ways to proceed. We write

$$s = \frac{P(R = 1|X = 1, Y = 1)/P(R = 1|X = 0, Y = 1)}{P(R = 1|X = 1, Y = 0)/P(R = 1|X = 0, Y = 0)} = \frac{q_{11}/q_{01}}{q_{10}/q_{00}}. \tag{16}$$

Suppose that amongst noncases, the response rate in the exposed group is q times that in the unexposed group (that is $q_{10}/q_{00} = q$), while amongst the cases, the response rate in the exposed group is $0.8q$ times that in the unexposed group (i.e., $q_{11}/q_{01} = 0.8q$). In this scenario, $s = 0.8$ and

$$\widehat{\text{OR}}^{*} = \frac{\widehat{\text{OR}}}{0.8} = \frac{2}{0.8} = 2.5, \qquad (17)$$

and we have underestimation because exposed cases were underrepresented in the original sample.

Discussion

In this article we have considered sensitivity analyses in a number of very simple scenarios. An extension would be to simultaneously consider the combined sensitivity to multiple sources of bias. We have also considered the sensitivity of point estimates only, and have not considered hypothesis testing or interval estimation. A comprehensive treatment of observational studies and, in particular, the sensitivity to various forms of bias may be found in [3]. The above derivations can be extended to various different modeling scenarios, for example, [5] examines sensitivity to unmeasured confounding in the context of Poisson regression in spatial epidemiology.

References

[1] Carroll, R.J., Ruppert, D. & Stefanski, L.A. (1995). *Measurement in Nonlinear Models*, Chapman & Hall/CRC Press, London.

[2] Cornfield, J., Haenszel, W., Hammond, E.C., Lillienfield, A.M., Shimkin, M.B. & Wynder, E.L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions, *Journal of the National Cancer Institute* **22**, 173–203.

[3] Rosenbaum, P.R. (2002). *Observational Studies, Second Studies*, Springer-Verlag, New York.

[4] Rothman, K.J. & Greenland, S. (1998). *Modern Epidemiology*, 2nd Edition, Lippincott-Raven, Philadelphia.

[5] Wakefield, J.C. (2003). Sensitivity analyses for ecological regression, *Biometrics* **59**, 9–17.

[6] White, J.E. (1982). A two-stage design for the study of the relationship between a rare exposure and a rare disease, *American Journal of Epidemiology* **115**, 119–128.

(*See also* **Clinical Trials and Intervention Studies**)

JON WAKEFIELD

Sensitivity Analysis in Observational Studies

PAUL R. ROSENBAUM

Volume 4, pp. 1809–1814

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Sensitivity Analysis in Observational Studies

Randomization Inference and Sensitivity Analysis

Randomized Experiments and Observational Studies

In a randomized experiment (*see* **Randomization**), subjects are assigned to treatment or control groups at random, perhaps by the flip of a coin or perhaps using random numbers generated by a computer [7]. Random assignment is the norm in **clinical trials** of treatments intended to benefit human subjects [21, 22]. Intuitively, randomization is an equitable way to construct treated and control groups, conferring no advantage to either group. At baseline before treatment in a randomized experiment, the groups differ only by chance, by the flip of the coin that assigned one subject to treatment, another to control. Therefore, comparing treated and control groups after treatment in a randomized experiment, if a common statistical test rejects the hypothesis that the difference in outcomes is due to chance, then a treatment effect is demonstrated. In contrast, in an observational study, subjects are not randomly assigned to groups, and outcomes may differ in treated and control groups for reasons other than effects caused by the treatment. Observational studies are the norm when treatments are harmful, unwanted, or simply beyond control by the investigator.

In the absence of random assignment, treated and control groups may not be comparable at baseline before treatment. Baseline differences that have been accurately measured in observed covariates can often be removed by **matching**, **stratification** or model based adjustments [2, 28, 29]. However, there is usually the concern that some important baseline differences were not measured, so that individuals who appear comparable may not be. A sensitivity analysis in an observational study addresses this possibility: it asks what the unmeasured covariate would have to be like to alter the conclusions of the study. Observational studies vary markedly in their sensitivity to hidden bias: some are sensitive to very small biases, while others are insensitive to quit large biases.

The First Sensitivity Analysis

The first **sensitivity analysis** in an observational study was conducted by Cornfield, et al. [6] for certain observational studies of cigarette smoking as a cause of lung cancer; see also [10]. Although the tobacco industry and others had often suggested that cigarettes might not be the cause of high rates of lung cancer among smokers, that some other difference between smokers and nonsmokers might be the cause, Cornfield, et al. found that such an unobserved characteristic would need to be a near perfect predictor of lung cancer and about nine times more common among smokers than among nonsmokers. While this sensitivity analysis does not rule out the possibility that such a characteristic might exist, it does clarify what a scientist must logically be prepared to assert in order to defend such a claim.

Methods of Sensitivity Analysis

Various methods of sensitivity analysis exist. The method of Cornfield, et al. [6] is perhaps the best known of these, but it is confined to binary responses; moreover, it ignores sampling variability, which is hazardous except in very large studies. A method of sensitivity analysis that is similar in spirit to the method of Cornfield et al. will be described here; however, this alternative method takes account of sampling variability and is applicable to any kind of response; see, for instance, [25, 26, 29], and Section 4 of [28] for detailed discussion. Alternative methods of sensitivity analysis are described in [1, 5, 8, 9, 14, 18, 19, 23, 24], and [33].

The sensitivity analysis imagines that in the population before matching or stratification, subjects are assigned to treatment or control independently with unknown probabilities. Specifically, two subjects who look the same at baseline before treatment – that is, two subjects with the same observed covariates – may nonetheless differ in terms of unobserved covariates, so that one subject has an odds of treatment that is up to $\Gamma \geq 1$ times greater than the odds for another subject. In the simplest randomized experiment, everyone has the same chance of receiving the treatment, so $\Gamma = 1$. If $\Gamma = 2$ in an observational study, one subject might be twice as likely as another to receive the treatment because of unobserved pretreatment differences. The sensitivity analysis asks how much hidden bias can be present – that is, how large can Γ be – before the qualitative conclusions of the study begin to change. A study is highly sensitive to hidden bias if the conclusions change for Γ just barely larger than 1, and it is insensitive if the conclusions change only for quite large values of Γ.

If $\Gamma > 1$, the treatment assignments probabilities are unknown, but unknown only to a finite degree measured by Γ. For each fixed $\Gamma \geq 1$, the sensitivity analysis computes bounds on inference quantities, such as P values or **confidence intervals**. For $\Gamma = 1$, one obtains a single P value, namely the P value for a randomized experiment [7, 16, 17]. For each $\Gamma > 1$, one obtains not a single P value, but rather an interval of P values reflecting uncertainty due to hidden bias. As Γ increases, this interval becomes longer, and eventually it become uninformative, including both large and small P values. The point, Γ, at which the interval becomes uninformative is a measure of sensitivity to hidden bias. Computations are briefly described in Section titled 'Sensitivity Analysis Computations' and an example is discussed in detail in Section titled 'Sensitivity Analysis: Example'.

Sensitivity Analysis Computations

The straightforward computations involved in a sensitivity analysis will be indicated briefly in the case of one standard test, namely Wilcoxon's signed rank test for matched pairs (*see* **Distribution-free Inference, an Overview**) [17]. For details in this case [25] and many others, see Section 4 of [28]. The null hypothesis asserts that the treatment is without effect, that each subject would have the same response under the alternative treatment. There are S pairs, $s = 1, \ldots, S$ of two subjects, one treated, one control, matched for observed covariates. The distribution of treatment assignments within pairs is simply the conditional distribution for the model in Section titled 'Methods of Sensitivity Analysis' given that each pair includes one treated subject and one control. Each pair yields a treated-minus-control difference in outcomes, say D_s. For brevity in the discussion here, the D_s will be assumed to be untied, but ties are not a problem, requiring only slight change to formulas. The absolute differences, $|D_s|$, are ranked from 1 to S, and Wilcoxon's signed rank statistic, W, is the sum of the ranks of the positive differences, $D_s > 0$.

For the signed rank statistic, the elementary computations for a sensitivity analysis closely parallel the elementary computations for a conventional analysis. This paragraph illustrates the computations and may be skipped. In a moderately large randomized experiment, under the null hypothesis of no effect, W is approximately normally distributed with expectation $S\,(S+1)\,/4$ and variance $S\,(S+1)\,(2S+1)\,/24$; see Chapter 3 of [17]. If one observed $W = 300$ with $S = 25$ pairs in a randomized experiment, one would compute $S\,(S+1)/4 = 162.5$ and $S\,(S+1)\,(2S+1)/24 = 1381.25$, and the deviate $Z = (300 - 162.5)/\surd(1381.25) = 3.70$ would be compared to a Normal distribution to yield a one-sided P value of 0.0001. In a moderately large observational study, under the null hypothesis of no effect, the distribution of W is approximately bounded between two Normal distributions, with expectations $\mu_{\max} = \lambda\, S\,(S+1)/2$ and $\mu_{\min} = (1-\lambda)\, S\,(S+1)/2$, and the same variance $\sigma^2 = \lambda\,(1-\lambda)\, S\,(S+1)\,(2S+1)/6$, where $\lambda = \Gamma/\,(1+\Gamma)$. Notice that if $\Gamma = 1$, these expressions are the same as in the randomized experiment. For $\Gamma = 2$ and $W = 300$ with $S = 25$ pairs, one computes $\lambda = 2/\,(1+2) = 2/3$, $\mu_{\max} = (2/3)\,25\,(25+1)/2 = 216.67$, $\mu_{\min} = (1/3)\,25\,(25+1)/2 = 108.33$, and $\sigma^2 = (2/3)\,(1/3)\,25\,(25+1)\,(2 \times 25+1)/6 = 1227.78$; then two deviates are calculated, $Z_1 = (300 - 108.33)\,/\surd(1227.78) = 5.47$ and $Z_2 = (300 - 108.33)/\surd(1227.78) = 2.38$, which are compared to a Normal distribution, yielding a range of P values from 0.00000002 to 0.009. In other words, a bias of magnitude $\Gamma = 2$ creates some uncertainty about the correct P value, but it would leave no doubt that the difference is significant at the conventional 0.05 level.

Just as W has an exact randomization distribution useful for small S, so too there are exact sensitivity bounds. See [31] for details including S-Plus code.

Sensitivity Analysis: Example

A Matched Observational Study of an Occupational Hazard

Studies of occupational health usually focus on workers, but Morton, Saah, Silberg, Owens, Roberts and Saah [20] were worried about the workers' children. Specifically, they were concerned that workers exposed to lead might bring lead home in clothes and hair, thereby exposing their children as well. They matched 33 children whose fathers worked in a battery factory to 33 unexposed control children of the

Table 1 Blood lead levels, in micrograms of lead per decaliter of blood, of exposed children whose fathers worked in a battery factory and age-matched control children from the neighborhood. Exposed father's lead exposure at work (high, medium, low) and hygiene upon leaving the factory (poor, moderate, good) are also given. Adapted for illustration from Tables 1, 2 and 3 of Morton, et al. (1982). Lead absorption in children of employees in a lead-related industry, *American Journal of Epidemiology* **115**, 549–555. [20]

Pair s	Exposure	Hygiene	Exposed child's Lead level μg/dl	Control child's Lead level μg/dl	Dose Score
1	high	good	14	13	1.0
2	high	moderate	41	18	1.5
3	high	poor	43	11	2.0
⋮	⋮	⋮	⋮	⋮	⋮
33	low	poor	10	13	1.0
Median			34	16	

same age and neighborhood, and used Wilcoxon's signed rank test to compare the level of lead found in the children's blood, measured in μ*g* of lead per decaliter (dl) of blood. They also measured the father's level of exposure to lead at the factory, classified as high, medium, or low, and the father's hygiene upon leaving the factory, classified as poor, moderate, or good. Table 1 is adapted for illustration from Tables 1, 2, and 3 of Morton, et al. (1982) [20]. The median lead level for children of exposed fathers was more than twice that of control children, 34 μg/dl versus 16 μg/dl.

If Wilcoxon's signed rank test W is applied to the exposed-minus-control differences in Table 1, then the difference is highly significant in a one-sided test, $P < 0.0001$. This significance level would be appropriate in a randomized experiment, in which children were picked at random for lead exposure. Table 2 presents the sensitivity analysis, computed as in Section titled 'Sensitivity Analysis Computations'. Table 2 gives the range of possible one-sided significance levels for several possible magnitudes of hidden bias, measured by Γ. Even if the matching of exposed and control children had failed to control an unobserved characteristic strongly related to blood lead levels and $\Gamma = 4.25$ times more common among exposed children, this would still not explain the higher lead levels found among exposed children.

Where Table 2 focused on significance levels, Table 3 considers the one sided 95% confidence interval, $[\widehat{\tau}_{\text{low}}, \infty)$, for an additive effect obtained by inverting the signed rank test [28]. If the data in Table 1 had come from a randomized experiment ($\Gamma = 1$) with an additive treatment effect τ, then we would be 95% confident that father's lead exposure had increased his child's lead level by $\widehat{\tau}_{\text{low}} = 10.5$ μg/dl [17]. In an observational study with $\Gamma > 1$, there is a range of possible endpoints for the 95% confidence interval, and Table 3 reports the smallest value in the range. Even if $\Gamma = 3$, we would

Table 2 Sensitivity analysis for one-sided significance levels in the lead data. For unobserved biases of various magnitudes, the table gives the range of possible significance levels

Γ	min	max
1	<0.0001	<0.0001
2	<0.0001	0.0018
3	<0.0001	0.0136
4	<0.0001	0.0388
4.25	<0.0001	0.0468
5	<0.0001	0.0740

Table 3 Sensitivity analysis for one-sided confidence intervals for an additive effect in the lead data. For unobserved biases of biases of various magnitudes, the table gives smallest possible endpoint for the one-sided confidence interval

Γ	min $\widehat{\tau}_{\text{low}}$
1	10.5
2	5.5
3	2.5
4	0.5
4.25	0.0
5	−1.0

Table 4 Sensitivity to hidden bias in four observational studies. The randomization test assuming no hidden bias is highly significant in all four studies, but the magnitude of hidden bias that could alter this conclusion varies markedly between the four studies

Treatment	$\Gamma = 1$	(Γ, max P value)
Smoking/Lung Cancer Hammond [11]	<0.0001	(5, 0.03)
Diethylstilbestrol/ vaginal cancer Herbst, et al. [12]	<0.0001	(7, 0.054)
Lead/Blood lead Morton, et al. [20]	<0.0001	(4.25, 0.047)
Coffee/MI Jick, et al. [15]	0.0038	(1.3, 0.056)

be 95% confident exposure increased lead levels by 2.5 μg/dl.

Studies Vary in Their Sensitivity to Hidden Bias

Studies vary markedly in their sensitivity to hidden bias. As an illustration, Table 4 compares the sensitivity of four studies, a study of smoking as a cause of lung cancer [11], a study of prenatal exposure to diethylstilbestrol as a cause of vaginal cancer [12], the lead exposure study [20], and a study of coffee as a cause of myocardial infarction [15].

If no effect is tested using a conventional test appropriate for a randomized experiment ($\Gamma = 1$), the results are highly significant in all four studies. The last column of Table 4 indicates sensitivity to hidden bias, quoting the magnitude of hidden bias $\Gamma \geq 1$ needed to produce an upper bound on the P value close to the conventional 0.05 level. The study [12] of the effects of diethylstilbestrol becomes sensitive at about $\Gamma = 7$, while the study [15] of the effects of coffee becomes sensitive at $\Gamma = 1.3$. A small bias could explain away the effects of coffee, but only an enormous bias could explain away the effects of diethylstilbestrol. The lead exposure study, although quite insensitive to hidden bias, is about halfway between these two other studies, and is slightly more sensitive to hidden bias than the study of the effects of smoking.

Reducing Sensitivity to Hidden Bias

Accurately predicting a highly specific pattern of associations between treatment and response is often said to strengthen the evidence that the effects of the treatment caused the association. For instance, Cook, Campbell, and Peracchio [3] write: ‘Conclusions are more plausible if they are based on evidence that corroborates numerous, complex, or numerically precise predictions drawn from a descriptive causal hypothesis.’ Hill [13] and Weiss [34] emphasized the role of a dose response relationship. Cook and Shadish [4] write: ‘Successful prediction of a complex pattern of multivariate results often leaves few plausible alternative explanations.’

Does successful prediction of a complex pattern of associations affect sensitivity to hidden bias? It may, or it may not, and the degree to which it has done so can be appraised using methods similar to those in Section titled ‘Sensitivity Analysis Computations’. See [27] and [30] for methods of analysis, and [32] for issues in research design. The issues will be illustrated using the example in Table 1.

Recall that exposed fathers were classified by their degree of exposure and their hygiene upon leaving the factory. If the fathers’ exposure to lead at work were the cause of the higher lead levels among exposed children, then one would expect more lead in the blood of children whose fathers had higher exposure and poorer hygiene. Here, exposed children are divided into three groups of roughly similar size. The 13 exposed children in the category (high exposure, poor hygiene) were assigned a score of 2.0. Low exposure with any hygiene was assigned a score of 1, as was good hygiene with any exposure, and there were 12 such exposed children. The remaining 8 exposed children in intermediate situations were assigned a score of 1.5; they had either high exposure with moderate hygiene or medium exposure with poor hygiene. (None of the 33 matched children had medium exposure with moderate hygiene, although one unmatched child not used here fell into this category.)

The coherent or dose signed rank statistic D gives greater weight to matched pairs with higher doses [27, 30]. Table 5 compares the sensitivity to hidden bias of the usual Wilcoxon signed rank test W, which ignores doses, to the sensitivity of the coherent signed rank statistic. In particular, Table 5 gives the upper bound on the one-sided significance level for testing no effect. For W, this is the same computation as in Table 2. In fact, the coherent pattern of associations is somewhat less sensitive to hidden bias in this example: the upper bound on the

Table 5 Coherent patterns of associations can reduce sensitivity to hidden bias. Upper bounds on one-sided significance levels in the lead data, ignoring and using dose information

Γ	Wilcoxon W	Coherent D
1	<0.0001	<0.0001
3	0.0136	0.0119
4.35	0.0502	0.0398
4.75	0.0645	0.0503

P value for W ignoring doses is just above 0.05 at $\Gamma = 4.35$, but using doses with D the corresponding value is $\Gamma = 4.75$.

Exposed children had higher lead levels than unexposed controls, and also exposed children with higher exposures had higher lead levels than exposed children with lower lead levels. A larger hidden bias is required to explain this pattern of associations than is required to explain the difference between exposed and control children. In short, accurate prediction of a pattern of associations may reduce sensitivity to hidden bias, and whether this has happened, and the degree to which it has happened, may be appraised by a sensitivity analysis.

Acknowledgment

This work was supported by grant SES-0345113 from the US National Science Foundation.

References

[1] Berk, R.A. & De Leeuw, J. (1999). An evaluation of California's inmate classification system using a generalized regression discontinuity design, *Journal of the American Statistical Association* **94**, 1045–1052.

[2] Cochran, W.G. (1965). The planning of observational studies of human populations (with Discussion), *Journal of the Royal Statistical Society, Series A* **128**, 134–155.

[3] Cook, T.D., Campbell, D.T. & Peracchio, L. (1990). Quasi-experimentation, in *Handbook of Industrial and Organizational Psychology*, M. Dunnette & L. Hough, eds, Consulting Psychologists Press, Palo Alto, pp. 491–576.

[4] Cook, T.D. & Shadish, W.R. (1994). Social experiments: some developments over the past fifteen years, *Annual Review of Psychology* **45**, 545–580.

[5] Copas, J.B. & Li, H.G. (1997). Inference for nonrandom samples (with discussion), *Journal of the Royal Statistical Society* **B 59**, 55–96.

[6] Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. & Wynder, E. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions, *Journal of the National Cancer Institute* **22**, 173–203.

[7] Fisher, R.A. (1935). *The Design of Experiments*, Oliver & Boyd, Edinburgh.

[8] Gastwirth, J.L. (1992). Methods for assessing the sensitivity of statistical comparisons used in title VII cases to omitted variables, *Jurimetrics* **33**, 19–34.

[9] Gastwirth, J.L., Krieger, A.M. & Rosenbaum, P.R. (1998b). Dual and simultaneous sensitivity analysis for matched pairs, *Biometrika* **85**, 907–920.

[10] Greenhouse, S. (1982). Jerome Cornfield's contributions to epidemiology, *Supplement of Biometrics* **38**, 33–45.

[11] Hammond, E.C. (1964). Smoking in relation to mortality and morbidity: findings in first thirty-four months of follow-up in a prospective study started in 1959, *Journal of the National Cancer Institute* **32**, 1161–1188.

[12] Herbst, A., Ulfelder, H. & Poskanzer, D. (1971). Adenocarcinoma of the vagina: Association of maternal stilbestrol therapy with tumor appearance in young women, *New England Journal of Medicine* **284**, 878–881.

[13] Hill, A.B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* **58**, 295–300.

[14] Imbens, G.W. (2003). Sensitivity to exogeneity assumptions in program evaluation, *American Economic Review* **93**, 126–132.

[15] Jick, H., Miettinen, O., Neff, R., Jick, H., Miettinen, O.S., Neff, R.K., Shapiro, S., Heinonen, O.P., Slone, D. (1973). Coffee and myocardial infarction, *New England Journal of Medicine*, **289**, 63–77.

[16] Kempthorne, O. (1952). *Design and Analysis of Experiments*, John Wiley & Sons, New York.

[17] Lehmann, E.L. (1998). *Nonparametrics: Statistical Methods Based on Ranks*, Prentice Hall, Upper Saddle River.

[18] Lin, D.Y., Psaty, B.M. & Kronmal, R.A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies, *Biometrics* **54**, 948–963.

[19] Manski, C. (1995). *Identification Problems in the Social Sciences*, Harvard University Press, Cambridge.

[20] Morton, D., Saah, A., Silberg, S., Owens, W., Roberts, M. & Saah, M. (1982). Lead absorption in children of employees in a lead-related industry, *American Journal of Epidemiology* **115**, 549–555.

[21] Peto, R., Pike, M., Armitage, P., Breslow, N., Cox, D., Howard, S., Mantel, N., McPherson, K., Peto, J. & Smith, P. (1976). Design and analysis of randomised clinical trials requiring prolonged observation of each patient, I, *British Journal of Cancer* **34**, 585–612.

[22] Piantadosi, S. (1997). *Clinical Trials*, Wiley, New York.

[23] Robins, J.M., Rotnitzkey, A. & Scharfstein, D. (1999). Sensitivity analysis for selection bias and unmeasured

confounding in missing data and causal inference models, in *Statistical Models in Epidemiology*, E. Halloran & D. Berry, eds, Springer, New York, pp. 1–94.

[24] Rosenbaum, P.R. (1986). Dropping out of high school in the United States: an observational study, *Journal of Educational Statistics* **11**, 207–224.

[25] Rosenbaum, P.R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies, *Biometrika* **74**, 13–26.

[26] Rosenbaum, P.R. (1995). Quantiles in nonrandom samples and observational studies, *Journal of the American Statistical Association* **90**, 1424–1431.

[27] Rosenbaum, P.R. (1997). Signed rank statistics for coherent predictions, *Biometrics* **53**, 556–566.

[28] Rosenbaum, P.R. (2002a). *Observational Studies*, Springer, New York.

[29] Rosenbaum, P.R. (2002b). Covariance adjustment in randomized experiments and observational studies (with Discussion), *Statistical Science* **17**, 286–327.

[30] Rosenbaum, P.R. (2003a). Does a dose-response relationship reduce sensitivity to hidden bias? *Biostatistics* **4**, 1–10.

[31] Rosenbaum, P.R. (2003b). Exact confidence intervals for nonconstant effects by inverting the signed rank test, *American Statistician* **57**, 132–138.

[32] Rosenbaum, P.R. (2004). Design sensitivity in observational studies, *Biometrika* **91**, 153–164.

[33] Rosenbaum, P.R. & Rubin, D.B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome, *Journal of the Royal Statistical Society Series B* **45**, 212–218.

[34] Weiss, N. (1981). Inferring causal relationships: elaboration of the criterion of 'dose-response.', *American Journal of Epidemiology* **113**, 487–90.

PAUL R. ROSENBAUM

Sequential Decision Making

Hans J. Vos

Volume 4, pp. 1815–1819

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Sequential Decision Making

Well-known examples of fixed-length mastery tests in the behavioral sciences include pass/fail decisions in education, certification, and successfulness of therapies. The fixed-length mastery problem has been studied extensively in the literature within the framework of (empirical) Bayesian decision theory [9] (*see* **Bayesian Item Response Theory Estimation**). In this approach, the following two basic elements are distinguished: A measurement model relating the probability of a correct response to student's (unknown) true level of functioning, and a loss structure evaluating the total costs and benefits for each possible combination of decision outcome and true level of functioning. Within the framework of Bayesian decision theory [2, 5], optimal rules (i.e., Bayes rules) are obtained by minimizing the posterior expected losses associated with all possible decision rules. Decision rules are hereby prescriptions specifying for each possible observed response pattern what action has to be taken. The Bayes principle assumes that prior knowledge about student's true level of functioning is available and can be characterized by a probability distribution called *the prior*. This prior probability represents our best prior beliefs concerning student's true level of functioning; that is, before any item yet has been administered.

The test at the end of the treatment does not necessarily have to be a fixed-length mastery test but might also be a sequential mastery test (SMT). In this case, in addition to the actions declaring mastery or nonmastery, also the action of to continue testing and administering another random item is available. Sequential mastery tests are designed with the goal of maximizing the probability of making correct classification decisions (i.e., mastery and nonmastery) while at the same time minimizing test length [6]. For instance, Ferguson [3] showed that average test lengths could be reduced by half without sacrificing classification accuracy.

The purpose of this entry is to derive optimal rules for SMT in the context of education using the framework of Bayesian sequential decision theory [2, 5]. The main advantage of this approach is that costs of testing (i.e., administering another random item) can be taken explicitly into account.

Bayesian Sequential Principle Applied to SMT

It is indicated in this section how the framework of Bayesian sequential decision theory, in combination with the binomial distribution for modeling response behavior (i.e., the measurement model) and adopting threshold loss for the loss function involved, is applied to SMT.

Framework of Bayesian Sequential Decision Theory

Three basic elements can be identified in Bayesian sequential decision theory. In addition to a measurement model and a loss function, costs of administering one additional item must be explicitly specified in this approach. Doing so, posterior expected losses corresponding to the mastery and nonmastery decisions can now be calculated straightforward at each stage of testing. As far as the posterior expected loss corresponding to continue testing concerns, this quantity is determined by averaging the posterior expected losses corresponding to each of the possible future classification outcomes relative to observing those outcomes (i.e., the posterior predictive distributions). Optimal rules (i.e., Bayesian sequential rules) are now obtained by minimizing the posterior expected losses associated with all possible decision rules at each stage of testing using techniques of backward induction (i.e., dynamic programming). This technique starts by considering the final stage of testing (where the option to continue testing is not available) and then works backward to the first stage of testing.

Notation

In order to classify within a reasonable period of time those students for whom the decision of declaring mastery or nonmastery is not as clear-cut, a sequential mastery test is supposed to have a maximum length of n $(n \geq 1)$. Let the observed item response at each stage of testing k $(1 \leq k \leq n)$ for a randomly sampled student be denoted by a discrete random variable X_k, with realization x_k. The observed response variables $X_1, \ldots, X_k$ are assumed to be independent and identically distributed for each value of k, and take the values 0 and 1 for respectively incorrect and correct responses to item k. Furthermore, let the observed number-correct score after

k items have been administered be denoted by a discrete random variable $S_k = X_1 + \cdots + X_k$, with realization $s_k = x_1 + \cdots + x_k$ $(0 \leq s_k \leq k)$.

Student's true level of functioning is unknown due to measurement and sampling error. All that is known is his/her observed number-correct score s_k. In other words, the mastery test is not a perfect indicator of student's true performance. Therefore, let student's (unknown) true level of functioning be denoted by a continuous random variable T on the latent proportion-correct metric, with realization t $(0 \leq t \leq 1)$.

Finally, a criterion level t_c $(0 \leq t_c \leq 1)$ on T must be specified in advance by the decision-maker using methods of standard-setting (e.g., [1]). A student is considered a true nonmaster and true master if his/her true level of functioning t is smaller or larger than t_c, respectively.

Threshold Loss and Costs of Testing

Generally speaking, as noted before, a loss function evaluates the total costs and benefits of all possible decision outcomes for a student whose true level of functioning is t. These costs may concern all relevant psychological, social, and economic consequences which the decision brings along. As in [6], here the well-known threshold loss function is adopted as the loss structure involved. The choice of this loss function implies that the 'seriousness' of all possible consequences of the decisions can be summarized by possibly different constants, one for each of the possible classification outcomes.

For the sequential mastery problem, following Lewis and Sheehan [6], a threshold loss function can be formulated as a natural extension of the one for the fixed-length mastery problem at each stage of testing k as shown in Table 1.

The value e represents the costs of administering one random item. For the sake of simplicity, these costs are assumed to be equal for each classification outcome as well as for each testing occasion. Applying an admissible positive linear transformation [7], and assuming the losses l_{00} and l_{11} associated with the correct classification outcomes are equal and take the smallest values, the threshold loss function in Table 1 was rescaled in such a way that l_{00} and l_{11} were equal to zero. Hence, the losses l_{01} and l_{10} must take positive values.

Table 1 Table for threshold loss function at stage k $(1 \leq k \leq n)$ of testing

	True level of functioning	
Decision	$T \leq t_c$	$T > t_c$
Declaring nonmastery	ke	$l_{01}+ ke$
Declaring mastery	$l_{10}+ ke$	ke

The ratio l_{10}/l_{01} is denoted as the loss ratio R, and refers to the relative losses for declaring mastery to a student whose true level of functioning is below t_c (i.e., false positive) and declaring nonmastery to a student whose true level of functioning exceeds t_c (i.e., false negative).

The loss parameters l_{ij} $(i, j = 0, 1; i \neq j)$ associated with the incorrect decisions have to be empirically assessed, for which several methods have been proposed in the literature. Most texts on decision theory, however, propose lottery methods [7] for assessing loss functions empirically. In general, the consequences of each pair of actions and true level of functioning are scaled in these methods by looking at the most and least preferred outcomes. But, in principle, any psychological scaling method can be used.

Measurement Model

Following Ferguson [3], in the present entry the well-known binomial model will be adopted for the probability that after k items have been administered, s_k of them have been answered correctly (*see* **Catalogue of Probability Density Functions**). Its distribution at stage k of testing for given student's true level of functioning t, $\mathrm{P}(s_k|t)$, can be written as follows:

$$\mathrm{P}(s_k|t) = \binom{k}{s_k} t^{s_k}(1-t)^{k-s_k}. \qquad (1)$$

If each response is independent of the other, and if student's probability of a correct answer remains constant, the distribution function of s_k, given true level of functioning t, is given by (1). The binomial model assumes that the test given to each student is a random sample of items drawn from a large (real or imaginary) item pool [10]. Therefore, for each student a new random sample of items must be drawn in practical applications of the sequential mastery problem.

Optimizing Rules for the Sequential Mastery Problem

In this section, it will be shown how optimal rules for SMT can be derived using the framework of Bayesian sequential decision theory. Doing so, given an observed item response vector $(x_1, \ldots, x_k)$, first the Bayesian principle will be applied to the fixed-length mastery problem by determining which of the posterior expected losses associated with the two classification decisions is the smallest. Next, applying the Bayesian principle again, optimal rules for the sequential mastery problem are derived at each stage of testing k by comparing this quantity with the posterior expected loss associated with the option to continue testing.

Applying the Bayesian Principle to the Fixed-length Mastery Problem

As noted before, the Bayesian decision rule for the fixed-length mastery problem can be found by minimizing the posterior expected losses associated with the two classification decisions of declaring mastery or nonmastery. In doing so, the posterior expected loss is taken with respect to the posterior distribution of T. It can easily be verified from Table 1 and (1) that mastery is declared when the posterior expected loss corresponding to declaring mastery is smaller than the posterior expected loss corresponding to declaring nonmastery, or, equivalently, when s_k is such that

$$\begin{aligned}(l_{10} + ke)\mathrm{P}(T \le t_c|s_k) + (ke)\mathrm{P}(T > t_c|s_k)\\ < (ke)\mathrm{P}(T \le t_c|s_k) + (l_{01} + ke)\mathrm{P}(T > t_c|s_k),\end{aligned} \tag{2}$$

and that nonmastery is declared otherwise. Rearranging terms, it can easily be verified from (2) that mastery is declared when s_k is such that

$$\mathrm{P}(T \le t_c|s_k) < \frac{1}{1+R}, \tag{3}$$

and that nonmastery is declared otherwise.

Assuming a beta prior for T, it follows from an application of Bayes' theorem (*see* **Bayesian Belief Networks**) that under the assumed binomial model from (1), the posterior distribution of T will be a member of the beta family again (the conjugacy property, see, e.g., [5]). In fact, if the beta function $B(\alpha, \beta)$ with parameters α and β (α, $\beta > 0$) is chosen as prior distribution (i.e., the natural conjugate of the binomial distribution) and student's observed number-correct score is s_k from a test of length k, then the posterior distribution of T is $B(\alpha + s_k, k - s_k + \beta)$. Hence, assuming a beta prior for T, it follows from (3) that mastery is declared when s_k is such that

$$B(\alpha + s_k, k - s_k + \beta) < \frac{1}{1+\mathrm{R}}, \tag{4}$$

and that nonmastery is declared otherwise.

The uniform distribution on the standard interval [0,1] as a noninformative prior will be assumed in this entry, which results as a special case of the beta distribution $B(\alpha, \beta)$ for $\alpha = \beta = 1$ (*see* **Catalogue of Probability Density Functions**). In other words, prior true level of functioning can take on all values between 0 and 1 with equal probability. It then follows immediately from (4) that mastery is declared when s_k is such that

$$B(1 + s_k, k - s_k + 1) < \frac{1}{1+\mathrm{R}}, \tag{5}$$

and that nonmastery is declared otherwise. The beta distribution has been extensively tabulated [8]. Normal approximations are also available [4].

Derivation of Bayesian Sequential Rules

Let $d_k(x_1, \ldots, x_k)$ denote the decision rule yielding the minimum of the posterior expected losses associated with the two classification decisions at stage k of testing, and let the posterior expected loss corresponding to this minimum be denoted as $V_k(x_1, \ldots, x_k)$. Bayesian sequential rules can now be found by using the following backward induction computational scheme: First, the Bayesian sequential rule at the final stage n of testing is computed. Since the option to continue testing is not available at this stage of testing, it follows immediately that the Bayesian sequential rule is given by $d_n(x_1, \ldots, x_n)$, and its corresponding posterior expected loss by $V_n(x_1, \ldots, x_n)$.

To compute the posterior expected loss associated with the option to continue testing at stage $(n - 1)$ until stage 0, the risk $R_k(x_1, \ldots, x_k)$ will be introduced at each stage k $(1 \le k \le n)$ of testing. Let the risk at stage n be defined as $V_n(x_1, \ldots, x_n)$. Generally, given response pattern $(x_1, \ldots, x_k)$, the risk at stage $(k-1)$ is then computed inductively as a function of the risk at stage k as:

$$R_{k-1}(x_1, \ldots, x_{k-1}) = \min\{V_{k-1}(x_1, \ldots, x_{k-1}),\\ E[R_k(x_1, \ldots, x_{k-1}, X_k)|x_1, \ldots, x_{k-1}]\}. \qquad (6)$$

The posterior expected loss corresponding to administering one more random item after $(k-1)$ items have been administered, $E[R_k(x_1, \ldots, x_{k-1}, X_k)|x_1, \ldots, x_{k-1}]$, can then be computed as the expected risk at stage k of testing as

$$E[R_k(x_1, \ldots, x_{k-1}), X_k|x_1, \ldots, x_{k-1}]\\ = \sum_{x_k=0}^{x_k=1} R_k(x_1, \ldots, x_k)P(X_k = x_k|x_1, \ldots, x_{k-1}), \qquad (7)$$

where $P(X_k = x_k|x_1, \ldots, x_{k-1})$ denotes the posterior predictive distribution of X_k at stage $(k-1)$ of testing. Computation of this conditional distribution is deferred until the next section. Note that (7) averages the posterior expected losses associated with each of the possible future classification outcomes with weights corresponding to the probabilities of observing those outcomes.

The Bayesian sequential rule at stage $(k-1)$ is now given by: Administer one more random item if $E[R_k(x_1, \ldots, x_{k-1}, X_k)|x_1, \ldots, x_{k-1}]$ is smaller than $V_{k-1}(x_1, \ldots, x_{k-1})$; otherwise, decision d_{k-1} $(x_1, \ldots, x_{k-1})$ is taken. The Bayesian sequential rule at stage 0 denotes the decision whether or not to administer at least one random item.

Computation of Posterior Predictive Distributions

As is clear from (7), the posterior predictive distribution $P(X_k = x_k|x_1, \ldots, x_{k-1})$ is needed for computing the posterior expected loss corresponding to administering one more random item at stage $(k-1)$ of testing. Assuming the binomial distribution as measurement model and the uniform distribution $B(1,1)$ as prior, it was shown (e.g., [5]) that $P(X_k = 1|x_1, \ldots, x_{k-1}) = (1 + s_{k-1})/(k+1)$, and, thus, that $P(X_k = 0|x_1, \ldots, x_{k-1}) = [1 - (1 + s_{k-1})/(k+1)] = (k - s_{k-1})/(k+1)$.

Determination of Appropriate Action for Different Number-correct Score

Using the general backward induction scheme discussed earlier, for a given maximum number n $(n \geq 1)$ of items to be administered, a program BAYES was developed to determine the appropriate action (i.e., nonmastery, continuation, or mastery) at each stage k of testing for different number-correct score s_k.

As an example, the appropriate action is depicted in Table 2 as a closed interval for a maximum of 20 items (i.e., $n = 20$). Students were considered as true masters if they knew at least 55% of the subject matter. Therefore t_c was fixed at 0.55. Furthermore, the loss corresponding to the false mastery decision was perceived twice as large as the loss corresponding to the false nonmastery decision (i.e., $R = 2$). On a scale in which one unit corresponded to the constant costs of administering one random item (i.e., $e = 1$), therefore, l_{10} and l_{01} were set equal to 200 and 100, respectively. These numerical values reflected the assumption that the losses corresponding to taking incorrect classification decisions were rather

Table 2 Appropriate action calculated by stage of testing and number-correct score

Stage of testing	Appropriate action by number-correct score		
	Nonmastery	Continuation	Mastery
0		0	
1		[0,1]	
2	0	[1,2]	
3	0	[1,3]	
4	[0,1]	[2,4]	
5	[0,1]	[2,4]	5
6	[0,2]	[3,5]	6
7	[0,2]	[3,5]	[6,7]
8	[0,3]	[4,6]	[7,8]
9	[0,4]	[5,7]	[8,9]
10	[0,4]	[5,7]	[8,10]
11	[0,5]	[6,8]	[9,11]
12	[0,5]	[6,8]	[9,12]
13	[0,6]	[7,9]	[10,13]
14	[0,7]	[8,9]	[10,14]
15	[0,7]	[8,10]	[11,15]
16	[0,8]	[9,10]	[11,16]
17	[0,9]	[10,11]	[12,17]
18	[0,10]	11	[12,18]
19	[0,11]		[12,19]
20	[0,12]		[13,20]

large relative to the costs of administering one random item.

As can be seen from Table 2, at least five random items need to be administered before mastery can be declared. However, in principle, nonmastery can be declared already after administering two random items. Also, generally a rather large number of items have to be answered correctly before mastery can be declared. This can be accounted for the relatively large losses corresponding to false positive decisions (i.e., 200) relative to the losses corresponding to false negative decisions (i.e., 100). In this way, relatively large posterior expected losses from taking false positive decisions can be avoided.

Discussion and Conclusions

In this entry, using the framework of Bayesian sequential decision theory, optimal rules for the sequential mastery problem (nonmastery, mastery, or to continue testing) in the context of education were derived. It should be emphasized, however, that the Bayes sequential principle is especially appropriate when costs of testing can be assumed to be quite large. For instance, when testlets (i.e., blocks of parallel items) rather than single items are considered. Also, the proposed strategy might be appropriate in the context of sequential testing problems in psychodiagnostics. Suppose that a new treatment (e.g., cognitive-analytic therapy) must be tested on patients suffering from some mental health problem (e.g., anorexia nervosa). Each time after having exposed a patient to the new treatment, it is desired to make a decision concerning the effectiveness/ineffectiveness of the new treatment or to continue testing and exposing the new treatment to another random patient suffering from the same mental health problem. In such clinical situations, costs of testing generally are quite large and the Bayesian sequential approach might be considered as an alternative to fixed-length mastery tests.

References

[1] Angoff, W.H. (1971). Scales, norms and equivalent scores, in *Educational Measurement*, 2nd Edition, R.L. Thorndike, ed., American Council on Education, Washington, pp. 508–600.

[2] DeGroot, M.H. (1970). *Optimal Statistical Decisions*, McGraw-Hill, New York.

[3] Ferguson, R.L. (1969). *The Development, Implementation, and Evaluation of a Computer-Assisted Branched Test for a Program of Individually Prescribed Instruction*, Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh.

[4] Johnson, N.L. & Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions*, Houghton Mifflin, Boston.

[5] Lehmann, E.L. (1959). *Testing Statistical Hypotheses*, 3rd Edition, Macmillan Publishers, New York.

[6] Lewis, C. & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test, *Applied Psychological Measurement* **14**, 367–386.

[7] Luce, R.D. & Raiffa, H. (1957). *Games and Decisions*, John Wiley & Sons, New York.

[8] Pearson, K. (1930). *Tables for Statisticians and Biometricians*, Cambridge University Press, London.

[9] van der Linden, W.J. (1990). Applications of decision theory to test-based decision making, in *New Developments in Testing: Theory and Applications*, R.K. Hambleton & J.N. Zaal, eds, Kluwer Academic Publishers, Boston, pp. 129–155.

[10] Wilcox, R.R. (1981). A review of the beta-binomial model and its extensions, *Journal of Educational Statistics* **6**, 3–32.

HANS J. VOS

Sequential Testing

D.S. Coad

Volume 4, pp. 1819–1820

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Sequential Testing

When a **clinical trial** is carried out, it is often desirable to stop the trial early if there is convincing evidence that one treatment is superior to the others or there is clearly no difference between the treatments. The use of sequential methods can lead to substantial reductions in the numbers of patients required compared with a fixed-sample design in order to achieve the same **power**. The methods involve the monitoring of a test statistic, and the trial is stopped as soon as this statistic crosses some stopping boundary. The stopping boundary is chosen in order that the trial has a given type I error probability and power for detecting a prespecified treatment difference. Since the development of sequential methods in the 1940s, there is now a wide range of methods available. The method to be used depends on such considerations as the type of patient response being observed, the testing problem under consideration, and how many treatments there are.

Most of the work on sequential testing in the 1950s and 1960s focused on the fully-sequential case, that is, the data are inspected after each new patient's response. Many of the early sequential methods developed for clinical trials are described in the classic book [1]. The often impractical nature of fully-sequential methods led to the development of group-sequential methods in the 1970s. With these methods, the data are inspected after each group of patient responses, and approaches are available for both equal and unequal group sizes. For a given type I error probability and power, the values of the stopping boundary are computed numerically for each of the planned *interim analyses* [4]. Two statistical packages are available for the design of sequential clinical trials–PEST 4 [2] and EaSt-2000 [3]. Both packages also incorporate methods for drawing statistical inferences upon termination, such as **confidence intervals**.

As an example, suppose that we wish to compare two treatments A and B when patient response is binary. Then, the usual measure of treatment difference is the *log odds ratio* (*see* **Odds and Odds Ratios**) $\theta = \log\{p_A q_B/(p_B q_A)\}$, where p_A and p_B are the success probabilities for the two treatments, and $q_A = 1 - p_A$ and $q_B = 1 - p_B$. Following [5], one approach is to use the statistics

$$Z = \frac{n_B S_A - n_A S_B}{n_A + n_B} \quad \text{and}$$
$$V = \frac{n_A n_B (S_A + S_B)(n_A + n_B - S_A - S_B)}{(n_A + n_B)^3}, \qquad (1)$$

where n_A and n_B are the numbers of patients on the two treatments, and S_A and S_B are the numbers of successes. For example, if the probability of success for treatment B is expected to be about $p_B = 0.6$ and we wish to test whether treatment A leads to an improvement of $p_A = 0.8$, then, for a given type I error probability of $\alpha = 0.05$ and power of $1 - \beta = 0.9$, we can use PEST 4 to obtain the stopping boundary for the test. After each group of patients, the value of Z is plotted against the value of V until a point falls outside of the stopping boundaries. Depending on which boundary is crossed first, we either conclude that treatment A leads to an improvement or that there is insufficient evidence of a treatment difference.

References

[1] Armitage, P. (1975). *Sequential Clinical Trials*, 2nd Edition, Blackwell, Oxford.

[2] Brunier, H.C. & Whitehead, J. (2000). *Pest 4.0 Operating Manual*, University of Reading.

[3] Cytel Software Corporation (2001). *EaSt-2001: Software for the Design and Interim Monitoring of Group Sequential Clinical Trials*, Cytel Software Corporation, Cambridge.

[4] Jennison, C. & Turnbull, B.W. (2000). *Group Sequential Methods with Applications to Clinical Trials*, Chapman and Hall, London.

[5] Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*, Revised 2nd Edition, Wiley, Chichester.

D.S. Coad

Setting Performance Standards: Issues, Methods

Barbara S. Plake

Volume 4, pp. 1820–1823

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Setting Performance Standards: Issues, Methods

Standardized tests are used for many purposes. Such purposes include which students are eligible for high school graduation, which applicants pass a licensure test, and who, among qualified applicants, will receive scholarships or other awards and prices. In order to make these decisions, typically one or more score values is identified from the possible test scores to be the 'passing score' or 'cutscore'.

Setting performance standards, or cutscores, is often referred to as '*standard setting*'. This is because, by determining the passing score, the 'standard' is set for the score needed to pass the test. Standard setting methods are often differentiated by whether they focus on the test takers (the examinee population) or on the test questions. Methods are presented that illustrate both of these approaches.

Examinee Focused Methods

Two different examinee focused methods are discussed. The first method includes strategies whereby examinees are assigned to performance categories by someone who is qualified to judge their performance. The second method uses the score distribution on the test to make examinee classifications.

When using the first of these methods, people who know the examinees well (for example, their teachers) are asked to classify them into performance categories. These performance categories could be simply 'qualified to pass' or 'not qualified to pass', or more complex, such as 'Basic', 'Proficient' and 'Advanced'. When people make these classifications, they do not know how the examinees did on the test. After examinees are classified into the performance categories, the test score that best separates the classification categories is determined.

The second general approach to setting cutscores, using examinee performance data, is through 'norm-based methods'. When using norm-based methods, the scores from the current examinee group are summarized, calculating the average (or mean) of the set of scores and some measure of how spread out across the score range the test scores fall (variability). In some applications, the cutscore is set at the mean of the score distribution or the mean plus or minus a measure of score variability (standard deviation). Setting the passing score above the mean (say one standard deviation above the mean) would, in a bell-shaped score distribution, pass about 15% of the examinees; likewise, setting the passing score one standard deviation below the mean would fail about 15% of the examinees.

Test-based Methods

Test-based methods for setting passing scores consider the questions that comprise the test. Before discussing test-based methods, it is necessary to know more about the kinds of questions that comprise the test. Many tests are composed of multiple-choice test items. These items have a question and then several (often four) answer choices from which the examinee selects the right or best answer. Multiple-choice test items are often favored because they are quick and easy to score and, with careful test-construction efforts, can cover a broad range of content in a reasonable length of testing time.

Other tests have questions that ask the examinee to write an answer, not select an answer from a set list (such as with multiple-choice items). Sometimes these types of questions are called *constructed-response* questions because the examinee is required to construct a response. Some agencies find these kinds of questions appealing because they are seen as more directly related, in some situations, to the actual work that is required.

The methods used for setting passing scores will vary based on the type of questions and tasks that comprise the test. In every case, however, a panel of experts is convened (called *subject matter experts*, or SMEs). Their task is to work with the test content in determining the recommended minimum passing score for the test, or in some situation, the multiple cutscores for making performance classifications for examinees (such as Basic, Proficient, and Advanced).

Multiple-choice Questions

Until recently, most of the tests that were used in standard setting contained only multiple-choice questions. The reasons for this were mostly because of the ability to obtain a lot of information about the examinees' skill levels in a limited amount of time. The ease and accuracy of scoring were also a consideration in the popularity of the multiple-choice test question in high-stakes testing. Two frequently used standard setting methods for multiple-choice questions are described.

Angoff Method. The most prevalent method for setting cutscores on multiple-choice tests for making pass/fail decisions is the Angoff method [1]. Using this method, SMEs are asked to conceptualize an examinee who is just barely qualified to function at the designated performance level (called the *minimally competent candidate*, or MCC). Then, for each item in the test, the SMEs are asked to estimate the probability that a randomly selected MCC would be able to correctly answer the item. The SMEs work independently when making these predictions. Once these predictions have been completed, the probabilities assigned to the items in the test are added together for each SME. These estimates are then averaged across the SMEs to determine the overall minimum passing score. Because the SMEs will vary in their individual estimates of the minimum passing score, it is possible to also compute the variability in their minimum passing score estimates. The smaller the variability, the more cohesive the SMEs are in their estimates of the minimum passing score. Sometimes this variability is used to adjust the minimum passing score to be either higher or lower than the average value.

In most application of the Angoff standard setting method, more than one set of estimates is obtained from the SME. The first set (described above) is often called *Round 1*. After Round 1, SMEs are typically given additional information. For example, they may be told the groups' Round 1 minimum passing score value and the range, so they can learn about the level of cohesion of the panel at this point. In addition, it is common for the SMEs to be told how recent examinees performed on the test. The sharing of examinee performance information is somewhat controversial. However, it is often the case that SMEs need performance information as a reality check. If data are given to the SMEs, then it is customary to conduct a second round of ratings, called *Round 2*. The minimum passing scores are then calculated using the Round 2 data in a manner identical to that for the Round 1 estimates. Again, panels' variability may be used to adjust the final recommended minimum passing score.

There have been many modifications to the Angoff method, so many in fact that there is not a single, standard set of procedures that define the Angoff standard setting method. Variations include whether or not performance data is provided between Rounds, whether or not there is more than one round, whether or not SMEs may discuss their ratings, and how the performance estimates are made by the SMEs. Another difference in the application of the Angoff method is the definitions for the skill levels for the MCC.

Bookmark Method. Another standard setting method that is used with multiple-choice questions (and also with mixed formats that include multiple-choice and constructed-response question) is called the *Bookmark Method* [2]. In order to conduct this method, the test questions have to be assembled into a special booklet with one item per page and organized in ascending order of difficulty (easiest to hardest). SMEs are given a booklet and asked to page through the booklet until they encounter the first item that they believe that the MCC would have less than at 67% chance of answering correctly. They place their bookmark on the page prior to that item in the booklet. The number of items that precedes the location of the bookmark represents an individual SME's estimate of the minimum passing score. The percent level (here identified as 67%) is called the *response probability* (RP); RP values other than 67% are sometimes used. Individual SME estimates of the minimum passing score are shared with the group, usually graphically. After discussion, SMEs reconsider their initial bookmark location. This usually continues through multiple rounds, with SME's minimum passing scores estimates shared after each round. Typically there is large diversity in minimum passing score estimates after round 1, but the group often reaches a small variability in minimum passing score estimates following the second or third round.

Constructed-response Questions

As stated previously, constructed-response questions ask the examinee to prepare a response to a question. This could be a problem-solving task for mathematics, preparing a cognitive map for a reading passage, presenting a critical reasoning essay on a contemporary problem. What is common across these tasks is that the examinee cannot select an answer but rather must construct one. Another distinguishing feature of constructed-response questions is that they typically are worth more than one point. There is often a scoring rubric or scoring system used to determining the number of points an examinee's answer will receive for each test question.

There are several methods for setting standards on constructed-response tests. One method, called *Angoff extension* [3] asks the SMEs to estimate the total number of points that the MCC is likely to earn out of the total available for that test question. The calculation of the minimum passing score is similar to that for the Angoff method, except that the total number of points awarded to each question is added to calculate the minimum passing score estimate for each SME. As with the Angoff method, multiple rounds are usually conducted with some performance data shared with the SMEs between rounds.

Another method used with constructed-response questions is called the *Analytical Judgment* (AJ) method [5]. For this method, SME read examples of examinee responses and assign them into multiple performance categories. For example, if the purpose were to set one cutscore (say for Passing), the categories would be 'Clearly Failing', 'Failing', 'Nearly Passing', 'Just Barely Passing', 'Passing', and 'Clearly Passing'. Usually a total of 50 examinee responses are collected for each constructed-response questions, containing examples of low, middle, and high performance. Scores on the examples are not shown to the SMEs. After the SMEs have made their initial categorizations of the example papers into these multiple performance categories, the examples that are assigned to the 'Nearly Passing' and 'Just Barely Passing' categories are identified. The scores on these examples are averaged and the average score is used as the recommended minimum passing score. Again, the variability of scores that were assigned to these two performance categories may be used to adjust the minimum passing score. An advantage of this method is that actual examinee responses are used in the standard setting effort. A disadvantage of the method is that it considers the examinees' test responses question by question, without an overall consideration of the examinees' test performance. A variation of this method, called the *Integrated Judgment Method* [4] has SMEs consider the questions individually and then collectively in making only one overall test classification decision into the above multiple categories.

Other methods exist for use with constructed-response questions but they are generally consistent with the two approaches identified above. More information about these and other standard setting methods can be found in Cizek's 2001 book titled *Setting Performance Standards: Concepts, Methods, and Perspectives*.

Conclusion

Tests are used for a variety of purposes. Because decisions are based on test performance, the score value for passing the test must be determined. Typically, a standard setting procedure is used to identify recommended values for these cutscores. The final decision about the value of the cutscore is a policy decision that should be made by the governing agency.

Tests serve an important purpose. It is desired to certify that the examinee does have the requisite skills and competencies needed to graduate from school programs, practice in an occupation or profession, or receive elevated status within a profession. If the passing scores on these tests are not set appropriately, there is no assurance that these outcomes will be achieved. Therefore, it is critically important that sound methods are used when determining these cutscores.

References

[1] Angoff, W.H. (1971). Scales, norms, and equivalent scores, 2nd Edition, in *Educational Measurement*, R.L. Thorndike, ed., American Council on Education, Washington.

[2] Cizek, G.J., ed. (2001). *Setting Performance Standards: Concepts, Methods, and Perspectives*, Lawrence Erlbaum, Mahwah.

[3] Hambleton, R.K. & Plake, B.S. (1995). Extended Angoff procedure to set standards on complex performance assessments, *Applied Measurement in Education* **8**, 41–56.

[4] Jaeger, R.M. & Mills, C.N. (2001). An integrated judgment procedure for setting standards on complex, large-scale assessments, in *Setting Performance Standards:*

Concepts, Methods, and Perspectives, G.J. Cizek, ed., Lawrence Erlbaum, Mahwah.

[5] Plake, B.S. & Hambleton, R.K. (2000). A standard setting method for use with complex performance assessments: categorical assignments of student work, *Educational Assessment* **6**, 197–215.

BARBARA S. PLAKE

Sex-Limitation Models

Thalia C. Eley

Volume 4, pp. 1823–1826

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Sex-Limitation Models

Sex differences are one of the most marked and frequently reported features of behavioral phenotypes. Many aspects of emotional difficulties within the internalizing arena are more prevalent in females than males. For example, depression is roughly twice as common in women than in men [3]. In contrast, males show higher levels of externalizing or behavioral problems, with conduct problems two to three times as common in males than females [4]. Another area in which males show higher rates than females is in learning and communication difficulties. For example, language delays are more common in boys than girls, and autism is also more prevalent in males than females. As a result of these widespread sex differences, behavioral scientists have focused considerable attention on trying to identify methods by which to explore their origins. Sex limitation models use a **structural equation model**-fitting approach with twin or adoption data to test for both *quantitative* and *qualitative* sex differences in the genetic and environmental etiology of the phenotype.

A first step in any examination of sex differences is to establish whether there are mean or variance differences between the two sexes. Or, alternatively, if the phenotype is a disorder, whether there are prevalence differences between the sexes. All these features can be left free to differ between the two sexes in structural equation models, providing a more accurate fit to the data. Having established these core differences, the next step is to examine whether there are either *quantitative* and/or *qualitative* sex differences in the genetic and environmental etiology. In twin studies (*see* **Twin Designs**) and **adoption studies**, the simple **ACE model** divides the sources of variance into additive genetic influence (A), common or shared environment (C: environmental influences that make family members similar to one another) and nonshared environment (E: environmental influences that make family members different from one another).

Quantitative Sex Differences

Quantitative sex differences refer to there being different relative proportions of genetic, shared, and nonshared environmental influence on the phenotype for the two sexes. Thus, with twin data, if there is a greater difference in resemblance between monozygotic (MZ) females and dizygotic (DZ) females than there is between the two groups in males, this indicates greater **heritability** for females. In order to test for differences of this kind, two parallel models must be run. First, a model in which A, C, and E are free to differ for males and females is run (heterogeneity or free model). Next, a model is run in which these parameters are fixed to be of the same size (homogeneity or constrained model). As the constrained model is ‘nested’ within the free model, the difference in fit between the two can be examined by looking at the change in chi-square, which is itself distributed as a chi-square with an associated degrees of freedom (difference between the degrees of freedom for the two models) and *P* value. A significant difference in the fit of these two models indicates a significant difference in the relative influence of A, C, and E on males and females for the trait.

In order to illustrate this type of effect, we take as an example some data published on antisocial behavior (ASB) in a sample of Swedish adolescents [2]. As can be seen in Figure 1, the intraclass correlations are given for 5 groups: MZ male, DZ male, MZ female, DZ female, DZ opposite-sex. One clear feature of these data is that while the DZ male and female correlations are similar, the MZ female correlation is significantly higher than the male MZ correlation. This indicates a greater genetic influence on females as compared

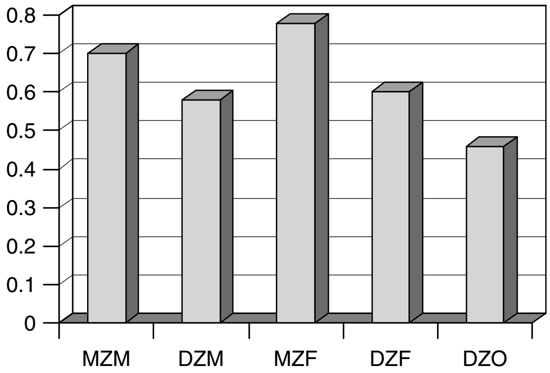


Figure 1 Within-pair correlations for nonaggressive antisocial behavior in young Swedish twins. Adapted from Eley, T.C., Lichtenstein, P. & Stevenson, J. (1999) [2]

Table 1 Quantitative and qualitative sex-limitation models for nonaggressive antisocial behavior in young Swedish Twins

	Males			Females			DZOS					
Model	a^2	c^2	e^2	a^2	c^2	e^2	r_A	r_C	AIC	χ^2	df	p
1[b]	0.47[a]	0.29[a]	0.24[a]	0.47[a]	0.29[a]	0.24[a]	0.5	1.0	4.83	28.83	12	0.01
2[c]	0.26[a]	0.47[a]	0.27[a]	0.60[a]	0.19[a]	0.21[a]	0.5	1.0	−2.82	15.18	9	0.09
3	0.30[a]	0.44[a]	0.26[a]	0.41[a]	0.37[a]	0.22[a]	0.14	1.0	−6.10	9.91	8	0.27
4	0.30[a]	0.44[a]	0.26[a]	0.41[a]	0.37[a]	0.22[a]	0.5	0.69	−6.10	9.91	8	0.27

[a]Denotes a parameter that could not be dropped from the model without significant deterioration in the fit.

[b]Model 1 fits significantly worse than models 2 ($p < .01$), 3 & 4 ($p < .001$).

[c]Model 2 fits significantly worse than models 3 & 4 ($p < .05$).

to males. In the constrained model, A, C, and E were estimated at 47, 29, and 24%, respectively. In contrast, for the sexes-free model, these parameters were 26, 47, and 27%, respectively, for males and 60, 19, and 21% for females. The difference in fit (chi-square) between the two models was 13.65 for 3 degrees of freedom, which is significant at the 1% level (see Table 1). Thus, there is a clear quantitative sex difference in the etiology of nonaggressive antisocial behavior in this sample of Swedish young people. However, there is also another interesting feature of the intraclasss correlations in Figure 1, which is that the DZ opposite-sex correlation is significantly lower than that for the same-sex pairs. This is indicative of a *qualitative* sex difference.

Qualitative Sex Differences

While it is clear that genes and environment could influence the two sexes to a different extent, what is also of interest is that there may be differences in the specific aspects of the genetic or environmental influence for males and females. In other words, there may be different genes that make antisocial behavior heritable in girls from those that influence this phenotype in boys. Qualitative sex differences take this model-fitting approach one step further and allow for the examination of whether the genetic or environmental influences on the phenotype are the *same* in both boys and girls. DZ opposite-sex pairs allow us to examine this issue because any decrease in resemblance between them relative to same-sex DZ pairs is indicative of a lower level of sharing of either genes or environment. This is tested for either by allowing the genetic correlation between members of DZ opposite-sex pairs to be free rather than fixed to 0.5 as is the case for same-sex pairs, or by leaving the shared environment correlation for DZO pairs free to be estimated rather than fixed at 1.0. Unfortunately, because there is only one more unit of information in this model (the DZO covariance), only one of these effects can be tested at any one time.

As noted above, in the data from the young Swedish twins, the DZO correlation was rather lower than that for the male or female DZ pairs (0.46 as compared to 0.58 and 0.60 respectively). The fit for the two models where either the genetic correlation or the shared environment correlation between members of the DZO pairs were left free to vary are given in the final two rows of Table 1. As can be seen, these models have identical fit, which is often the case, as it is difficult to distinguish between these two models. However, both of the qualitative sex difference models fit significantly better than the model with quantitative sex differences only, and the genetic and shared environment correlations are estimated at 0.14 (as compared to 0.5) and 0.69 (as compared to 1.0), respectively. This indicates that there are somewhat different genetic and/or shared environmental influences on nonaggressive antisocial behavior in girls and boys. Furthermore, in this model, there are still significant differences between the relative impact of genes and environment on males and females, with estimates of 30, 44, and 26%, respectively, for genes, shared, and nonshared environment in boys, and 41, 37, and 22%, respectively, in girls.

Sex Limitation and Extreme Scores

The models described above refer to sex differences in the origins of individual differences in the normal range. There are two approaches that can be taken when examining sex differences in more extreme phenotypes. First, with disorders, a threshold model can be undertaken. This can allow for sex differences in the threshold on the estimated liability (above which individuals express the disorder), and also both quantitative and qualitative sex differences on the causes of variation in this underlying liability distribution. This approach can also be used for data that are very skewed, and that are best modeled as a series of categories with thresholds in between each on an underlying liability. However, if the phenotype is a continuous measure (e.g., depression symptoms), in addition to examining sex effects on the distribution of the full range of scores it may be of interest to examine sex effects on extreme scores (i.e., those with high depressive symptoms, thus at high risk for disorder). To examine quantitative sex differences in extreme scores, a regression approach is used [1] (*see* **DeFries–Fulker Analysis**) in which the co-twins' scores are predicted from the probands' scores (those in the extreme group), the degree of genetic relatedness (1.0 for MZ pairs, 0.5 for DZ pairs), sex, and the interactions between sex and both the proband score and the degree of genetic relatedness. The interaction between sex and proband score provides an overall indication of whether there are male–female differences in twin resemblance, and the interaction between sex and genetic relatedness indicates the difference in heritability between the two sexes. As noted in **DeFries–Fulker Analysis**, the core feature of this approach is that following transformation, mean scores for the co-twin groups fall between 0 and 1 and can be interpreted in a similar way to correlations. A model-fitting approach to this method also allows for the testing of qualitative sex differences [5]. The likelihood of both quantitative and qualitative sex differences is indicated by the relative size of MZ versus DZ co-twin means in males versus females, and in the comparison of DZO co-twin means with the same-sex DZ co-twins means. Figure 2 illustrates a pattern of transformed co-twin means indicative of both quantitative and qualitative sex differences. Examination of the data implies a **heritability** for males of around 80%, as compared to 40% for the females. Furthermore, the DZO co-twin mean is much lower than either the male or female DZ mean, indicating a qualitative sex difference in the influence of genes and environment on extreme group membership for this trait.

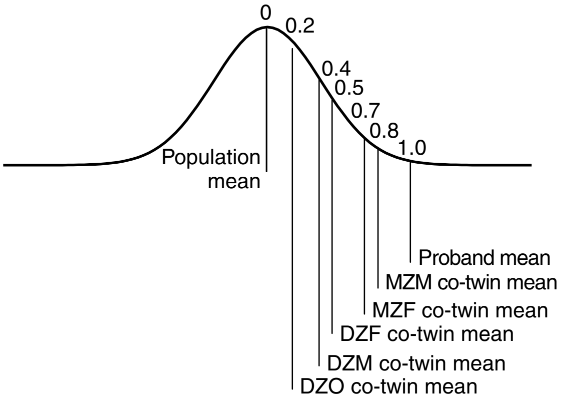


Figure 2 Hypothetical distribution of transformed population, proband, and co-twin means indicating both quantitative and qualitative sex differences

Summary

In addition to the basic sex effects (mean, prevalence, and variance difference), a model-fitting approach can be used to test for two main types of sex difference: quantitative sex differences (level of genetic and environmental influence on males versus females) and qualitative sex differences (different genes impact on males versus females). These models can be tested for variation in the full range of scores, for the liability underlying a disorder, or for the membership of an extreme group defined as being at one or other end of a normally distributed trait. Such findings can be informative particularly with regard to molecular genetics. If there are different genes impacting on a trait in males versus females, then molecular genetic work needs to be undertaken on the two sexes independently. Similarly, for such finding with regard to environmental influence, social researchers need to examine the two sexes separately.

References

[1] DeFries, J.C., Gillis, J.J. & Wadsworth, S.J. (1998). Genes and genders: a twin study of reading disability, in *Dyslexia and Development: Neurobiological Aspects of Extra-Ordinary Brains*, A.M., Galaburda, ed., Harvard University Press, Cambridge, pp. 186–344.

[2] Eley, T.C., Lichtenstein, P. & Stevenson, J. (1999). Sex differences in the aetiology of aggressive and non-aggressive antisocial behavior: results from two twin studies, *Child Development* **70**, 155–168.

[3] Hankin, B.L., Abramson, L.Y., Moffitt, T.E., Silva, A., McGee, R. & Angell, K.E. (1998). Development of depression from preadolescence to young adulthood: emerging gender differences in a 10-year longitudinal study, *Journal of Abnormal Psychology* **107**, 128–140.

[4] Moffitt, T.E., Caspi, A., Rutter, M. & Silva, P.A. (2001). *Sex Differences in Antisocial Behaviour: Conduct Disorder, Delinquency and Violence in the Dunedin Longitudinal Study*, Cambridge University Press, Cambridge.

[5] Purcell, S. & Sham, P.C. (2003). A model-fitting implementation of the DeFries-Fulker model for selected twin data, *Behavior Genetics* **33**, 271–278.

THALIA C. ELEY

Shannon, Claude E

SANDY LOVIE

Volume 4, pp. 1827–1827

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Shannon, Claude E

Born: April 30, 1916, in Michigan, USA.
Died: February 24, 2001, in Massachusetts, USA.

For most psychologists, Shannon is known and admired for only one piece of work, his 1948 paper on information theory and the mathematics of communication systems (see the accessible version in Shannon and Weaver, [5]). Although information theory approaches to cognition flourished in the 1950s and 1960s, culminating in Garner's textbook on information and structure dated 1962 [2], it is now only of historical interest, although one could speculate that the recent interest in complexity theory might once again have psychologist's dusting off their copies of '*A Mathematical Theory of Communication*'. However, at that time, information theory appeared to offer a universal way of defining and assessing information and processing capacity, leading people like George Miller to argue in 1956 for the existence of the 'magical number 7′ as the capacity limit to the human information processing system [3]. None of this would have been possible without the somewhat eccentric and reclusive electrical engineer Shannon.

Shannon's early education was at the University of Michigan, where he obtained BSc degrees in mathematics and electrical engineering in 1936, while he had obtained an MSc and a PhD by 1940 from the Massachusetts Institute of Technology (MIT), with a pioneering thesis on the Boolean analysis of logical switching circuits for the former degree, and one on population genetics for the latter. He had also worked at MIT with Vannevar Bush (who was later to invent *hypertext* as a way of structuring and accessing knowledge) on an early form of computer, the *differential analyzer*. Shannon's next move was to the prestigious Bell Laboratories of AT&T Bell Telephones in New Jersey, where he stayed until 1972, latterly as a consultant. In the interim, he became a visiting member of the faculty at MIT in 1956, and then Donner Professor of Science there from 1958 onward until his retirement two decades later.

Earlier on in 1948 he had published his key paper on the mathematics of information and communication where he defined the information in any message as its predictability, thus turning it into a *probabilistic* measure. He also noted that all signals could be represented digitally to any degree of precision, that is, by binary digits or 'bits', an abbreviation that was also claimed by **John Tukey**. Thus, information could be represented as the sum of the ($\log_2$) of the probabilities of the events in an array of signals. The difference in stimulus and response information also defined the channel capacity of the system, and much work was done by psychologists in the 1950s and 1960s to measure this for many types of stimuli. Shannon also showed that the addition of extra bits in a message that was subject to noise or interference improved the reception of that message, leading to the concept of *redundancy*, an idea exploited by Attneave [1] in perception and by Miller [4] for the recall of simple strings.

Shannon was a somewhat retiring scientist who did most of his work behind closed doors. He also alarmed his colleagues at both Bell and MIT by riding a unicycle from his small collection of such machines down the corridors of these august institutions. It is even rumored that one unicycle was equipped with an asymmetric hub, which attracted crowds to see him progress in an up and down, sinusoidal fashion along the corridor! His other early contributions were to computer encryption, and the creation of unbreakable codes for military use, again drawing on the ideas first set out so eloquently in 1948. Later efforts were concentrated on AI and computing, in particular the formal outline of a practical Turing machine, ideas that had to await the age of the solid state device before seeing their implementation.

References

[1] Attneave, F. (1959). *Applications of Information Theory to Psychology*, Holt, Rinehart & Winston, New York.

[2] Garner, W.R. (1962). *Uncertainty and Structure as Psychological Concepts*, Wiley, New York.

[3] Miller, G.A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information, *Psychological Review* **63**, 81–97.

[4] Miller, G.A. (1958). Free recall of redundant strings of letters, *Journal of Experimental Psychology* **56**, 485–491.

[5] Shannon, C.E. & Weaver, W. (1964). *A Mathematical Theory of Communication*, University of Illinois Press, Urbana.

Sandy Lovie

Shared Environment

DANIELLE M. DICK

Volume 4, pp. 1828–1830

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Shared Environment

Behavior genetic twin studies partition variance in a behavior into genetic and environmental influences. Environmental influences can be further divided into shared (also called common) environment and **nonshared** (or unique) **environment**. When behavior geneticists refer to shared environment, they are referring to environmental influences that make sibling more similar to one another. Examples of shared environmental influences could include growing up in the same home, shared parental rules and upbringing, shared family experiences, sharing the same school and community, and peers that are common to both siblings.

In general, genetically informative studies of behavioral variation have not found a strong role for shared environment. In fact, behavior genetic research consistently demonstrates little or no effect of shared environment for many outcomes, including personality, psychopathology, and adult IQ [4]. Studies of children and adolescents have found more consistent evidence of shared environmental effects, as one might expect when children are still living together at home. Shared environment is widely accepted as an important influence on conduct problems and juvenile delinquency, and IQ in childhood [2, 4, 7, 8, 10]. Additionally, for substance use behaviors, the initiation of substance use appears to show evidence of shared environmental effects, but once an individual begins to use the substance, genetic influences gradually assume a greater role in impacting the behavior. In other words, parents and peers may have a big effect on whether (or when) one starts to drink or smoke, but once they initiate, an individual's own specific genetic predispositions assume greater influence on their subsequent use. This is evident in a longitudinal study of alcohol use among Finnish adolescents, whose frequency of alcohol use was measured at ages 16, 17, and 18.5. At age 16, more than 50% of the variation in frequency of alcohol use was attributed to shared environmental effects, but by age 18.5, less than a third was, as genetic influences had assumed a much greater role [5]. This is a common finding in the literature: shared environmental influences tend to decrease across the life span. As siblings increasingly gain independence from their families, make their own decisions, grow into adults, and start their own careers and families, the impact of shared environment is largely displaced by genetic influences and unique environmental experiences.

So how can it be that there is so little evidence for strong shared environmental influences on behavior? Part of the confusion stems from the fact that shared environment – as behavior geneticists use the term – is not necessarily environmental effects that siblings objectively share, but rather, environmental events that make them more similar. Therefore, it is possible that an environmental event, such as parental divorce, could be objectively shared by both siblings, and yet have different effects on each child. For example, one sibling might react to the divorce by trying to please the parents by being on his/her best behavior in an attempt to reunite the parents. However, another sibling might react to the divorce by acting out, engaging in substance use and delinquent behavior, in a rebellious attempt to 'get back at' the parents. In this case, the environmental event was shared by the siblings, but it had the effect of making them different from one another rather than more similar. In this case, parental divorce would be classified as a 'nonshared environment' because it made the siblings' behavior more diverse. Thus, shared environment is not synonymous with family environment, and a lack of evidence supporting shared environmental effects should not be interpreted as evidence that family influences are not important. Familial influences may be nonshared, in the respect that they serve to make siblings different, rather than similar. Furthermore, nonfamilial influences, such as peer groups, can be shared and serve to make siblings more alike. The distinction between objectively shared environments and the term 'shared environment' as used by behavior geneticists is important when interpreting the literature [9].

Another reason that shared environmental influences might not be widespread in the literature is that **classic twin studies** are not particularly powerful for detecting shared environmental influences. The problem is that the 'a' and 'c' parameters, representing additive genetic, and common (shared) environmental effects, respectively, are highly correlated. **Power** analyses have indicated that in order to detect a common environmental effect of even 50%, more than 100 pairs of twins of each zygosity are needed. As the influence of common environment decreases, the number of pairs needed to detect the effect increases exponentially. Power is further decreased when the outcome is binary, such as affected or unaffected status. **Gene-environment correlation** can also contribute to inflated estimates of genetic influence at the expense of shared environment.

Despite these limitations, behavior genetic studies are advancing to better characterize shared environmental influences. While traditional behavior genetic designs modeled shared environment latently, twin researchers are increasingly *measuring* aspects of the shared environment and incorporating specific environmental measures into genetically informative designs. These models have more power to detect environmental effects, even when these environments only have small effects. In a study by Kendler and colleagues [3], parental loss was included in the classic biometrical twin model and contributed significantly to the variance in major depression – despite the fact that shared environmental influences were not significant when modeled latently. We have studied effects of parental monitoring and home atmosphere on behavior problems in 11- to 12-year-old Finnish twins; both parental monitoring and home atmosphere contributed significantly to the development of children's behavior problems, accounting for 2 to 5% of the total variation, and as much as 15% of the total common environmental effect. Recent research in the United Kingdom found neighborhood deprivation influenced behavior problems, too, accounting for about 5% of the effect of shared environment [1]. Incorporation of specific, measured environments into genetically informative designs offers a powerful technique to study and specify environmental effects.

Another new development in studying the shared environment has been to partition the shared environment into more distinct components. As mentioned previously, shared environmental effects can include everything from parents and peers, to school and community influences. In a unique design used by our research group in studying Finnish twins, a classmate control of the same gender and age was included for each twin. All members of each dyad shared the same neighborhood, school, and classroom, but only the co-twins shared common household experience. In this way, the classic shared environment component could be separated into family environment and school/neighborhood environment. Studying a large sample of 11- to 12-year-old same-sex Finnish twins, sampled from more than 500 classrooms throughout Finland, we found that for some behaviors, including early onset of smoking and drinking, there were significant correlations for both control-twin and control–control dyads. This demonstrates that for these behaviors, the shared environment includes significant contributions from nonfamilial environments – schools, neighborhoods, and communities [6].

In conclusion, shared environment refers to any environmental event that makes siblings more similar to each other. It can include family, peer, school, and neighborhood effects; however, each of these potential environmental effects can also be nonshared if they have the effect of making siblings different from one another. New developments in behavior genetics are making it possible to specify shared environments of importance and to tease apart familial and nonfamilial effects.

References

[1] Caspi, A., Taylor, A., Moffitt, T.E. & Plomin, R. (2000). Neighborhood deprivation affects children's mental health: environmental risks identified in a genetic design, *Psychological Science* **11**, 338–342.

[2] Goldsmith, H.H. & Bihun, J.T. (1997). Conceptualizing genetic influences on early behavioral development, *Acta Paediatrica* **422**, p. 54–59.

[3] Kendler, K.S., Neale, M.C., Kessler, R.C., Heath, A.C. & Eaves, L.J. (1992). Childhood parental loss and adult psychopathology in women: a twin study perspective, *Archives of General Psychiatry* **49**, 109–116.

[4] McGue, M. & Bouchard, T.J. (1998). Genetic and environmental influences on human behavioral differences, *Annual Review of Neuroscience* **21**, p. 1–24.

[5] Rose, R.J., Dick, D.M., Viken, R.J. & Kaprio, J. (2001). Gene-environment interaction in patterns of adolescent drinking: regional residency moderates longitudinal influences on alcohol use, *Alcoholism: Clinical and Experimental Research* **25**, 637–643.

[6] Rose, R.J., Viken, R.J., Dick, D.M., Bates, J., Pulkkinen, L. & Kaprio, J. (2003). It does take a village: nonfamilial environments and children's behavior, *Psychological Science* **14**, 273–277.

[7] Rose, R.J., Dick, D.M., Viken, R.J., Pulkkinen, L., Nurnberger Jr, J.I. & Kaprio, J. (2004). Genetic and environmental effects on conduct disorder, alcohol dependence symptoms, and their covariation at age 14, *Alcoholism: Clinical and Experimental Research* **28**, 1541–1548.

[8] Rutter, M., Silberg, J., O'Connor, T. & Simonoff, E. (1999). Genetics and child psychiatry: II. Empirical research findings, *Journal of Child Psychology and Psychiatry* **40**, 19–55.

[9] Turkheimer, E. & Waldron, M. (2000). Nonshared environment: a theoretical, methodological, and quantitative review, *Psychological Bulletin* **126**, 78–108.

[10] Waldman, I.D., DeFries, J.C. & Fulker, D.W. (1992). Quantitative genetic analysis of IQ development in young children: multivariate multiple regression with orthogonal polynomials, *Behavior Genetics* **22**, 229–238.

DANIELLE M. DICK

Shepard Diagram

Jan de Leeuw

Volume 4, pp. 1830–1830

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Shepard Diagram

In a general nonmetric scaling situation (*see* **Multidimensional Scaling**), using the Shepard–Kruskal approach, we have data $y_i, \ldots, y_n$ and a model $f_i(\theta)$ with a number of free parameters θ. Often this is a nonmetric multidimensional scaling model, in which the model values are distances, but linear models and inner product models can be and have been treated in the same way. We want to choose the parameters in such a way that the rank order of the model approximates the rank order of the data as well as possible.

In order to do this, we construct a loss function of the form

$$\sigma(\theta, \hat{y}) = \sum_{i=1}^{n} w_i(\hat{y}_i - f_i(\theta))^2,$$

where the w_i are known weights. We then minimize σ over all $\hat{y}$ that are monotone with the data y and over the parameters θ (*see* **Monotonic Regression**).

After we have found the minimum, we can make a **scatterplot** with the data y on the horizontal axis and the model values f on the vertical axis. This is what we would also do in linear or nonlinear regression analysis. In nonmetric scaling, however, we also have the $\hat{y}$, which are computed by **monotone regression**. We can add the $\hat{y}$ to vertical axis and use them to draw the best-fitting monotone step function through the scatterplot. This shows the **optimal scaling** of the data, in this case the monotone transformation of the data, which best fits the fitted model values. The scatterplot with y and f, and $\hat{y}$ drawn in, is called the *S*hepard diagram. In Figure 1, we show an example from a nonmetric analysis of the classical Rothkopf Morse code confusion data [2]. The stimuli are 36 Morse code signals. The raw data are the proportions p_{ij}, which signals i and j were judged to be the same by over 500 subjects. Dissimilarities were computed using the transformation

$$\delta_{ij} = -\frac{1}{2}\log\frac{p_{ij}p_{ji}}{p_{ii}p_{jj}},$$

which is suggested by both Shepard's theory of stimulus generalization and by Luce's choice model for discrimination (see [1] for details). A nonmetric scaling analysis in two dimensions of these dissimilarities gives the Shepard diagram in Figure 1.

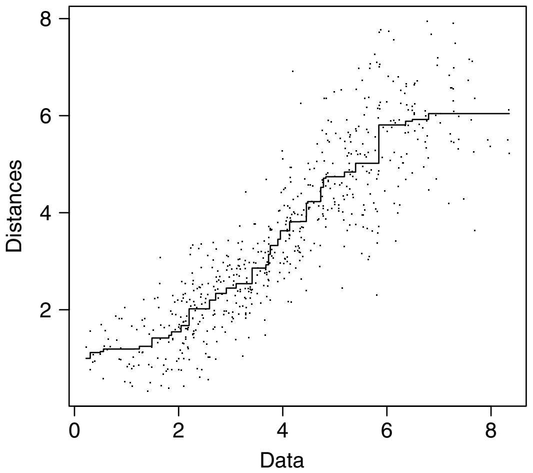


Figure 1 Shepard diagram Morse code data

References

[1] Luce, R.D. (1963). Detection and recognition, in *Handbook of Mathematical Psychology*, Vol. 1, Chap. 3, R.D. Luce, R.R. Bush & E. Galanter, eds, Wiley, pp. 103–189.

[2] Rothkopf, E.Z. (1957). A measure of stimulus similarity and errors in some paired associate learning, *Journal of Experimental Psychology* **53**, 94–101.

JAN DE LEEUW

Sibling Interaction Effects

Dorret I. Boomsma

Volume 4, pp. 1831–1832

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Sibling Interaction Effects

The effects of social interaction (*see* **Social Interaction Models**) among siblings to individual differences in behavior were first discussed by Eaves [3] and later by Carey [1] and others. In the context of behavior genetic research, social interaction effects reflect that alleles may cause variation in one or more traits of individuals carrying these alleles, but may also, through social interaction, influence the phenotypes of individuals who do not carry them [4]. Social interactions between siblings thus create an additional source of variance and generate genotype-environment covariance if the genes causing the social interaction overlap with the genes that influence the phenotype under study.

Social interaction effects between siblings can either be cooperative (imitation) or competitive (contrast), depending on whether the presence in the family of, for example, a high-scoring sibling inhibits or facilitates the behavior of the other siblings. Cooperation implies that behavior in one sibling leads to similar behavior in the other siblings. In the case of competition, the behavior in one child leads to the opposite behavior in the other child.

In the classical **twin design**, cooperation or positive interaction leads to increased twin correlations for both monozygotic (MZ) and dizygotic (DZ) twins. The relative increase is larger for DZ than for MZ correlations, and the pattern of correlations thus resembles the pattern that is seen if a trait is influenced by common environmental factors. Positive interactions have been observed for traits such as antisocial tendencies [2]. Negative sibling interaction, or competition, will result in MZ correlations, which are more than twice as high as DZ correlations, a pattern also seen in the presence of genetic **dominance**. Such a pattern of correlations has been reported in genetic studies of Attention Deficit Hyperactivity Disorder (ADHD) and related phenotypes in children (e.g., [6]). In adults, a pattern of high MZ and low DZ correlations has been observed for anger [7].

In data obtained from parental ratings of the behavior of their children, the effects of cooperation and competition may be mimicked (e.g., [8]). When parents are asked to evaluate and report upon their children's phenotype, they may compare the behavior. In this way, the behavior of one child becomes the standard against which the behavior of the other child is rated. Parents may stress either similarities or differences between children, resulting in an apparent cooperation or competition effect.

The presence of an interaction effect, either true sibling interaction or rater bias, is indicated by differences in MZ and DZ variances. If the interaction effect is cooperative the variances of MZ and DZ twins are both inflated, and this effect is greatest on the MZ variance. The opposite is observed if the effect is competitive; MZ and DZ variances are both deflated and again this effect is greatest on the MZ variance. In addition to heterogeneity in MZ and DZ variances, the presence of interaction affects MZ and DZ correlations. Under competition MZ correlations are much larger than DZ correlations. Under cooperation MZ correlations are less than twice the DZ correlation. These patterns of correlations are not only consistent with contrast and cooperation effects, but also with genetic nonadditivity (e.g., genetic dominance) and common environmental influences, respectively. In order to distinguish between these alternatives, it thus is crucial to consider MZ and DZ variance-covariance structures in addition to MZ and DZ correlations.

Rietveld et al. [5] carried out a simulation study to determine the statistical **power** to distinguish between the two alternatives of genetic dominance and social interaction. The results showed that when both genetic dominance and contrast effects are present, genetic dominance is more likely to go undetected in the classical twin design than the interaction effect. Failure to detect the presence of genetic dominance leads to slightly biased estimates of additive genetic, unique environmental, and interaction effects (*see* **ACE Model**). Competition is more easily detected in the absence of genetic dominance. If the significance of the interaction effect is evaluated while also modeling genetic dominance, small contrast effects are likely to go undetected, resulting in a relatively large bias in estimates of the other parameters. Alternative designs, such as including pairs of unrelated siblings reared together to the classical twin study, or including the nontwin siblings of twins, increases the statistical power to detect contrast effects as well as the power to distinguish between genetic dominance and contrast effects.

Sibling interaction will go undetected in the classical twin design if the genes responsible for the interaction differ from the genes which influence the trait. In such cases, a comparison with data from singletons may permit further investigation. In parental ratings, the question whether an interaction effect represents true sibling interaction or rater bias also must be resolved through the collection of additional data, for example, from teachers.

References

[1] Carey, G. (1986). Sibling imitation and contrast effects, *Behavior Genetics* **16**, 319–341.

[2] Carey, G. (1992). Twin imitation for antisocial behavior: implications for genetic and family environment research, *Journal of Abnormal Psychology* **101**, 18–25.

[3] Eaves, L.J. (1976). A model for sibling effects in man, *Heredity* **36**, 205–214.

[4] Eaves, L.J., Last, K.A., Young, P.A. & Martin, N.G. (1978). Model-fitting approaches to the analysis of human behavior, *Heredity* **41**, 249–320.

[5] Rietveld, M.J.H., Posthuma, D., Dolan, C.V. & Boomsma, D.I. (2003). ADHD: sibling interaction or dominance: an evaluation of statistical power, *Behavior Genetics* **33**, 247–255.

[6] Rietveld, M.J.H., Hudziak, J.J., Bartels, M., van Beijsterveldt, C.E.M. & Boomsma, D.I. (2004). Heritability of attention problems in children. Longitudinal results from a study of twins age 3 to 12, *Journal of Child Psychology and Psychiatry* **45**, 577–588.

[7] Sims, J., Boomsma, D.I., Carrol, D., Hewitt, J.K. & Turner, J.R. (1991). Genetics of type A behavior in two European countries: evidence for sibling interaction, *Behavior Genetics* **21**, 513–528.

[8] Simonoff, E., Pickles, A., Hervas, A., Silberg, J.L., Rutter, M. & Eaves, L. (1998). Genetic influences on childhood hyperactivity: contrast effects imply parental rating bias, not sibling interaction, *Psychological Medicine* **28**, 825–837.

Dorret I. Boomsma

Sign Test

Shlomo Sawilowsky

Volume 4, pp. 1832–1833

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Sign Test

The logic of the nonparametric Sign test is 'almost certainly the oldest of all formal statistical tests as there is published evidence of its use long ago by **J. Arbuthnott** (1710)!' [6, p. 65]. The modern version of this procedure is generally credited to **Sir Ronald A. Fisher** in 1925. Early theoretical development with applications appeared in [1].

The Sign test is often used to test a population median hypothesis, or with matched data as a test for the equality of two population medians. It is based upon the number of values above or below the hypothesized median. Gibbons [4] positioned it as a counterpart of the parametric one-sample *t* Test. However, Hollander and Wolfe [5] stated 'generally, (but not always) the sign test is less efficient than the **signed rank test**'.

Assumptions

Data may be discrete or continuous variables. It is assumed that the data are symmetric about the median, with no values equal to the hypothesized population median. There are a variety of procedures to handle values located at the median, including (a) deleting them from the analysis and (b) alternating the assignment for or against the null hypothesis.

Hypotheses

The null hypothesis being tested is $H_0 : M = M_0$, where M is the population median and M_0 is a hypothesized value for that parameter. The nondirectional alternative hypothesis is $H_1 : M \neq M_0$. Directional alternative hypotheses are of the form $H_1 : M < M_0$ and $H_1 : M > M_0$.

Procedure

Each x_i is compared with M_0. If $x_i > M_0$, then a plus sign '+' is recorded. If $x_i < M_0$, then a minus sign '−' is recorded. In this way, all data are reduced to '+' and '−' signs. If the alternative hypothesis is $H_1 : M > M_0$, the logic of the test would indicate that there will be more plus signs than minus signs. If there are values equal to the median, count half of them as plus and half as minus.

Test Statistic

The test statistic is the number of '+' signs or the number of '−' signs. If the expectation under the alternative hypothesis is that there will be a preponderance of '+' signs, the test statistic is the number of '−' signs. Similarly, if the expectation is a preponderance of '−' signs, the test statistic is the number of '+' signs. If the test is two-tailed, use the smaller of the two counts. Thus,

$$S = \text{the number of ` + ' or ` − ' signs}$$
$$\text{(depending upon the context).}$$

Large Sample Size

The large sample approximation is given by

$$S^* = \frac{S - (n/2)}{\sqrt{n/4}},$$

where S is the test statistic and n is the sample size. S^* is compared to the standard normal z scores for the appropriate α level. Monte Carlo simulations conducted by Fahoome and Sawilowsky [3] and Fahoome [2] indicated that n should not be less than 50.

Example

Consider the following sample data ($n = 15$). The null hypothesis is $H_0 : M = 18$, which is to be tested against the alternative hypothesis $H_0 : M \neq 18$.

20	33	4	34	13	6	29	17	39	26
+	+	−	+	−	−	+	−	+	+

13	9	33	16	36
−	−	+	−	+

Each of the scores is assigned a plus or a minus, depending on whether the score is above or below the median. The number of minuses is 7 and the number of pluses is 8. Thus, choose $S = 7$. Using a table of critical values, S is not significant and the null hypothesis cannot be rejected on the basis of the evidence provided by the sample.

References

[1] Dixon, W.J. & Mood, A.M. (1946). The statistical sign test, *Journal of the American Statistical Association* **41**, 557–566.

[2] Fahoome, G. (2002). Twenty nonparametric statistics and their large-sample approximations, *Journal of Modern Applied Statistical Methods* **1**(2), 248–268.

[3] Fahoome, G. & Sawilowsky, S. (2000). Twenty nonparametric statistics, in *Annual Meeting of the American Educational Research Association, SIG/Educational Statisticians*, New Orleans.

[4] Gibbons, J.D. (1971). *Nonparametric Methods for Quantitative Analysis*, Holt, Rinehart, & Winston, New York.

[5] Hollander, M. & Wolfe, D. (1973). *Nonparametric Statistical Methods*, John Wiley & Sons, New York.

[6] Neave, H.R. & Worthington, P.L. (1988). *Distribution-Free Tests*, Unwin Hyman, London.

(*See also* **Distribution-free Inference, an Overview**)

SHLOMO SAWILOWSKY

Signal Detection Theory

John T. Wixted

Volume 4, pp. 1833–1836

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Signal Detection Theory

Signal-detection theory is one of psychology's most notable achievements, but it is not a theory about typical psychological phenomena such as memory, attention, vision or psychopathology (even though it applies to all of those areas and more). Instead, it is a theory about how we use evidence to make decisions. The evidence could be almost anything (e.g., the intensity of an auditory perception, the strength of a retrieved memory, or the number of symptoms suggestive of schizophrenia), and the task of the decision-maker is to decide whether or not enough evidence exists to declare the presence of the condition in question (e.g., the presence of a tone or a remembered item or a disease).

What makes decisions like these difficult is that we sometimes think we hear sounds that did not, in fact, occur. Similarly, faces sometimes seem familiar even upon first encounter, and symptoms of schizophrenia can be exhibited even by people who do not suffer from the disease. That being the case, we cannot make a positive decision merely because there is the slightest evidence pointing in the positive direction. Instead, there must be *enough* evidence, and signal-detection theory is all about understanding the process of deciding that there is, indeed, sufficient evidence to make a positive decision.

Signal-detection theory was initially introduced to the field of psychology in the area of psychophysics (Green & Swets, 1966), where the prototypical task was to decide whether a tone was presented on a particular trial or not. Since then, the basic detection framework has been much more broadly applied, such that it is now the major decision theory in tasks as diverse as recognition memory and diagnostic radiology. To illustrate the basic principles of signal-detection theory, a standard recognition memory task will be considered, though many other domains of application would do just as well. In a typical recognition task, participants are presented with a list of stimuli (e.g., a series of faces) followed by a recognition test involving items that appeared on the list (the targets) randomly intermixed with items that did not appear on the list (the lures, which are also known as distractors). The recognition test is the signal-detection task in this example. In the simplest case, the targets and lures are presented one at a time for a yes/no recognition decision ('yes' means that the item is judged to have appeared on the list), and there is an equal number of each. The proportion of targets that receive a correct 'yes' response is the hit rate (HR), and the proportion of lures that receive an *incorrect* 'yes' response is the false alarm rate (FAR).

Although a test item falls into one category or the other (i.e., the item is either a target or a lure), a signal-detection analysis of the recognition task begins with the assumption that the test items vary *continuously* along a psychologically meaningful dimension. That dimension need not be named, but leaving it unspecified often seems somehow wrong to those who are new to the theory. Thus, for the sake of illustration only, we might name that dimension 'familiarity' (at least for the specific case of recognition memory). When presented with a test item on a recognition test, the participant will experience some sense of familiarity, and this holds true even for the lures. The familiarity of the lures might be low, on average, but some sense of familiarity will occur for each one (perhaps because the face somewhat resembles a face that appeared on the list or because it resembles an acquaintance, etc.). Moreover, the lures will not all generate the exact same low level of familiarity. Instead, they will generate a range of familiarity values (i.e., a distribution of values). Some of that variability arises because items presumably differ in inherent familiarity (e.g., an average-looking face might seem more familiar than a strange-looking face). Variability can also arise from internal sources. That is, even two faces with the same inherent levels of familiarity may generate different internal responses due to moment-to-moment processing differences. Thus, variability in subjective evidence always exists, even in the simplest auditory detection experiment in which the same physical stimulus is presented on many trials.

The left distribution in Figure 1 represents the hypothetical distribution of familiarity values associated with the lures on a recognition test. The mean of that distribution occurs at a relatively low point on the familiarity scale (labeled μ_{Lure}, although its true value is unknown), and the distribution itself simply reflects the fact that some of the lures have a higher familiarity value than the mean whereas others have a lower familiarity value than the mean. This distribution is analogous to the 'noise' distribution in an

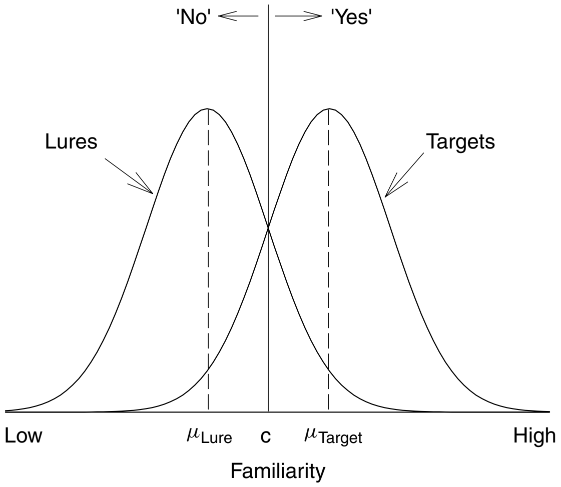


Figure 1 Prototypical signal-detection model of a Yes/No recognition memory test

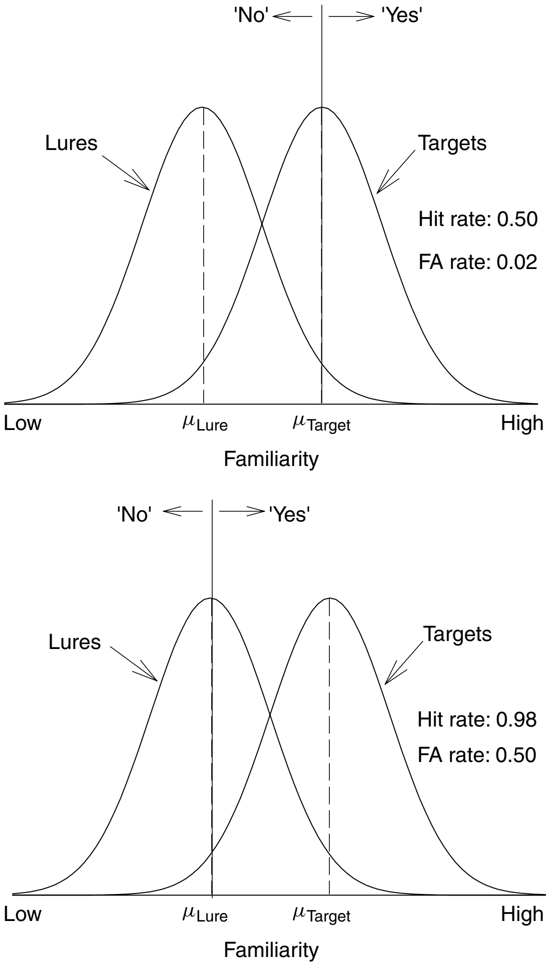


Figure 2 An illustration of conservative (upper panel) and liberal (lower panel) placements of the decision criterion

auditory perception experiment (i.e., trials in which a to-be-detected tone is not presented).

The targets that are presented during the recognition test also have some mean familiarity value (labeled μ_{Target}), but that mean value is relatively high because the faces were recently seen on a list. Once again, though, all of the faces do not have the same high familiarity value. Instead, there is a distribution of familiarity values about the mean. In the hypothetical example shown in Figure 1, the distribution of familiarity values associated with the targets is also assumed to be Gaussian (i.e., normal), and it is assumed to have the same standard deviation as the lure distribution. This distribution is analogous to the 'signal' (i.e., white noise plus tone) distribution in an auditory detection experiment. Note that the equal-variance assumption is not required, and it is often not true in practice, but Figure 1 depicts the simplest version of detection theory.

The crux of the decision problem is this: how familiar must a test item be before it is declared to have appeared on the list? It is somewhat frustrating to realize that there is no obvious answer to this question, and it is up to the participant to pick a criterion familiarity value above which items receive a 'yes' response. In the hypothetical example illustrated in Figure 1, the participant has placed the decision criterion at the point labeled 'c' on the familiarity scale, which happens to be halfway between the means of the target and lure distributions. In a case like that, the HR would be 0.84 (i.e., 84% of the targets yield a familiarity greater than c) and the FAR would be 0.16 (i.e., 16% of the lures yield a familiarity greater than c), where the HR is the proportion of targets that receives a correct 'yes' response and the FAR is the proportion of lures that receives an incorrect 'yes' response. But a more conservative participant, illustrated in the upper panel of Figure 2, might place the criterion at a higher point on the familiarity scale, in which case both the HR and the FAR would both be lower. This conservative participant requires that a test item generate a familiarity value greater than μ_{Target} before saying 'yes'. Only 50% of targets and approximately 2% of the lures exceed that familiarity value, so the HR for this participant would be 0.50 and the FAR would be 0.02. A liberal participant, illustrated in the lower panel of Figure 2, might instead only require that a test item generate a familiarity value greater than μ_{Lure} before saying 'yes,' in which case the HR would be 0.98 and the FAR would be 0.50. In practice, participants exhibit variability just like this (i.e., some have high hit and FARs whereas others have low hit and FARs).

What is the appropriate measure of memory performance for these participants? A natural choice would be to use the percentage of correct responses, but the problem with that choice is that it does not remain constant as bias changes. The neutral, conservative and liberal observers in the example above all have the same memory (i.e., the distributions are the same distance apart for each) – they differ only in their willingness to say 'yes'. That is, they differ in *bias*, not in their ability to discriminate targets from lures. If half the test items are targets and half are lures, then percent correct is equal to the average of the HR and the correct rejection rate, where the correct rejection rate is equal to 1 minus the FAR. For the neutral, conservative and liberal observers, percent correct is equal to 84%, 74% and 74%, respectively. The value should remain constant because memory remains constant, and the fact that it does not reveals a flaw with that measure.

A better approach would be to use the distance between the means of the target and lure distributions as a measure of discriminability and to use the location of the decision criterion as a measure of bias. An intuitively appealing measure like the percentage of correct responses conflates these two separable properties of discriminative performance. The measure of discriminability derived from detection theory is d', which is the distance between the means of the target and lure distributions *in standard deviation units* (not in units of familiarity). In the example shown in Figure 1, d' is equal to 2.0. That is, the means are 2 standard deviations apart. Had the items on the list been given less study time, d' would be less than 2.0 (down to a minimum of 0, at which point chance responding would prevail). Had the items been given more study time, d' would be greater than 2.0 (up to a practical maximum of about 4, at which point very few mistakes would be made).

The formula for computing d' is z(*HR*) - z(*FAR*), where z(p) is the z-score associated with the cumulative normal probability of p. Thus, for the neutral participant, $d' = \text{z}(0.84) - \text{z}(0.16)$, which is approximately equal to 1 −1, or 2. For the conservative participant, $d' = \text{z}(0.50) - \text{z}(0.02)$, which is also approximately equal to 2. And for the liberal participant, $d' = \text{z}(0.98) - \text{z}(0.50)$, which, again, is approximately 2.0. Thus, detection theory provides a means of computing discriminability *independent of bias*. In this hypothetical example, the hit and FARs vary across observers, but the ability to discriminate a target from a lure (i.e., the distance between the means of the target and lure distributions) remains constant.

A plot of the HR vs. the FAR over different levels of bias is called the *Receiver Operating Characteristic* (ROC; *see* **Receiver Operating Characteristics Curves**). The typical ROC traces out a curvilinear path (some sense of this can be obtained by plotting the 3 pairs of hit and FARs discussed above), and the entire path represents a single value of d' over a continuous range of bias. Note that d' is an excellent choice of dependent measure when the equal-variance model applies. When an unequal-variance model is applicable, other detection-related measures are more appropriate [1, 2].

Detection theory also provides various ways to specifically quantify the degree of bias. One common measure is C, which is the distance from the point of intersection between the two distributions (i.e., from the midpoint) to the location of the criterion (again, in standard deviation units). The computational formula for C is: $-0.5^*[\text{z}(HR) + \text{z}(FAR)]/d'$. This value will be zero for the unbiased case (i.e., criterion midway between the distributions), but it will be positive for more conservative (i.e., higher) placements of the criterion and negative for more liberal (i.e., lower) placements of the criterion. For the neutral, conservative and liberal responders considered above, the d' is 2.0 for all three and the corresponding C values are 0, 0.5, and −0.5. Thus, for any pair of hit and FARs, one can compute a bias-free discriminability measure (d') and a measure of the participant's degree of bias (C).

It is important to emphasize that d' and C are measured *in standard deviation units*. We do not really know anything about the underlying distributions (i.e., we do not know their means or their standard deviations), but we can compute d' and C

from the obtained hit and FARs nonetheless. It seems like magic until you realize, for example, that if a participant has a HR of 0.84 and a FAR of 0.16, there is no way to draw the equal-variance Gaussian signal-detection depiction of those values except as drawn in Figure 1. The distributions must be placed two standard deviations apart, and the criterion must be placed halfway between. No other arrangement would correspond to a HR of 0.84 and a FAR of 0.16 (and this remains true even if we do not know what to name the decision axis, which has been named '*familiarity*' in Figure 1 for the sake of illustration only).

The great value of signal-detection theory lies not only in its ability to separate discriminability and bias measures (which a measure like percent correct cannot do) but also in its ability to conceptualize the underlying decision processes associated with an extremely wide range of endeavors. The conceptual utility of graphical depictions such as those shown in Figures 1 and 2 is hard to overstate whether the topic in question is perception, memory, or psychiatric diagnosis (to name just a few areas of application).

References

[1] Green, D.M. & Swets, J.A. (1966). *Signal Detection Theory and Psychophysics*, Wiley, New York.

[2] Macmillan, N.A. & Creelman, C.D. (1991). *Direction Theory: A User's Guide*, Cambridge University Press, Cambridge.

JOHN T. WIXTED

Signed Ranks Test

SHLOMO SAWILOWSKY AND GAIL FAHOOME

Volume 4, pp. 1837–1838

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Signed Ranks Test

The nonparametric [4] Signed Ranks procedure provides a test of a hypothesis about the magnitude of the location parameter for a sample. It can be employed with a single sample to test a hypothesis about the median of the population sampled or with the differences between paired samples to test a hypothesis about the median of the population of such differences.

Although the nonparametric Wilcoxon Rank Sum test (*see* **Wilcoxon–Mann–Whitney Test**) is considerably more powerful than the parametric independent samples t Test under departures from population normality, the Wilcoxon Signed Ranks test presents more modest power advantages over the paired samples t Test.

Assumptions

It is assumed that the paired differences are independent, and originate from a symmetric, continuous distribution.

Hypotheses

The null hypothesis is $H_0 : \theta = \theta_0$, which is tested either against the nondirectional alternative $H_1 : \theta \neq \theta_0$ or against one of the directional alternatives $H_1 : \theta < \theta_0$ or $H_1 : \theta > \theta_0$.

Procedure

In the one-sample case, compute the differences, D_i, by the formula

$$D_i = x_i - \theta_0 \tag{1}$$

and in the paired samples case

$$D_i = (x_{i2} - x_{i1}) - \theta_0, \tag{2}$$

where, for example, x_{i1} could be a 'before' or 'pretest' score and x_{i2} an 'after' or 'posttest' score.

Next, assign ranks, R_i, to the absolute values of these differences in ascending order, keeping track of the individual signs.

Test Statistic

The test statistic is the sum of either the positive ranks or the negative ranks. If the alternative hypothesis suggests that the sum of the positive ranks should be larger, then

$$T^- = \text{the sum of ranks of the negative differences.} \tag{3}$$

If the alternative hypothesis suggests that the sum of the negative ranks should be larger, then

$$T^+ = \text{the sum of ranks of the positive differences.} \tag{4}$$

For a two-tailed test, T is the smaller of the two rank sums. The total sum of the ranks is

$$\frac{N(N+1)}{2},$$

which gives the following relationship

$$T^+ = \frac{N(N+1)}{2} - T^-. \tag{5}$$

Large Sample Sizes

The large sample approximation is

$$z = \frac{T - \dfrac{N(N+1)}{4}}{\sqrt{\dfrac{N(N+1)(2N+1)}{24}}}, \tag{6}$$

where T is the test statistic. The resulting z is compared to the standard normal z for the appropriate alpha level. Monte Carlo simulations conducted by [3] and [2] indicated that the large sample approximation requires a minimum sample size of 10 for $\alpha = 0.05$ and 22 for $\alpha = 0.01$. Among others, [1] provides tables to be used with smaller sample sizes.

Example

A two-sided Wilcoxon Signed Rank test, with $\alpha = 0.05$, is computed on the following data set.

Test Score	87	90	88	88	89	91	89
Retest Score	90	85	94	97	96	90	99
D_i	3	−5	6	9	7	−1	10
R_i	2	3	4	6	5	1	7
Negative Ranks		3				1	

The obtained value, T, is the sum of the two negative ranks: $3 + 1 = 4$. The critical value for $N = 7$ is 3. Because $4 > 3$, the null hypothesis is rejected in favor of the alternative hypothesis that the median change in scores, from test to retest, differs from zero.

References

[1] Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Edition, Wiley, New York.

[2] Fahoome, G. (2002). Twenty nonparametric statistics and their large-sample approximations, *Journal of Modern Applied Statistical Methods* **1**(2), 248–268.

[3] Fahoome, G., & Sawilowsky, S. (2000). Twenty nonparametric statistics, in *Annual Meeting of the American Educational Research Association, SIG/Educational Statisticians*, New Orleans.

[4] Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics* **1**, 80–83.

(*See also* **Distribution-free Inference, an Overview**)

Shlomo Sawilowsky and Gail Fahoome

Simple Random Assignment

Chau Thach and Vance W. Berger

Volume 4, pp. 1838–1840

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Simple Random Assignment

There are two types of simple random assignment – unrestricted random assignment and the random allocation rule (*see* **Randomization**). Unrestricted random assignment occurs when each subject in a randomized study is assigned to one of K possible treatments with a fixed probability, such as $1/K$, independent of all previous assignments in the study. Neither the total sample size of the study nor the sample size for each treatment group needs to be known in advance to use this procedure. For a study with two treatments and equal allocation to each treatment, unrestricted random assignment is equivalent to using a simple toss of a fair coin to allocate each subject to a treatment.

In practice, allocation is usually conducted not with coins but rather using random numbers generated from a computer. In contrast to unrestricted random assignment, the random allocation rule does require the total sample size of the study and the sample sizes in each treatment group to be known in advance. A randomly chosen subset of the subjects in the study is assigned to one treatment group. If there are only two treatment groups, then the unselected subjects comprise the other treatment group. If there are more than two treatment groups, then another randomly chosen subset is assigned to another treatment group, and so on, with the remaining subjects assigned to the last treatment group. For example, if four subjects were to be randomized to two groups, with two subjects to be randomized to each group, then there would be six possible pairs of subjects to make up the first group: $\{1, 2\}$, $\{1, 3\}$, $\{1, 4\}$, $\{2, 3\}$, $\{2, 4\}$, and $\{3, 4\}$. By the random allocation rule, each of these six groups is selected with the same probability, specifically 1/6. In general, if there are two treatment groups and equal allocation to each, then the random allocation rule has $(2\mathrm{n})!/(\mathrm{n}!)^2$ possible allocation sequences and unrestricted randomization has 2^{2n} possible allocation sequences, where there are $2n$ subjects in total, and n are assigned to each group for the random allocation rule. With unrestricted randomization, n is the expected size of each treatment group.

For both types of simple random assignment, the marginal (unconditional) probability of assignment to a treatment is the same for each subject. For two treatments with equal allocation, each subject has the same probability, $1/2$, of being assigned to a treatment, over the set of all possible permutations, with both types of simple random assignment. With unrestricted random assignment, each subject still has the same constant conditional probability of assignment to each treatment, even given all the previous assignments. However, with the random allocation rule, the conditional probability of assignment to a treatment is not constant for each subject. For example, if four subjects were to be randomized to two groups using the random allocation rule and the first two subjects have both been randomized to the first treatment group, then the next two subjects would have no chance of being in the first treatment group.

One benefit of unrestricted random assignment is that it is the only allocation method to completely eliminate the type of selection bias that results from the random allocation process being predictable, thereby allowing an investigator to enter a specific subject into the study based on the concordance between this subject's characteristics and the next treatment to be assigned [2]. Note that many randomized studies are unmasked so that the prior assignments are known. Even those that are masked in theory may not be fully masked, so that at least some of the prior assignments are known. In such a case, any patterns created by restrictions on the random allocation allow for prediction of the future assignments [1]. It is in this context that unrestricted random assignment can be appreciated for its lack of any restrictions on the random allocation, and hence its resistance to selection bias [3].

For the random allocation rule, there is a small chance for selection bias in an unmasked study. Certainly, the last allocation is predictable, and depending on the allocation sequence, more allocations may also be predictable. The maximum number of predictable allocations is the size of the largest treatment group. For example, if four subjects were to be randomized to two groups and the first two subjects have been randomized to the first treatment group, then the investigator would know that the next two subjects would be in the second treatment group. The third and fourth allocations would be predictable. But if the allocation sequence were instead ABAB, then only the fourth allocation would be predictable [3]. The opportunity for selection bias can be eliminated if all the subjects are allocated simultaneously rather than sequentially as they enter the study [1].

Although complete random assignment is perhaps the easiest allocation method to understand and implement, it is not widely used in practice. This is due mainly to the possibility of treatment imbalance – imbalances in the number of subjects assigned to each treatment – at any stage in the study. The imbalance in the numbers allocated to each group can be substantial if the sample size is small, and even if the sample size is large, it is still possible that there would be a gross imbalance during the early stages of the study. Imbalances in the size of the treatment groups can reduce the **power** to detect any true differences between the two treatment groups. The reduction in power is small, except for extreme imbalances. For example, if a study has 90% power with equal balance between treatments (1 : 1 assignment), then the power is reduced to about 85% if the imbalance in treatments is as extreme as 7 : 3 or greater [5].

If 20 subjects are being randomized to two treatments with 1 : 1 assignment (10 to each group on average, but this perfect balance would not be forced), then the probability of obtaining a 3 : 2 imbalance (12 subjects being assigned to one treatment and eight subjects being assigned to the other treatment) or worse is about 50%. With 100 subjects, this probability reduces to 5% [4]. As the sample size increases even further, this probability approaches 0%. These imbalances can reduce the power of the study to detect any true difference, but this concern can be addressed by using random allocation. However, both types of simple random assignment can be subject to accidental and chronological bias [3]. Accidental bias occurs when the subjects enrolled at one point in time differ systematically from those subjects enrolled at a later time but in the same study.

Consider the myriad number of ways that ambient conditions can change during the course of a study – study personnel may leave and be replaced, economic factors may change, new legislation may take effect during the study, flu season or allergy season may occur during one part of the study but not during another part of the study, and so on. These systematic differences that occur over time may not be measurable, or the variables that do measure them may not be recorded. This becomes a concern if a disproportionate number of subjects at one time are enrolled to one group, and a disproportionate number of subjects at another time are enrolled to the other group. This is chronological bias [3], and more restrictions on the random allocation than simply terminal balance tend to be needed to minimize or eliminate it. We see, then, that there is a trade-off, in that more restrictions are needed to control chronological bias, yet fewer restrictions are required to control selection bias that arises from prediction of future assignments. It is generally impossible to simultaneously eliminate both [3], although a few exceptions to this rule of thumb exist [1]. In most cases, there will be covariate imbalances across the treatment groups, whether simple random assignment is used or not. However, complete random assignment has the smallest chance for accidental bias, while random allocation has a slightly larger chance [6].

Treatment imbalances do not invalidate the statistical tests that are generally performed to compare the treatment groups as long as they are random. Chronological bias, while called a bias, is actually a random error that is just as likely to favor one treatment group as it is to favor another. If it is the cause of a covariate imbalance, then standard inference is still unbiased. However, selection bias is a true bias, and so if it is the cause of a covariate imbalance, then standard comparisons are not unbiased, and standard tests are not valid. This, and the fact that unrestricted random assignment eliminates selection bias, is one reason to consider using it.

References

[1] Berger, V.W. & Christophi, C.A. (2003). Randomization technique, allocation concealment, masking, and susceptibility of trials to selection bias, *Journal of Modern Applied Statistical Methods* **2**(1), 80–86.

[2] Berger, V.W. & Exner, D.V. (1999). Detecting selection bias in randomized clinical trials, *Controlled Clinical Trials* **20**, 319–327.

[3] Berger, V.W., Ivanova, A. & Deloria-Knoll, M. (2003). Enhancing allocation concealment through less restrictive randomization procedures, *Statistics in Medicine* **22**(19), 3017–3028.

[4] Friedman, L.M., Furgerg, C.D. & DeMets, D.L. (1998). *Fundamentals of Clinical Trials*, Springer, New York.

[5] Lachin, J.M. (1988). Statistical properties of randomization in clinical trials, *Controlled Clinical Trials* **9**, 289–311.

[6] Lachin, J.M. (1988). Properties of simple randomization in clinical trials, *Controlled Clinical Trials* **9**, 312–326.

Chau Thach and Vance W. Berger

Simple Random Sampling

Vance W. Berger and Jialu Zhang

Volume 4, pp. 1840–1841

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Simple Random Sampling

When a census of the entire population of interest is difficult to obtain, a sample is often used. There are many sampling designs that can be used to obtain a sample that would hopefully be representative of the population, and among these sampling designs, several allocate the same inclusion probability to each unit in the population (*see* **Survey Sampling Procedures**). Of course, for this to be the case, one would need the sampling frame from which units are selected to match the population of interest. Otherwise, any unit not in the sampling frame cannot be selected, or, rather, is selected with probability zero.

If the sampling frame is equivalent to the population of interest, and if each unit in the population (or sampling frame) has the same inclusion probability, and if all the units are independent, then we have a simple random sample. The independence refers to the selection probability, so that knowledge that not only does each unit have a common selection probability but also each pair of units has a (different) common selection probability. Another way to formulate this is to state that each sample consisting of a given number of units has the same probability of becoming the selected sample.

Simple random sampling treats each element in the population as being equally important; if this is true, then it would make sense that each unit would be sampled with equal probability. Sometimes this is not the case. For example, the population may be heterogeneous, with small but important subgroups that need to be represented adequately. In such a case, one might want to over-sample from these small subgroups. Also, if it is known that the subgroups are relatively equally well represented in the population, then this fact might be exploited to ensure balance in the sample, as well as in the population. For example, if sampling is undertaken without regard to gender, then it is possible that a gross imbalance would occur in the sample, in which case one gender would have more precise estimation than the other. Specifying a common number of males and females, and possibly selecting simple random samples from each, would guard against this potential problem. Of course, such a compound simple random sample does not itself constitute a simple random sample, because not all subgroups would have the same probability of being selected. For example, a subgroup with unequal numbers of males and females would have probability zero of being selected.

Simple random sampling may be conducted with replacement or without replacement. As the name would suggest, simple random sampling with replacement involves sampling with replacement from the population. Each element in the population may be selected into the sample more than once. Consider, for example, an urn with 20 numbered balls. To obtain a sample of five balls by simple random sampling with replacement, one needs to draw a ball randomly from the urn five times, but after each selection the ball is put back into the urn before the next draw. Therefore, the sampling population remains the same for each draw. If we denote the size of the total population by N, then each unit has a chance of $1/N$ to be selected into the sample at each selection. A specific sample of size n has a chance of $(1/N)^n$ of being selected. This design has been modified to create the play-the-winner rule of patient allocation to clinical trials. Specifically, after a ball is selected, one replaces not only the selected ball but also more balls of the same color or of a different color depending on the outcomes [2].

In simple random sampling without replacement, units are sampled from the population without replacement. Consider the same example as above. To obtain a sample of 5 balls from the urn with 20 numbered balls by simple random sampling without replacement, one draws a ball randomly from the urn 5 times. But now the ball is not replaced after the selection, so after each draw, the population is different from what it was during the previous draw. After 5 draws, only 15 balls are left in the urn. And unlike the simple random sampling with replacement, the 5 balls in the sample are always distinct. With a population of size N, the chance of obtaining a specific sample of size n is $1/\binom{N}{n}$. The probability of obtaining a specific sample is now different from what it was when sampling with replacement. Note that even a simple random sample, be it with replacement or without, cannot confer independence to multiple observations taken from the same unit that was selected [1].

References

[1] Max, L. & Onghena, P. (1999). Some issues in the statistical analysis of completely randomized and repeated measures designs for speech, language, and hearing research, *Journal of Speech, Language, and Hearing Research* **42**, 261–270.

[2] Rosenberger, W. & Lachin, J.M. (2002). *Randomization in Clinical Trials: Theory and Practice*, John Wiley & Sons, New York.

Vance W. Berger and Jialu Zhang

Simple V Composite Tests

Ranald R. Macdonald

Volume 4, pp. 1841–1842

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Simple V Composite Tests

In a statistical test to test if a statistic S is significantly bigger than some fixed value c, the chance or **sampling distributions** of S based on an appropriate null hypothesis has to be assumed. This usually takes the form of supposing that S is c plus a variable error with an expected value of 0. Thus, under the null hypothesis the expected value of S is c and all the variation in the observed value of S is due to sampling error. The P value is then the probability of getting an S as or more extreme than was observed from this distribution (in a two-tailed test) (*see* **Classical Statistical Inference: Practice versus Presentation**). A more usual way of presenting this is to suppose that c is the value of an unknown parameter θ that is involved in the specification of the sampling distribution and that the null hypothesis H_0 is $\theta = c$. This hypothesis is called a simple hypothesis if θ is the only unknown parameter in the sampling distribution. Examples of tests of simple hypotheses include testing that a penny is fair on the basis of the proportion of heads in 100 tosses or that the mean IQ of a class is equal to 100, where IQs are supposed to be normally distributed with a standard deviation of 15.

Where the sampling distribution has more than one unknown parameter, the null hypothesis H_0 $\theta = c$ is said to be composite. Thus, a test of whether the mean IQ of a class is equal to 100, where IQs are supposed to be normally distributed with a standard deviation to be estimated from the data, is a composite hypothesis. However, composite hypotheses can involve several parameters. A composite null hypothesis has the general form $\theta_1 = c_1, \theta_2 = c_2, \ldots \theta_k = c_k$ with k degrees of freedom, where the θs are unknown parameters and the cs fixed values. The general form incorporates hypotheses like $\theta_1 = \theta_2 = \theta_3, \ldots = \theta_k$ as these can be rewritten as $\theta_1 - \theta_2 = 0, \theta_2 - \theta_3 = 0 \ldots \theta_{k-1} - \theta_k = 0$ with $k - 1$ degrees of freedom. This is because the sampling distribution can be reexpressed using the differences between adjacent θs plus the average value of θ instead of the θs themselves. Examples of tests of composite hypotheses include **analyses of variance**, which test the hypothesis that the means of several groups are the same, chi-square tests with several degrees of freedom, and **meta analyses**, which test the hypothesis that there is an effect present in a number of studies.

Statistical tests are also used to test nonparametric hypotheses. This term is somewhat loose and can be used in at least two ways [2]. First, it can apply to hypotheses where the probability distribution of the data is not specified. Examples here include using a Mann Whitney test (*see* **Wilcoxon–Mann–Whitney Test**) to test for differences between two sets of data that have been assumed to have been randomly sampled from identically shaped but unknown distributions and tests based on **bootstrap** methods where the sampling distributions are generated by randomly sampling the observed data to generate empirical distributions (see [1] and **Bootstrap Inference**). A second usage is that a nonparametric hypothesis can take the form that a set of data is consistent with some probability model or law with an unknown parameter or parameters. For example, one might wish to test the hypothesis that a set of data is normally distributed where both the mean and standard deviation are unknown. Such hypotheses are too complex to be entirely evaluated by the results of a single statistical test (*see* **Model Evaluation**; **Model Selection**).

Finally, it should be noted that all statistical tests of hypotheses whether simple, composite, or nonparametric, ultimately reduce the data to a single statistic. Even where a statistical test is derived from multivariate sampling distributions, the test can be seen as assessing whether its P value, itself a univariate statistic, is too small to be attributed to chance. On the other hand, if a composite null hypothesis has several degrees of freedom, it may be decomposed in a number of ways into independent component null hypotheses each with 1 degree of freedom (*see* **Multiple Comparison Procedures**). By testing these component hypotheses, more information about how the overall hypotheses are violated can be obtained though this raises the possibility of artifactually increasing the achieved levels of significance (*see* **Multiple Testing**). It should also be noted that rejection of any component hypothesis itself implies a rejection of the overall hypothesis. Thus, despite the problems of multiple testing, where one has good reason for expecting that an effect with several degrees of freedom will conform to a particular pattern, the probability of rejecting the overall null hypothesis can be increased by using a test that incorporates these expectations.

References

[1] Seigel, S. & Castellan, N.J. (1988). *Nonparametric Statistics for the Behavioral Sciences*, McGraw Hill, New York.

[2] Stuart, A., Ord, J.K. & Arnold, S. (1999). *Kendall's Advanced Theory of Statistics*, 6th Edition, Vol. 2a, *Classical Inference and the Linear Model*, Arnold, London.

(*See also* **Directed Alternatives in Testing**)

RANALD R. MACDONALD

Simulation Methods for Categorical Variables

Stefan von Weber and Alexander von Eye

Volume 4, pp. 1843–1849

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Simulation Methods for Categorical Variables

The simulation of real processes is a strong and legitimate tool in the hand of the research worker [1]. The parameters and variables used in simulation models can be discretely or continuously distributed. Under many conditions, discretely distributed variables are called **categorical variables**. They are called *categorical random variables* if they possess random characteristics. The following list presents examples of tasks that can meaningfully be approached using categorical random variables

1. *Monte Carlo Sampling* [9, 10]: the computer draws random samples from a population; sampling can be uniform, stratified, or by way of **bootstrap** resampling.
2. *Statistical Design or Simulation Design* [8, 11]: the following is a typical question for which simulation methods are employed: which sample size minimizes the cost of an expensive series of experiments. Costs can arise from the number of trials or the sample size, but also from the loss that results from falsely rejecting true hypotheses.
3. *Numerical mathematics* [4, 13]: Calculation of distributions that are either algebraically not tractable or can be calculated only with great effort; examples of such distributions include mixtures of various discrete distributions (*see* **Finite Mixture Distributions**). Statistical research concerning the performance of tests or the power of tests heavily relies on the means of simulations (e.g., [14]).
4. *Stochastic processes of national or business economy*: Simulation studies are used to test models of changes in population parameters and their impact on a national economy. Categorical random variates in such a study are, for instance, the number and the gender distribution of children in families.
5. *Random sampling to circumvent problems with complete enumeration* [12, 15]: For reasons of time or computer capacity, complete enumeration often causes problems. Therefore, random combinations are often drawn. Examples of applications include the optimization of the order of steps in a process, the optimization of breeding processes, or the minimization of costs. The categorical random variates in these examples are the order of objects, the genetic patterns of the parental population, and the costs that come with combinations of parameters.

Discrete Distributions

A random variable that can assume the discrete values $0, 1, 2, \ldots, n$ has a discrete distribution. In applied research, the discrete distribution, that is, the set of probabilities of patterns of variable categories, is almost always found by analyzing the data of a sample. Examples of such data include responses to questionnaires, code patterns in observational studies, or measurements that are categorized. Other discrete variables describe, for example, the gender distribution at universities over the last 25 years, or the distribution of countries that visiting students come from.

Discrete distributions are often derived by operations on discrete variables, for example, the sum function of one or more categorical variables. For example, the sum of 5 dichotomous variables that are scored as 0 and 1 is a binomially distributed random variable (*see* **Catalogue of Probability Density Functions**) with the six possible outcomes $0, 1, 2, \ldots, 5$. An example of such a distribution describes the number of girls in families with 5 children.

One of the best known ways to create a *multivariate discrete distribution* is to cross two or more discrete variables. This operation yields a **contingency table**. Suppose we study the relationship between customer gender (G; female $= 1$; male $= 2$) and time of the day of shopping (T; $0 - 8$ A.M. $=$ 0; 8 A.M. $- 4$ P.M. $= 1$; 4 P.M. $- 12$ P.M. $= 2$). Crossing G with T results in the two-dimensional $G \times T$ contingency table.

Elements of a Simulation Study

Most simulations involve using computers. Researchers write programs using general purpose software such as FORTRAN, software specialized for mathematical problems such as MAPLE [3], or the programming tools provided by the general purpose statistical software packages such as SAS or SPSS [13]. The integral parts of a simulation study with categorical random variables are:

1. description of the question that the simulation will answer; examples of such questions are the estimation of the **power** of a statistical test, the estimation of the β-error, or the examination of relationships in cross-classifications;
2. a model with a comprehensive list of the factors under study, for example, all categorical variables, the α-error, the sample size, the smallest difference of interest between a cell probability and the probability under independence;
3. a good random number generator that creates uniformly distributed random numbers; alternatively, for large samples, a generator that creates normally distributed random numbers;
4. algorithms that generate the desired discrete distribution from the uniformly distributed random numbers, for instance, the random frequencies of a contingency table;
5. algorithms for the analysis of individual trials, for example, the test of a particular hypothesis for a particular cross-classification; storing of the results of the tests for the individual trial;
6. algorithms that summarize the results for the individual trials; for example, these algorithms describe the number of trials in which a hypothesis was rejected.

Generation of Discrete Random Numbers for Small Samples

Many programming environments provide acceptable generators for uniformly distributed random numbers. However, new generators are constantly being developed [7]. Most generators allow one to create reproducible series of random numbers. These are series that are the same each time the generator is invoked. However, most generators also allow one to create series that are hard to reproduce. For this option, a random function is used that determines the seed or the beginning of a series. As a compromise, one can reject the first k numbers from a series, with k determined anew for each simulation run.

If random numbers are needed that reflect a particular distribution, for example, the binomial distribution with n classes and an *a priori* specified probability for the occurrence of the target element A of {A, B}, a number of options exists. The three most important ones are:

1. *Using existing programs.* This option implies using a developer environment that provides the needed random number generators. For example, the environment MAPLE provides the uniform generator *random*; SAS provides the binomial generator *RanBin*; NEWRAN provides the binomial generator *Binomial*; or CenterSpace provides the binomial generator *RandGenBinomial.*
2. *Transforming uniform random numbers.* Here, one typically starts from an existing uniform random number generator that a programming environment such as Turbo Pascal, C, FORTRAN, or C++ makes available. The random numbers from these generators are then transformed such that the desired distribution results. The probabilities of the categories of the targeted distribution must be determined *a priori*.
3. *Implementing stochastic processes.* First, a drawing process is started using the uniform random number generator. Then, a stochastic process is used to select numbers such that the desired distribution results. Here again, the probabilities of the categories of the targeted distribution must be determined *a priori*.

The transformation of an uniformly distributed random number from the [0, 1] interval to any discretely distributed random variable with n classes (= categories) is a mapping of the [0, 1] interval onto the probability axis of the cumulative sum distribution. If the discrete probabilities are $0 < p_1 < p_2 < \cdots < p_n = 1.0$, the simplest (but by no means the fastest) transformation method involves a series of if-statements. The following example uses the language of Turbo Pascal. Any other programming language could be used. The function RANDOM that is used in this example yields uniformly distributed random numbers from the [0, 1] interval, where the number 1 does not occur. The cumulative probabilities must be given in the vector $\boldsymbol{p}$. That is, the programmer must either initialize the vector with the correct probabilities, or write program code that results in these probabilities. The first element of the vector $\boldsymbol{p}$ is p_1, the second is $p_1 + p_2$, and so on. The last element has the value $p_1 + p_2 + \cdots + p_n$. This value is always 1.0, that is, the simulation is exhaustive. In the example, k is the running index, and the quantity U is a dummy variable. The number of categories is n. The resulting quantity x has the desired distribution.

```
U:=RANDOM
for k:=n downto 1 do if
  (U<p[k]) then x:=k;
```

The stochastic drawing process of the binomial distribution is, for example, the drawing of white or black balls from an urn with returning each ball to the urn after registering its color, that is, with replacement. The probability of drawing a white ball is p, that of a black ball is $1 - p$. One generates a uniformly distributed random number $0 \leq \mathrm{U} \leq 1$, using the function RANDOM, and takes 'white' or sets event $= 1$ if $\mathrm{U} < p$. Alternatively, one takes 'black' or event $= 0$ if $\mathrm{U} \geq p$. The desired n-class distribution is found by adding the n drawing events with values of 0 or 1. The categorical random variable is binomially distributed with he $n + 1$ categories being $0, 1, 2, \ldots, n$. The following code (in Turbo Pascal) illustrates this procedure:

```
x:=0;
for k:=1 to n do if (RANDOM<p)
  then x:=x+1;
```

Whereas the first example (the transformation procedure above) is applicable to *any* distribution with a finite number of categories (as was mentioned above, the probabilities of the categories must be known), this second procedure creates only binomial distributions.

Generation of Discrete Random Numbers for Large Samples

If the expected frequency of a category is greater than 9, one can save computing time by using an approximative solution. Naturally, approximative solutions imply compromises concerning the characteristics of the resulting distribution. To give an example, we consider the generation of $2 \times 2 \times 2$ contingency tables with eight cells and multinomial distribution (*see* **Catalogue of Probability Density Functions**). The sample size N is assumed to be greater than 100. The probability p_j of Cell j should be known in this case. Therefore, the expectancy $e_j = Np_j$ is known also. Each cell frequency n_j of the eight cells is binomially distributed with $N + 1$ categories $0, 1, 2, \ldots, N$, according to the cell probability p_j. The variance of a binomially distributed frequency is $\sigma^2 = Np_j(1 - p_j)$. Neglecting the mostly insignificant skew of the binomial distribution, and also neglecting the mostly small deviation of the value $1 - p_j$ from the value 1, we can use the normal distribution of the cell frequency with mean Np_j and variance $\sigma^2 = Np_j$ as a sufficiently good approximation of the binomial distribution of a cell frequency.

To illustrate this method of creating binomially distributed cell frequencies, we now describe the calculation of a single discrete random cell frequency n_j by the following Turbo Pascal code. The quantity r is a real **dummy variable**, the variable j indicates the cell number.

```
r:=N*p[j];
  (* Expectation of cell j*)
r:=r+NORMRAND*sqrt(r);
  (* Adding standard deviation *)
if (r<0) then r:=0; (* Check
  against negative number *)
n[j]:=round(r); (* make it discrete
  and store *)
```

The sum of all cell frequencies n_j, S, is the desired sample size, N.

If there is no generator available that creates $N\{0; 1\}$, that is, normally distributed random numbers, one can use the method of summing 12 uniformly distributed random numbers from the interval [0, 1]. The Turbo Pascal code for this procedure is

```
Function NORMRAND : real;
  (* normally {N,0;1}
    distributed using the central
    limit theorem of Gauss *)
var sum: real;
  i: integer;
begin
  sum:=0;
  for i:=1 ro 12 do sum:=sum+
  RANDOM; NORMRAND:=sum-6;
end;
```

The procedure proposed by Box and Muller (1957) works with nearly the same speed. It uses two random numbers, $r1$ and $r2$, with uniform [0, 1] distribution, and transforms them to be a normally distributed $N\{0; 1\}$ random number. The transformation uses the mathematical functions square root, natural logarithm, and sine. The generated values are not limited to the interval between -6 and $+6$, as are the values generated by the method of summing 12 uniformly distributed random numbers. The Turbo Pascal code for Box and Muller's method follows.

```
Function NORMRAND : real;
  (* normally N{0; 1}
    distributed random numbers
    using a transformation by
    Box and Muller *)
var r1 : real;
  (* Number pi is pre-defined
    in Turbo Pascal *)
begin
     repeat r1:=RANDOM
       until r1>0;
     NORMRAND:=sqrt(−2*ln(r1))*
       sin(2*pi*RANDOM);
end;
```

Generation of Other Discrete Distributions

In behavioral research, a small number of basic theoretical discrete distributions is used. Among these are the binomial, the multinomial, the product-multinomial, the hypergeometric, and the Poisson distributions [6] (*see* **Catalogue of Probability Density Functions**). If one draws samples from samples from a population, we find ourselves in a situation in which frequencies follow both the binomial and the hypergeometric distributions. This case is of practical interest, because questionnaires are often administered in small units of larger entities, for example, departments of a college, small subsidiary plants.

The Poisson distribution is a special case because the number of categories is not restricted *a priori*. If categories exist with numbers $k \gg \lambda$, where λ is the expectation, then these categories come with very small probabilities, one truncates the distribution by eliminating large values of k. The remaining probabilities are then distributed proportionally to the first n categories. When programming Poisson processes, the transformation method is preferred over programming as a stochastic process.

In the following paragraphs, we illustrate the creation of a binomial distribution in a two-dimensional cross-classification. In the first step, we use the transformation method to generate Poisson-distributed random numbers.

In the following example, we simulate the answers that exactly 100 women and 100 men gave to a question. The question concerned a symptom such as headaches. The answers were scaled as 1 = none to 5 = severe. We assume $\lambda_f = 2.5$ for the female respondents, and $\lambda_m = 3.5$ for the male respondents. The following steps are performed to generate the code for the 2×5 contingency table CT[i, k], with $i = 1, 2$ and $k = 1, \ldots, 5$.

1. Set all cells of CT[i, k] to zero.
2. Compute, using the Poisson formula $P_k = \lambda^k/k!e^{-\lambda}$ for $k = 1, 2, \ldots, 5$, the 2×5 class probabilities, and store them into the vectors P_f for women and P_m for men.
3. Calculate the two sums of probabilities S_{Pf} and S_{Pm} from the two vectors P_f and P_m to truncate the distributions. Divide the probabilities in vector P_f by S_{Pf}, and P_m by S_{Pm} (renorming by prorating).
4. Calculate the cumulative probabilities in the vectors P_f and P_m using commands such as 'for k := 2 to 5 do P[k] := P[k] + P[k − 1];'
5. Repeat 100 times the algorithm 'U := RANDOM; for k := 5 downto 1 do if (U < Pf[k]) then x := k; CT[1,x] := CT[1,x] + 1;'. This algorithm yields the simulated frequencies for the female respondents.
6. Repeat 100 times the algorithm 'U := RANDOM; for k := 5 downto 1 do if (U < Pf[k]) then x := k; CT[2,x] := CT[2,x] + 1;'. This algorithm yields the simulated frequencies for the male respondents.
7. After completing Step 6, the contingency table is ready for further analysis, for example, for calculation of a test statistic that describes the association between Gender and headaches.

In the following sections, we illustrate the stochastic drawing process of creating random numbers for discrete distributions for the case in which the class probabilities are unknown. The drawing process for the *hypergeometric distribution* (*see* **Catalogue of Probability Density Functions**) uses the urn example, just as for the binomial distribution, but without replacing the ball. As a consequence, each drawing changes the probabilities of the balls that remain in the urn. Let N_p be the total number of balls (respondents, responses, observations) in a subpopulation, for example, the employees of a factory, and the number of white balls i_p, and the number of black balls $j_q = N_p - j_p$. The start probability p of white balls is $p = j_p/N_p$. The following code, in Turbo Pascal, illustrates the drawing process for the generation of hypergeometrically distributed random numbers x with value range $0, 1, \ldots, \min(n, j_p)$. The parameters of this process are the *a priori* probability p, the number n of categories, and the size N_p of the subpopulation.

```
x:=0;
for k:=1 to n do
begin if RANDOM<p then begin
  x:=x+1; jp=jp-1; end;
     p:=jp/(Np-k);
end;
```

An extension of the binomial distribution is the *multinomial distribution*. This is the distribution of variables with more than two categories. Examples of such variables are the die with six possible numbers and the $k = 5$ possible categories for a question in a questionnaire. In simulations, the researcher has to determine the probabilities $p_1, p_2, \ldots, p_k$ of the alternatives. The simplest option is to specify a uniform distribution, as what one would expect of a good die. Generally, finding these probabilities is often the goal of the simulation process. The random variate X is, after k drawings. k-dimensional, that is, ${x_1}^n$ is the frequency of Alternative $A_1, {x_2}^n$ is the frequency of Alternative A_2, and so on. The generation of a multinomial distribution can be performed using the above transformation procedure. We compute the vector p of the cumulative probabilities, set all elements of vector x to zero, and apply the random generation n times. After the n drawings, we find the k multinomially distributed frequencies as elements of vector x, that is, $x_1, x_2, \ldots, x_k$, each of which has the value range $0, 1, \ldots, n$. This procedure is illustrated by the following Turbo Pascal code.

```
U:=RANDOM;
for k:=5 downto 1 do if
  (U<p[k]) then j:=k;
x[j]:=x[j]:+1;
```

A *product-multinomial distribution* results from simultaneously analyzing two or more multinomial distributions, each of which having the same number of categories. The drawings of the different multinomial distributions are performed separately. The drawing process is the same as described above for the multinomial distribution. The result of a product-multinomial drawing can most conveniently be presented in the form of a cross-tabulation. For example, 50 men and 100 women respond to the same question using a 5-category response format. The resulting table has 2×5 cells. Let the first row contain the response frequencies of the male respondents. These frequencies can assume the values $0, 1, 2, \ldots, 50$. This applies accordingly to the response frequencies for the women, in the second row. The total sum of all frequencies is 150. The row totals are 50 and 100, respectively.

A combined *binomial-hypergeometric distribution* results from drawing $n < N_p$ respondents from a finite subpopulation of size N_p, without replacement. The Turbo Pascal program given below illustrates this procedure. For this program, we need two loop counters, i and j, the *a priori* probability p for the occurrence of Alternative A_1 from $\{A_1, A_2\}$ in the finite basic population, the size N_p of our subpopulation, the number of classes, n, a dummy vector $\boldsymbol{b}$ for storing the N_p scores of the alternatives A_1/A_2. From the vector $\boldsymbol{b}$, the algorithm draws randomly the n respondents. Variable x has the desired binomial-hypergeometric distribution with value range $0, 1, 2, \ldots, n$.

```
For j:=1 to Np do (* Create
  binomially distributed 0 or 1 *)
  if RANDOM<p then b[j]:=1
    else b[j]:=0;
i:=0;
x:=0;
repeat
  j:=round(Np*RANDOM+0.5);
    (* Random access index *)
  if ((j>0) and (j<=Np))
    then (* Check of
    index range *)
  begin
     if b[j]>=0 then
       (* Check whether
       selected already *)
     begin
       x:=x+b[j];
         (* add 0 or 1 *)
       b[j]:= --1;
         (* mark proband
         as selected *)
       i:=i+1;
         (* count selected
         probands *)
     end;
  end;
until (i=n);
  (* Stop with case n *)
```

Analysis and Registration of the Elementary Events and Statistical Summary

The analysis of a single drawing varies depending on the aims of a study [5]. In most studies, results are derived directly from the random numbers. In our example with the Poisson distribution, one could calculate a X^2-statistic for the cross-classification with the random numbers. The same applies to the example in which a product-multinomial distribution was simulated. Alternatives include studies in which running processes are simulated, for example, the evolution of a population. Here, we have a starting state, and after a finite number of drawings, we reach the end state. In this case, the computation of statistical summaries is already part of the analysis of the individual drawing. However, the result of the individual drawing is not of interest per se. One needs the summary of large numbers of drawings for reliable results.

The statistical summary describes the result of the simulation. In our example with the two-dimensional cross-classifications, the summary presents the number N_A of times the null hypothesis was rejected in relation to the total number of all simulation trials, N_R. A second result can be the Type II error (β or the complement of the statistical power of a test), calculated as $\beta = 1 - N_A/N_R$. Typically, simulation studies are meaningful because of the variation of parameters. In our examples with the two-dimensional cross-classification, we can, for instance, increase the sample size from $N = 20$ to $N = 200$ in steps of $\Delta N = 20$. Based on the results of this variation in N, we can produce graphs that show the degree to which the β-error depends on the sample size, N, while holding all other parameters constant (for an example, see [14]). Using this example, one can also show that the individual table is not really of interest. For the individual table, the β-error can only be either 0% or 100%, because we accept the correct alternative hypothesis or we reject it.

Using Different Computers and Parallel Computing

Simulation experiments of statistical problems are prototypical tasks for multiple computers and parallel computing. Suppose researchers have written a computer program or have produced it using a developer environment, they can run the program under different parameter combinations using different computers. This way, the time needed for a study can be considerably shortened, the number of parameters can be increased, or the range of parameter variation can be broadened.

If a computer with multiple processors is available, for example, a machine with 256 processor units, one can employ software for parallel computing. The task of the programmer here is to start 256 trials using 256 random number processes (using 256 different seeds). At the end, the 256 resulting matrices need to be summarized.

Presenting Discrete Distributions

For most observed discrete distributions, theoretical distributions are defined. (*see* **Catalogue of Probability Density Functions**) These include the binomial, the hypergeometric and the multinomial distributions. The graphical representation of the *density* distribution is, because of the categorical nature of the distributions, a bar diagram (*see* **Bar Chart**) rather than a smooth curve. The probability of each class or category is shown by the corresponding height of a bar. These can be compared with bars whose height is determined by the theoretical distribution, that is, the comparison bars function as expected values. This way, the results of a simulation study can be evaluated with reference to some theoretical distribution.

The graphical representation of contingency tables with many cells is more complex. Here, the mosaic methods proposed by Friendly [2] are most useful.

References

[1] Drasgow, F. & Schmitt, N. eds, (2001). *Measuring and Analyzing Behavior in Organizations*, Wiley, New York.

[2] Friendly, M. (1994). Mosaic displays for multi-way contingency tables, *Journal of the American Statistical Association* **89**, 190–200.

[3] Heck, A. (1993). *Introduction to MAPLE*, Springer Verlag, New York.

[4] Indurkhya, A. & von Eye, A. (2000). The power of tests in configural frequency analysis, *Psychologische Beiträge* **42**, 301–308.

[5] Klein, K.J. & Koslowski, S.W.J., eds, (2000). *Multilevel Theory, Research, and Methods in Organizations*, Jossey Bass, New York.

[6] Mann, P.D. (1995). *Statistics for Business and Economics*, Wiley, New York.

[7] Marsaglia, G. (2000). *The Monster-A Random Number Generator with Period Over 10^{2857} times Longer as Long as the Previous Touted Longest-Period One*, Boeing Scientific Research Laboratories, Seattle. Unpublished paper.

[8] Mason, R.L., Gunst, R.F. & Hess, J.L. (1989). *Statistical Design and Analysis of Experiments*, Wiley, New York.

[9] Nardi, P.M. (2002). *Doing Survey Research-A guide to Quantitative Methods*, Allyn & Bacon, Boston.

[10] Noreen, E.W. (1989). *Computer Intensive Methods for Hypothesis Testing*, Wiley, New York.

[11] Rasch, D., Verdooren, L.R. & Gowers, J.I. (1999). *Fundamentals in the Design and Analysis of Experiments and Surveys*, Oldenbourg, München.

[12] Russell, S. & Norvig, P. (2003). *Artificial Intelligence-A Modern Approach*, 2nd Edition, Prentice Hall, Englewood Cliffs.

[13] Stelzl, I. (2000). What sample sizes are needed to get correct significance levels for log-linear models? A Monte Carlo study using the SPSS procedure "Hiloglinear", *Methods of Psychological Research-Online* **5**, 95–116.

[14] von Weber, S., von Eye, A. & Lautsch, E. (2004). The type II error of measures for the analysis of 2 × 2 tables, *Understanding Statistics* **3**(4), 259–282.

[15] Wardrop, R.L. (1995). *Statistics: Learning in the Presence of Variation*, William C. Brown, Dubuque.

STEFAN VON WEBER AND
ALEXANDER VON EYE

Simultaneous Confidence Interval

Chris Dracup

Volume 4, pp. 1849–1850

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Simultaneous Confidence Interval

It is sometimes desirable to construct **confidence intervals** for a number of population parameters at the same time. The problems associated with this are demonstrated most simply by considering intervals that are independent (in the sense that they are constructed from independent data sources). If two independent 95% confidence intervals are to be computed, then the probability that both will contain the true value of their estimated parameter is $0.95 \times 0.95 = 0.9025$. The probability that three such intervals will all contain their true parameter value falls to $0.95 \times 0.95 \times 0.95 = 0.857$. It is apparent that the more such intervals are to be computed, the smaller will be the probability that they will all contain their true parameter value. With independent confidence intervals, it would be relatively easy to increase the level of confidence for the individual intervals so as to make the overall confidence equal to 0.95. With two intervals, each could be constructed to give 97.47% confidence, since $0.9747 \times 0.9747 = 0.950$. With three intervals, each could be constructed to give 98.30% confidence, since $0.9830 \times 0.9830 \times 0.9830 = 0.950$.

With truly independent confidence intervals, which are addressing very different questions, researchers are not likely to be overconcerned about having a simultaneous confidence statement for all their estimated parameters. However, in the behavioral sciences, the construction of simultaneous confidence intervals is most commonly associated with experiments containing k conditions ($k > 2$). In such situations, researchers often wish to estimate the difference in population means for each pair of conditions, and to express these as confidence intervals. The number of such pairwise comparisons is equal to $k(k-1)/2$, and this figure increases rapidly as the number of conditions increases. Furthermore, the desired intervals are not statistically independent as the data from each condition contribute to the construction of more than one interval (*see* **Multiple Testing**).

The simple multiplication rule used above with independent confidence intervals is no longer appropriate when the confidence intervals are not independent, but the approach is basically the same. The individual intervals are constructed at a higher confidence level in such a way as to provide, say, 95% confidence that all the intervals will contain their true parameter value. One of the most commonly used approaches employs the studentized range statistic [1, 2], and is the confidence interval version of Tukey's Honestly Significant Difference test (*see* **Multiple Comparison Procedures**). It will serve to illustrate the general approach.

For any pair of conditions, i and j, in a simple k-group study, the simultaneous 95% confidence interval for the difference in their population means is given by

$$(\bar{X}_i - \bar{X}_j) - q_{0.05,k,N-k}\sqrt{\frac{MS_{\text{error}}}{n}} \le \mu_i - \mu_j \le (\bar{X}_i - \bar{X}_j) + q_{0.05,k,N-k}\sqrt{\frac{MS_{\text{error}}}{n}} \quad (1)$$

where MS_{error} is the average of the sample variances of the k conditions, n is the number of observations in each condition (assumed equal), and $q_{0.05,k,N-k}$ is the critical value of the studentized range statistic for k conditions, and $N-k$ degrees of freedom, at the 0.05 significance level. By way of illustration, consider an experiment with $k = 3$ conditions each with $n = 21$ randomly assigned participants. With $k = 3$ and $N - k = 60$, the critical value of the studentized range statistic at the 0.05 level is 3.40. If the means of the samples on the dependent variable were 77.42, 69.52 and 64.97, and the MS_{error} was 216.42, then the simultaneous 95% confidence intervals for the difference between the population means would be

$$\begin{aligned} -3.01 &\le \mu_1 - \mu_2 \le 18.81 \\ 1.54 &\le \mu_1 - \mu_3 \le 23.36 \\ -6.36 &\le \mu_2 - \mu_3 \le 15.46 \end{aligned} \quad (2)$$

If confidence intervals are constructed in this way for all the pairs of conditions in an experiment, then the 95% confidence statement applies to all of them simultaneously. That is to say, over repeated application of the method to the data from different experiments, for 95% of the experiments *all* the intervals constructed will include their own true population difference. The additional confidence offered by such techniques is bought at the expense of a rather wider interval for each pair of conditions than would be required if that pair were the sole interest of the researcher. In the example, only the confidence interval for the difference in means between conditions 1 and 3 does not include zero. Thus, Tukey's Honestly Significant Difference test would show only these two conditions to differ significantly from one another at the 0.05 level.

bsa588Scheffé [2] provided a method for constructing simultaneous confidence intervals for all possible contrasts (actually an infinite number) in a k-group study, not just the simple comparisons between pairs of conditions. But, once again, this high level of protection is bought at the price of each interval being wider than it would have been if it had been the only interval of interest to the researcher. As an example of the particular flexibility of the approach, consider an experiment with five conditions. Following inspection of the data, it may occur to the researcher that the first two conditions actually have a feature in common, which is not shared by the other three conditions. Scheffé's method allows a confidence interval to be constructed for the difference between the population means of the two sets of conditions (i.e., for $\mu_{1\&2} - \mu_{3\&4\&5}$), and for any other contrast that might occur to the researcher. If Scheffé's method is used to construct a number (possibly a very large number) of 95% confidence intervals from the data of a study, then the method ensures that in the long run for at most 5% of such studies would any of the calculated intervals fail to include its true population value.

Like Tukey's Honestly Significant Difference test, Scheffé's test (*see* **Multiple Comparison Procedures**) is usually applied in order to test nil null hypotheses (*see* **Confidence Intervals**). However, it can be applied to test whether two sets of conditions have means that differ by some value other than zero, for example, $\mathrm{H}_0 : \mu_{1\&2} - \mu_{3\&4\&5} = 3$. Any hypothesized difference that is not included in the calculated simultaneous 95% confidence interval would be rejected at the 0.05 level by a Scheffé test.

References

[1] Kendall, M.G., Stuart, A. & Ord, J.K. (1983). *The Advanced Theory of Statistics*, *Vol. 3, Design and Analysis, and Time-Series*, 4th Edition, Griffin, London.

[2] Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance, *Biometrika* **40**, 87–104.

Chris Dracup

Single-Case Designs

Patrick Onghena

Volume 4, pp. 1850–1854

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Single-Case Designs

Single-case designs refer to research designs that are applied to experiments in which one entity is observed repeatedly during a certain period of time, under different levels ('treatments') of at least one independent variable. The essential characteristics of such single-case experiments are (a) that only one entity is involved (single-case), and (b) that there is a manipulation of the independent variable(s) (experiment). These characteristics imply that (c) the entity is exposed to all levels of the independent variable (like in a within-subject design), and (d) that there are repeated measures or observations (like in a longitudinal or time series design). In this characterization of single-case designs, we use the generic term 'entity' to emphasize that, although 'single-case' is frequently equated with 'single-subject', also experiments on a single school, a single family or any other specified 'research unit' fit the description [1, 15, 18, 20].

A Long Tradition and a Growing Impact

Single-case designs have a long history in behavioral science. Ebbinghaus' pivotal memory research and Stratton's study on the effect of wearing inverting lenses are classical nineteenth century experiments involving a single participant that had a tremendous impact on psychology [6]. During the twentieth century, the influential work of Skinner [27] and Sidman [26] provided the major impetus for continued interest in these kind of designs, and it was their methodological approach that laid the foundations of the current popularity of single-case research in behavior modification and clinical psychology [1, 18], neuropsychology [3, 31], psychopharmacology [4, 5], and educational research [20, 22].

In these areas, single-case research is evidently one of the only viable options if rare or unique conditions are involved, but it also seems to be the research strategy of first choice in research settings where between-entity variability is considered negligible or if demonstration in a single case is sufficient to confirm the existence of a phenomenon or to refute a supposedly universal principle. Another potential motivation to embark upon single-case research is its (initial) focus on the single case, which mimics the care for the individual patient that is needed in clinical work. In such applied settings, generalization is sometimes only a secondary purpose, which can be achieved by replication and aggregation of single-case results. In addition, the replicated single-case designs model is much more consistent with the way in which consecutive patients are entered into **clinical trials** than the random sampling model underlying many group designs and standard statistical techniques.

Single-case Designs, Case Studies, and Time Series

Single-case research using experimental designs should not be confused with case study research or observational time series research. In a traditional case study approach, a single phenomenon is also studied intensively, but there is not necessarily a purposive manipulation of an independent variable nor are there necessarily repeated measures (*see* **Case Studies**). Furthermore, most case studies are reported in a narrative way while results of single-case experiments usually are presented numerically or graphically, possibly accompanied by sophisticated statistical analyses. In observational time series research, there are also repeated measures and very often complex statistical analyses, but the main difference with single-case experiments lies in the absence of a designed intervention. Although time series **intervention analyses** can be applied to results from simply designed **interrupted time series** experiments (if the number of observations is large enough), a designed intervention is not crucial for time series research. Incidentally, the vast majority of applications of **time series analysis** concern mere observational series.

To Randomize or not to Randomize

An important consideration in designing a single-case experiment is whether or not to randomize (i.e., to randomly assign measurement occasions to treatments) [10]. This randomization provides statistical control over both known and unknown **confounding variables** that are time-related (e.g., 'history' and 'maturation'), and very naturally leads to a statistical test based on the randomization as it was implemented in the design, a so-called **randomization test** [8, 9, 28]. In this way, randomization can improve both the internal and the statistical-conclusion validity of the study (*see* **Internal Validity**). In a nonrandomized single-case experiment (e.g., in an operant response-guided experiment), one has to be very cautious when attributing outcome changes to treatment changes. In a nonrandomized intervention study, for example, one usually does not control the variables that covary with the intervention, or with the decision to intervene, making it very difficult to rule out response-guided biases [28] or **regression artifacts**. Therefore, the control aspect of randomization might be considered as essential to single-case experiments as it is to multiple-case experiments and **clinical trials** [10, 11, 28].

We can distinguish two important schedules by which randomization can be incorporated into the design of a single-case experiment. In the first schedule, the treatment alternation is randomly determined and this gives rise to the so-called *alternation designs*. In the second schedule, the moment of intervention is randomly determined, and this gives rise to the so-called *phase designs*. Both randomized alternation and randomized phase designs will be presented in the next two sections, followed by a discussion of two types of replications: simultaneous and sequential replications (*see* **Interrupted Time Series Design**).

Randomized Alternation Designs

In alternation designs, any level of the independent variable could be present at each measurement occasion. For example, in the *completely randomized single-case design*, the treatment sequence is randomly determined only taking account of the number of levels of the independent variable and the number of measurement occasions for each level. If there are two levels (A and B), with three measurement occasions each, then complete randomization implies a random selection among twenty possible assignments.

If some sequences of a complete randomization are undesirable (e.g., AAABBB), then other families of alternation designs can be devised by applying the classical randomization schemes known from group designs. For example, a *randomized block single-case design* is obtained in the previous setting if the treatments are paired, with random determination of the order of the two members of the pair. The selection occurs among the following eight possibilities: ABABAB, ABABBA, ABBAAB, ABBABA, BABABA, BABAAB, BAABBA, and BAABAB.

Although this randomized block single-case design is rampant in double-blind single-patient medication trials [16, 17, 23], it is overly restrictive if one only wants to avoid sequences of identical treatments. Therefore, a randomized version of the **alternating treatments design** was proposed, along with an algorithm to enumerate and randomly sample the set of acceptable sequences [25]. For the previous example, a constraint of two consecutive identical treatments at most results in six possibilities in addition to the eight possibilities listed for the randomized block single-case design: AABABB, AABBAB, ABAABB, BBABAA, BBAABA, and BABBAA.

This larger set of potential randomizations is of paramount importance to single-case experiments because the smallest P value that can be obtained with the corresponding randomization test is the inverse of the cardinality of this set. If the set is too small, then the experiments have zero statistical **power**. Randomized alternating treatments designs may guarantee sufficient power to detect treatment effects while avoiding awkward sequences of identical treatments [12, 25].

It should be noted that more complex alternation designs also can be constructed if there are two or more independent variables. For example, a *completely randomized factorial single-case design* follows on from crossing the levels of the independent variables involved.

Randomized Phase Designs

If rapid and frequent alternation of treatments is prohibitive, then researchers may opt for a phase design. In phase designs, the complete series of measurement occasions is divided into treatment phases and several consecutive measurements are taken in each phase. The simplest phase design is the *AB design* or the basic **interrupted time series design**. In this design, the first phase of consecutive measurements is taken under one condition (e.g., a baseline or control condition) and the second phase under another condition (e.g., a postintervention or treatment condition). All sorts of variations and extensions of this AB design can be conceived of:

ABA or *withdrawal and* **reversal designs**, *ABAB, ABABACA designs*, and so on [1, 15, 18, 20].

In phase designs, the order of the phases is fixed, so the randomization cannot be applied to the treatment sequence like in alternation designs. However, there is one feature that can be randomized without distorting the phase order and that is the moment of phase change (or the moment of intervention). In such randomized phase designs, only the number of available measurement occasions, the number of phases (c.q., treatments), and the minimum lengths of the phases should be specified [9, 24]. A randomized AB design with six measurement occasions and with at least one measurement occasion in each phase, for example, implies the following five possibilities: ABBBBB, AABBBB, AAABBB, AAAABB, and AAAAAB. There are, of course, many more repeated measurements in the phase designs of typical applications (e.g., by using a diary or psychophysical measures) and often it is possible to add phases and thereby increase the number of phase changes. In fact, a large number of measurement occasions and/or more than one phase change is necessary to obtain sufficient statistical power in these designs [14, 24].

Simultaneous and Sequential Replication Designs

Replication is the obvious strategy for demonstrating or testing generalizability of single-case results. If the replications are planned and part of the design, then the researcher has the option to conduct the experiments at the same time or conduct them one by one.

Simultaneous replication designs are the designs in which the replications (alternation or phase single-case designs) are carried out at the same time. The most familiar simultaneous replication design is the *multiple baseline across participants design* (*see* **Multiple Baseline Designs**). In such a design, several AB phase designs are implemented simultaneously and the intervention is applied for each separate participant at a different moment [1]. The purpose of the simultaneous monitoring is to control for historical confounding variables. If an intervention is introduced in one of the phase designs and produces a change for that participant, while little or no change is observed for the other participants, then it is less likely that other external events are responsible for the observed change, than if this change was observed in an isolated phase design.

Randomization can be introduced very easily in simultaneous replication designs by just applying the randomization schedules in the several phase and/or alternation designs separately [24]. In addition, between-case constraints can be imposed, for example, to avoid simultaneous intervention points or to obtain systematic staggering of the intervention in the multiple baseline across participants design [19].

If the replications are carried out one by one, then we have a *sequential replication design.* Also for these designs, a lot of options are available to the applied researcher, depending on whether the separate designs should be very similar (e.g., the same number of measurement occasions for all or some of the designs) and whether between-case comparisons are part of the design [22].

The statistical power of the corresponding randomization tests for both simultaneous and sequential replication designs is already adequate (>0.80) for designs with four participants and a total of twenty measurement occasions (for a range of effect sizes (*see* **Effect Size Measures**), autocorrelations and significance levels likely to be relevant for behavioral research) [13,21]. In addition, the sequential replication design also provides an opportunity to apply powerful nonparametric or parametric meta-analytic procedures (*see* **Meta-Analysis**) [2, 7, 24, 29, 30].

References

[1] Barlow, D.H. & Hersen, M., eds (1984). *Single-Case Experimental Designs: Strategies for Studying Behavior Change*, 2nd Edition, Pergamon Press, Oxford.

[2] Busk, P.L. & Serlin, R.C. (1992). Meta-analysis for single-case research, in *Single-Case Research Design and Analysis: New Directions for Psychology and Education*, T.R. Kratochwill & J.R. Levin, eds, Lawrence Erlbaum, Hillsdale, pp. 187–212.

[3] Caramazza, A., ed. (1990). *Cognitive Neuropsychology and Neurolinguistics: Advances in Models of Cognitive Function and Impairment*, Lawrence Erlbaum, Hillsdale.

[4] Conners, C.K. & Wells, K.C. (1982). Single-case designs in psychopharmacology, in *New Directions for Methodology of Social and Behavioral Sciences*, Vol. 13, *Single-Case Research Designs*, A.E. Kazdin & A.H. Tuma, eds, Jossey-Bass, San Francisco, pp. 61–77.

[5] Cook, D.J. (1996). Randomized trials in single subjects: the N of 1 study, *Psychopharmacology Bulletin* **32**, 363–367.

[6] Dukes, W.F. (1965). $N = 1$, *Psychological Bulletin* **64**, 74–79.

[7] Edgington, E.S. (1972). An additive method for combining probability values from independent experiments, *Journal of Psychology* **80**, 351–363.

[8] Edgington, E.S. (1986). Randomization tests, in *Encyclopedia of statistical sciences*, Vol. 7, S. Kotz & N.L. Johnson, eds, Wiley, New York, pp. 530–538.

[9] Edgington, E.S. (1995). *Randomization Tests*, 3rd Edition, Dekker, New York.

[10] Edgington, E.S. (1996). Randomized single-subject experimental designs, *Behaviour Research and Therapy* **34**, 567–574.

[11] Ferron, J., Foster-Johnson, L. & Kromrey, J.D. (2003). The functioning of single-case randomization tests with and without random assignment, *Journal of Experimental Education* **71**, 267–288.

[12] Ferron, J. & Onghena, P. (1996). The power of randomization tests for single-case phase designs, *Journal of Experimental Education* **64**, 231–239.

[13] Ferron, J. & Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs, *Journal of Experimental Education* **70**, 165–178.

[14] Ferron, J. & Ware, W. (1995). Analyzing single-case data: the power of randomization tests, *Journal of Experimental Education* **63**, 167–178.

[15] Franklin, R.D., Allison, D.B., & Gorman, B.S., eds (1996). *Design and Analysis of Single-Case Research*. Lawrence Erlbaum, Mahwah.

[16] Gracious, B. & Wisner, K.L. (1991). Nortriptyline in chronic fatigue syndrome: a double blind, placebo-controlled single-case study, *Biological Psychiatry* **30**, 405–408.

[17] Guyatt, G.H., Keller, J.L., Jaeschke, R., Rosenbloom, D., Adachi, J.D. & Newhouse, M.T. (1990). The n-of-1 randomized controlled trial: clinical usefulness–our three-year experience, *Annals of Internal Medicine* **112**, 293–299.

[18] Kazdin, A.E. (1982). *Single-Case Research Designs: Methods for Clinical and Applied Settings*, Oxford University Press, New York.

[19] Koehler, M.J. & Levin, J.R. (1998). Regulated randomization: a potentially sharper analytical tool for the multiple-baseline design, *Psychological Methods* **3**, 206–217.

[20] Kratochwill, T.R. & Levin, J.R., eds (1992). *Single-Case Research Design and Analysis: New Directions for Psychology and Education*, Lawrence Erlbaum, Hillsdale.

[21] Lall, V.F. & Levin, J.R. (2004). An empirical investigation of the statistical properties of generalized single-case randomization tests, *Journal of School Psychology* **42**, 61–86.

[22] Levin, J.R. & Wampold, B.E. (1999). Generalized single-case randomization tests: flexible analyses for a variety of situations, *School Psychology Quarterly* **14**, 59–93.

[23] Mahon, J., Laupacis, A., Donner, A. & Wood, T. (1996). Randomised study of n of 1 trials versus standard practice, *British Medical Journal* **312**, 1069–1074.

[24] Onghena, P. (1992). Randomization tests for extensions and variations of ABAB single-case experimental designs: a rejoinder, *Behavioral Assessment* **14**, 153–171.

[25] Onghena, P. & Edgington, E.S. (1994). Randomization tests for restricted alternating treatments designs, *Behaviour Research and Therapy* **32**, 783–786.

[26] Sidman, M. (1960). *Tactics of Scientific Research: Evaluating Experimental Data in Psychology*, Basic Books, New York.

[27] Skinner, B.F. (1938). *The Behavior of Organisms: An Experimental Analysis*, Appleton-Century-Crofts, New York.

[28] Todman, J.B. & Dugard, P. (2001). *Single-Case and Small-n Experimental Designs: A Practical Guide to Randomization Tests*, Lawrence Erlbaum, Mahwah.

[29] Van den Noortgate, W. & Onghena, P. (2003). Combining single-case experimental studies using hierarchical linear models, *School Psychology Quarterly* **18**, 325–346.

[30] Van den Noortgate, W. & Onghena, P. (2003). Hierarchical linear models for the quantitative integration of effect sizes in single-case research, *Behavior Research Methods, Instruments, & Computers* **35**, 1–10.

[31] Wilson, B. (1987). Single-case experimental designs in neuropsychological rehabilitation, *Journal of Clinical and Experimental Neuropsychology* **9**, 527–544.

PATRICK ONGHENA

Single and Double-blind Procedures

ARNOLD D. WELL

Volume 4, pp. 1854–1855

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Single and Double-blind Procedures

Suppose we want to test the effectiveness of a drug administered in the form of a pill. We could randomly assign patients to two groups, a treatment group that received the drug and a control group that did not, and then measure changes in the patients (*see* **Clinical Trials and Intervention Studies**). However, this design would not be adequate to evaluate the drug because we could not be sure that any changes we observed in the treatment condition were a result of the drug itself. It is well known that patients may derive some benefit simply from the belief that they are being treated, quite apart from any effects of the active ingredients in the pill.

The study could be improved by giving patients in the control group a **placebo**, pills that have the same appearance, weight, smell, and taste as those used in the drug condition but do not contain any active ingredients. Our goal would be to have a situation in which any systematic differences between groups were due only to the effects of the drug itself. We could assign patients randomly to the groups so that there were no systematic differences in patient characteristics between them. However, in order to rule out possible bias resulting from patients' expectations, it is important that they not be informed as to whether they have been assigned to the drug or the placebo/control condition. By not informing patients about condition assignments, we are employing what is called a *single-blind procedure*.

Another possible source of bias may occur because of what may be referred to as experimenter effects (*see* **Expectancy Effect by Experimenters**). The expectations and hopes of researchers may influence the data by causing subtle but systematic differences in how the researchers interact with patients in different conditions, as well as how the data are recorded and analyzed. Moreover, not only do the patient and experimenter effects described above constitute possible sources of bias that might threaten the validity of the research, the two types of bias can interact. Even if the patients are not explicitly told about the details of the research design and condition assignments, they may learn about them through interactions with the research personnel who are aware of this information, thereby reintroducing the possibility of bias due to patient expectations. If we try to rule out these sources of bias by withholding information from both subjects and researchers, we are said to be using a *double-blind procedure*. The purpose of double-blind procedures is to rule out any bias that might result from knowledge about the experimental conditions or the purpose of the study by making both participants and researchers 'blind' to such information.

These ideas generalize to many kinds of behavioral research. The behavior of participants and researchers may be influenced by many factors other than the independent variable – see for example, [1]. Factors such as information or misinformation about the purpose of the experiment may combine with motivation to obtain particular findings, or to please the experimenter, thereby causing the data to differ from what they would be without this information.

Reference

[1] Rosenthal, R. & Rosnow, R.L. (1969). *Artifacts in Behavioral Research*, Academic Press, New York.

Arnold D. Well

Skewness

KARL L. WUENSCH

Volume 4, pp. 1855–1856

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Skewness

In everyday language, the terms ‘skewed’ and ‘askew’ are used to refer to something that is out of line or distorted on one side. When referring to the shape of frequency or probability distributions, ‘skewness’ refers to asymmetry of the distribution. A distribution with an asymmetric tail extending out to the right is referred to as ‘positively skewed’ or ‘skewed to the right’, while a distribution with an asymmetric tail extending out to the left is referred to as ‘negatively skewed’ or ‘skewed to the left’.

Karl Pearson [2] first suggested measuring skewness by standardizing the difference between the mean and the mode, that is, $sk = (\mu - \text{mode})/\sigma$. Population modes are not well estimated from sample modes, but one can estimate the difference between the mean and the mode as being three times the difference between the mean and the median [3], leading to the following estimate of skewness: $sk_{\text{est}} = (3(M - \text{median}))/s$. Some statisticians use this measure but with the ‘3’ eliminated.

Skewness has also been defined with respect to the third moment about the mean: $\gamma_1 = (\sum(X - \mu)^3)/n\sigma^3$, which is simply the expected value of the distribution of cubed z scores. Skewness measured in this way is sometimes referred to as ‘Fisher’s skewness’. When the deviations from the mean are greater in one direction than in the other direction, this statistic will deviate from zero in the direction of the larger deviations. From sample data, Fisher’s skewness is most often estimated by $g_1 = (n \sum z^3)/((n - 1)(n - 2))$. For large sample sizes ($n > 150$), g_1 may be distributed approximately normally, with a standard error of approximately $\sqrt{(6/n)}$. While one could use this sampling distribution to construct confidence intervals for or tests of hypotheses about γ_1, there is rarely any value in doing so.

It is important for behavioral researchers to notice skewness when it appears in their data. Great skewness may motivate the researcher to investigate outliers. When making decisions about which measure of location to report (means being drawn in the direction of the skew) and which inferential statistic to employ (one that assumes normality or one that does not), one should take into consideration the estimated skewness of the population. Normal distributions have zero skewness. Of course, a distribution can be perfectly symmetric but far from normal. Transformations commonly employed to reduce (positive) skewness include square root, log, and reciprocal transformations.

The most commonly used measures of skewness (those discussed here) may produce surprising results, such as a negative value when the shape of the distribution appears skewed to the right. There may be superior alternative measures not in common use [1].

References

[1] Groeneveld, R.A. & Meeden, G. (1984). Measuring skewness and kurtosis, *The Statistician* **33**, 391–399.

[2] Pearson, K. (1895). Contributions to the mathematical theory of evolution, II: skew variation in homogeneous material, *Philosophical Transactions of the Royal Society of London* **186**, 343–414.

[3] Stuart, A. & Ord, J.K. (1994). *Kendall’s Advanced Theory of Statistics*, *Vol. 1, Distribution Theory*, 6th Edition, Edward Arnold, London.

(*See also* **Kurtosis**)

KARL L. WUENSCH

Slicing Inverse Regression

Chun-Houh Chen

Volume 4, pp. 1856–1863

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Slicing Inverse Regression

Dimensionality and Effective Dimension Reduction Space

Dimensionality is an issue that often arises and sets a severe limitation in the study of every scientific field. In the routine practice of regression analysis (*see* **Multiple Linear Regression**), the curse of dimensionality [11] may come in at the early exploratory stage. For example, a 2-D or **3-D scatterplot** can be successfully applied to examine the relationship between the response variable and one or two input variables. But, when the dimension of regressors gets larger, this graphical approach could become laborious, and it is important to focus only on a selective set of projection directions. In the parametric regression setting, simple algebraic functions of $\mathbf{x}$ are used to construct a link function for applying the **least squares** or **maximum likelihood** methods. In the **nonparametric regression** setting, the class of fitted functions is enlarged. The increased flexibility in fitting via computation intensive smoothing techniques, however, also increases the modeling difficulties that are often encountered with larger number of regressors.

Li [12] introduced the following framework for dimension reduction in regression:

$$Y = g(\beta_1'\mathbf{x}, \dots \beta_k'\mathbf{x}, \varepsilon). \qquad (1)$$

The main feature of (1) is that g is completely unknown and so is the distribution of ε, which is independent of the p-dimensional regressor $\mathbf{x}$. When k is smaller than p, (1) imposes a dimension reduction structure by claiming that the dependence of Y on $\mathbf{x}$ only comes from the k variates, $\beta_1'\mathbf{x}, \dots, \beta_k'\mathbf{x}$, but the exact form of the dependence structure is not specified. Li called this k-dimensional space spanned by the k β vectors the e.d.r. (effective dimension reduction) space and any vector in this space is referred to as an e.d.r. direction. The aim is to estimate the base vectors of the e.d.r. space. The notion of e.d.r. space and its role in regression graphics are further explored in Cook [4]. The primary goal of Li's approach is to estimate the e.d.r. directions first so that it becomes easier to explore data further with either the graphical approach or the nonparametric curve-smoothing techniques.

Special Cases of Model (1)

Many commonly used models in regression can be considered as special cases of model (1). We separate them into one-component models ($k = 1$) and the multiple-component models ($k > 1$). One-component models ($k = 1$) include the following:

1. Multiple linear regression. $g(\beta'\mathbf{x}, \varepsilon) = a + \beta'\mathbf{x} + \varepsilon$.
2. Box–Cox transformation. $g(\beta'\mathbf{x}, \varepsilon) = h_\lambda(a + \beta'\mathbf{x} + \varepsilon)$, where $h_\lambda(\cdot)$ is the power transformation function with power parameter λ given by

$$h_\lambda(t) = \begin{cases} (t^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, \\ \ln(\mathrm{t}) & \text{if } \lambda = 0. \end{cases} \qquad (2)$$

3. Additive error models. $g(\beta'\mathbf{x}, \varepsilon) = h(\beta'\mathbf{x}) + \varepsilon$, where $h(\cdot)$ is unknown.
4. Multiplicative error models. $g(\beta'\mathbf{x}, \varepsilon) = \mu + \varepsilon h(\beta'\mathbf{x})$, where $h(\cdot)$ is usually assumed to be known.
   Multiple-component models ($k > 1$) include the following:
5. **Projection pursuit regression** (Friedman and Stuetzle [8]). $g(\beta_1'\mathbf{x}, \dots \beta_k'\mathbf{x}, \varepsilon) = h_1(\beta_1'\mathbf{x}) + \cdots + h_r(\beta_r'\mathbf{x}) + \varepsilon$, where r may be unequal to k.
6. Heterogeneous error models. $g(\beta_1'\mathbf{x}, \beta_2'\mathbf{x}, \varepsilon) = h_1(\beta_1'\mathbf{x}) + \varepsilon h_2(\beta_2'\mathbf{x})$.

More detailed discussions about these models can be found in [4, 15].

An Example

The following example will be used to illustrate the concept and implementation of SIR throughout this article. Six independent standard normal random variables, $\mathbf{x} = (x_1, \dots, x_6)$, with 200 observations each are generated. The response variable Y is generated according to the following two-component model:

$$\begin{aligned} y &= g(\beta_1'\mathbf{x}, \beta_2'\mathbf{x}, \varepsilon) \\ &= \frac{\beta_1'\mathbf{x}}{0.5 + (\beta_2'\mathbf{x} + 1.5)^2} + 0 \cdot \varepsilon, \end{aligned} \qquad (3)$$

where $\beta_1' = (1, 1, 0, 0, 0, 0)$ and $\beta_2' = (0, 0, 1, 1, 0, 0)$. We employ this noise free model for an easier explanation.

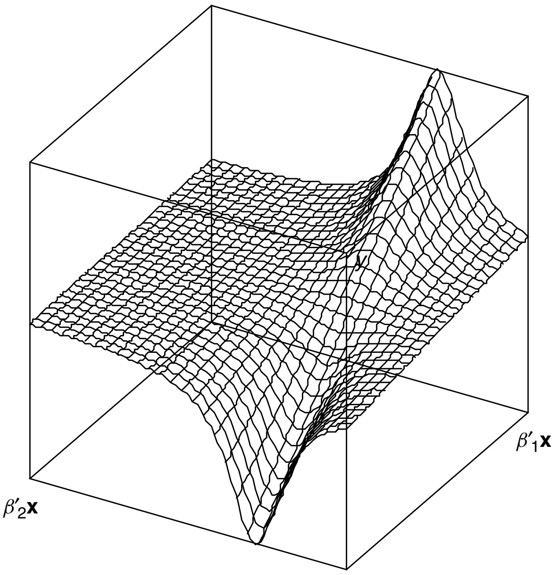


Figure 1 The response surface for function in (2)

Contour Plots and Scatterplot Matrix

The response surface of (2) is depicted in Figure 1. A different way to visualize the structure of the response variable Y is to overlay the scatter plot of $\beta_1'\mathbf{x}$ and $\beta_2'\mathbf{x}$ with the contours of (3), (Figure 2). Because the vectors $\beta_1' = (1, 1, 0, 0, 0, 0)$ and $\beta_2' = (0, 0, 1, 1, 0, 0)$ are not given, how to identify these components is the most challenging issue in a regression problem. One possible way of constructing a contour plot of the response variable on the input variables is through the scatter-plot matrix of paired variables in $\mathbf{x}$. Here, we show only three scatter plots of (x_1, x_2), (x_3, x_4), and (x_5, x_6). The upper panel of Figure 3 gives the standard scatter plots of the three paired variables. Since they are all independently generated, no interesting information can be extracted from these plots.

We can bring in the contour information about the response variable to these static scatter plots through color linkage. However, because of the black–white printing nature of the Encyclopedia, the range information about Y is coded in black and white. In the lower panel of Figure 3, each point in the scatter plots shows the relative intensity of the corresponding magnitude of the response variable Y. The linear structure of Y relative to (x_1, x_2) (the first component) can be easily detected from this plot. There also appears to be some nonlinear structure in the scatterplot of (x_3, x_4) (the second component);

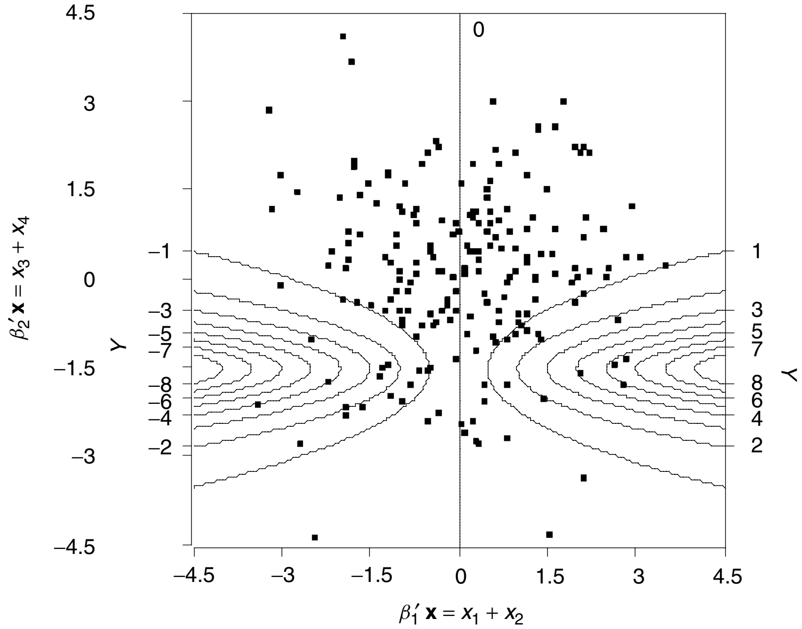


Figure 2 Scatter plot of $\beta_1'\mathbf{x}$ and $\beta_2'\mathbf{x}$ with contours of Y from (2)

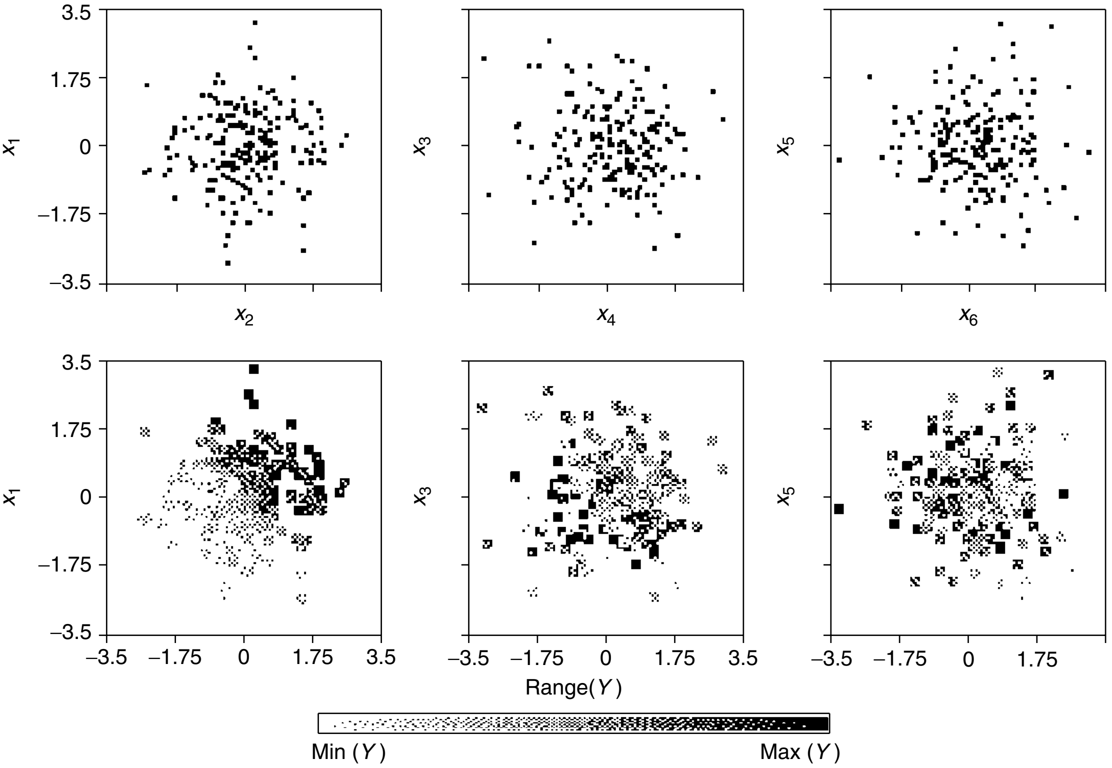


Figure 3 Upper panel: Standard scatter plots of (x_1, x_2), (x_3, x_4), and (x_5, x_6). Lower panel: Same sets of scatter plots as in upper panel except that each point is coded with the relative intensity showing the corresponding value of the response variable Y

but it is not as visible as the first component. No interesting pattern can be identified from plots related to (x_5, x_6) as one would expect.

Principal Component Analysis (PCA) and Multiple Linear Regression (MLR)

For our regression problem, the matrix map of the raw data matrix (Y, **x**) is plotted as Figure 4(a). In a matrix map, each numerical value in a matrix is represented by a color dot (gray shade). Owing to lack of color printing, ranges for all the variables, (Y, **x**) are linearly scaled to [0, 1] and coded in a white to black-gray spectrum. Please see Chen et al. [3] for an introduction to matrix map visualization.

Even for regression problems, **principal component analysis** (PCA) is often used to reduce the dimensionality of **x**. This is not going to work for our example since all the input variables are independently generated. A PCA analysis utilizes only the variation information contained in the input variables. No information about the response variable Y is taken into consideration in constructing the **eigenvalue** decomposition of the covariance matrix of **x** (*see* **Correlation and Covariance Matrices**). The sample covariance matrix for the six input variables in the example is also plotted as a matrix map in Figure 4(b). Since these variables are generated independently with equal variances, no structure with interesting pattern is anticipated from this map. The PCA analysis will produce six principal components with nearly equal eigenvalues. Thus, no reduction of dimension can be achieved from the performance of a PCA analysis on this example. What the SIR analysis does can be considered as a weighted PCA analysis that takes the information of response variable Y into account.

Multiple linear regression (MLR) analysis, on the other hand, can pick up some partial information from the two-component model in (2). MLR studies

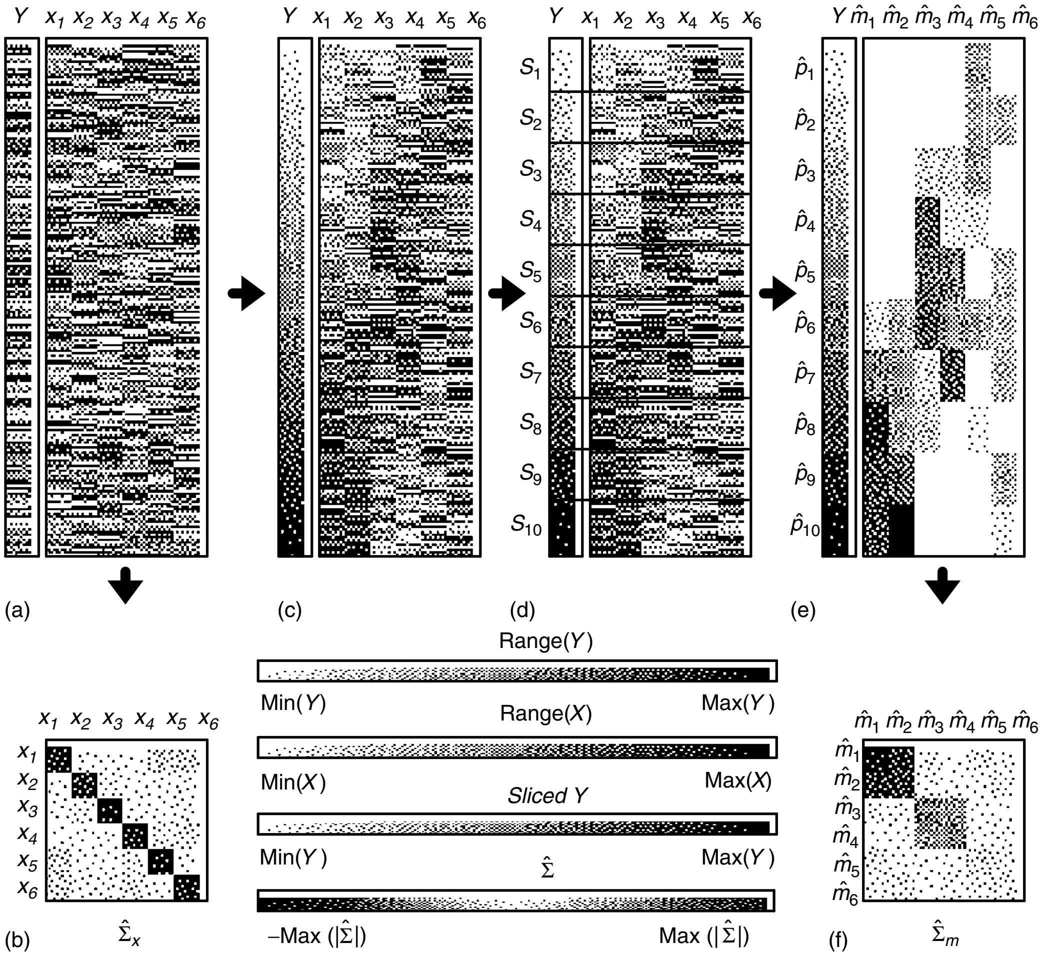


Figure 4 Matrix map of the raw data matrix (Y, $\mathbf{x}$) with a PCA analysis and the SIR algorithm. (a) Original (unsorted) matrix map. (b) Sample covariance matrix of $\mathbf{x}$ in (a), $\hat{\Sigma}_{\mathbf{x}}$. (c) Sorted (by rank of Y) map. (d) Sliced sorted map. (e) Map for sliced mean matrix, $\hat{m}$. (f) Sample covariance matrix of $\hat{m}$, $\hat{\Sigma}_m$

the linear relationship of a linear combination of the input variables $\mathbf{x}$ to the response variable Y. For our example, MLR will identify the linear relationship of Y to the first component $\beta_1'\mathbf{x}$, but not the nonlinear structure of the second component $\beta_2'\mathbf{x}$ on Y.

Implementation and Theoretical Foundation of SIR

Inverse Regression

Conventional functional-approximation and curve-smoothing methods regress Y against $\mathbf{x}$ (forward regression, $\mathrm{E}(Y|\mathbf{x})$). The contour plot in Figure 2 and gray-shaded scatter plots in Figure 3 give the hint to the basic concept of SIR; namely, to reverse the role of $\mathbf{x}$ and Y as in the general forward regression setup. We treat Y as if it were the independent variable and treat $\mathbf{x}$ as if it were the dependent variable. SIR estimates the e.d.r. directions based on inverse regression. The inverse regression curve $\eta(y) = \mathrm{E}(\mathrm{x}|Y = y)$ is composed of p simple regressions, $\mathrm{E}(x_j|y)$, $j = 1, \ldots, p$. Thus, one essentially deals with p one-dimension to one-dimension regression problems, rather than a high-dimensional forward regression problem. Instead of asking the question 'given $\mathbf{x} = \mathbf{x}_o$, what value will Y take'? in the forward regression framework, SIR rephrase the problem as 'given $Y = y$, what values will $\mathbf{x}$ take'? Instead of local smoothing, SIR intends to gain global insight on how Y changes as $\mathbf{x}$ changes by studying the reverse – how does the associated $\mathbf{x}$ region vary as Y varies.

The SIR Algorithm

Following are the steps in conducting the SIR analysis on a random sample $(Y_i, x_i), i = 1, \ldots, n$. Figure 4(a) gives the matrix map of our example with $n = 200$ and $p = 6$. No interesting pattern is expected from Figure 4(a) because the observations are listed in a random manner.

1. Sort the n observations (Y_i, x_i), $i = 1, \ldots, n$ according to the magnitudes of Y_i's. For our example, Figure 4(c) shows a smoothed spectrum of the ranks of Y_is with corresponding linear relationship of (x_1, x_2) and nonlinear structure of (x_3, x_4). The sorted (x_5, x_6) do not carry information on the Y_is.
2. Divide the range of Y into H slices S_h, for $h = 1, \ldots, H$. Let $\hat{p}_h$ $(= 0.1$ in this example) be the proportion of Y_is falling into the hth slice. $H = 10$ slices are used in our example, yielding 20 observations per slice.
3. Compute the sample mean of the $\mathbf{x}_i$s for each slice, $\hat{m}_h = (n\hat{p}_h)^{-1} \sum_{Y_i \in S_h} \mathbf{x}_i$, and form the weighted covariance matrix, $\hat{\Sigma}_m = \sum_{h=1}^{H} \hat{p}_h$ $(\hat{m}_h - \bar{\mathbf{x}})(\hat{m}_h - \bar{\mathbf{x}})$, where $\bar{\mathbf{x}}$ is the sample mean of all x_is. Figure 4(e) gives the matrix map of the corresponding slice means. The linear and nonlinear structures of the Y_i's on (x_1, x_2, x_3, x_4) are even more apparent now.
4. Estimate the covariance matrix of $\mathbf{x}$ with

$$\hat{\Sigma}_{\mathbf{x}} = n^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}}). \quad (4)$$

Find the SIR directions by conducting the eigenvalue decomposition of $\hat{\Sigma}_m$ with respect to $\hat{\Sigma}_{\mathbf{x}}$: for $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$, solve

$$\hat{\Sigma}_m b_j = \hat{\lambda}_j \hat{\Sigma}_{\mathbf{x}} \hat{b}_j, \quad j = 1, \ldots, p. \quad (5)$$

The weighted covariance matrix $\hat{\Sigma}_m$ in Figure 4(f) shows a strong two-component structure compared to that of the sample covariance matrix $\hat{\Sigma}_{\mathbf{x}}$ in Figure 4(b).

5. The ith eigenvector $\boldsymbol{b}_i$ is called the ith *SIR direction*. The first few (two for the example) SIR directions can be used for dimension reduction.

Table 1 The first two eigenvectors (with standard deviations and ratios) and eigenvalues with P values of SIR for model (2)

First vector	(−0.72 −0.68 −0.01 −0.01 0.07 0.01)
S.D.	(0.04 0.04 0.04 0.04 0.04 0.04)
Ratio	(−17.3 −16.1 −0.2 −0.1 1.9 0.3)
Second vector	(−0.01 0.08 −0.67 −0.72 0.05 −0.02)
S.D.	(0.09 0.10 0.09 0.09 0.09 0.09)
Ratio	(−0.2 0.9 −7.6 −7.7 0.5 −0.2)
Eigenvalues	(0.76 0.38 0.06 0.03 0.02 0.02)
P values	(0.0 3.8E-7 0.62 NA NA NA)

The standard SIR output of the example is summarized in Table 1 along with graphical comparisons of Y against two true e.d.r. directions with two estimated directions illustrated in Figure 5.

Some Heuristics

Let us take another look at the scatter plot of $\beta_1'\mathbf{x}$ and $\beta_2'\mathbf{x}$ along with the contours of Y in Figure 2. When the range of Y is divided into H slices, it is possible to use either equal number of observations per slice or equal length of interval per slice. We took the former option here, yielding 20 observations per slice. This is illustrated in Figure 6(a), with the contour lines drawn from the generated Y_is. The mean of the 20 observations contained in a slice is marked by a square with a number indicating which slice it comes from. The location variation of these sliced means along these two components suggests that the slicing-mean step of SIR actually is exploiting the relationship structure of Y against the two correct components of $\mathbf{x}$. This further leads to the well-structured sliced mean matrix shown in Figure 4(e) and the two-component weighted covariance matrix $\hat{\Sigma}_m$ shown in Figure 4(f).

Theoretical Foundation of SIR

The computation of SIR is simple and straightforward. The first three steps of the SIR algorithm

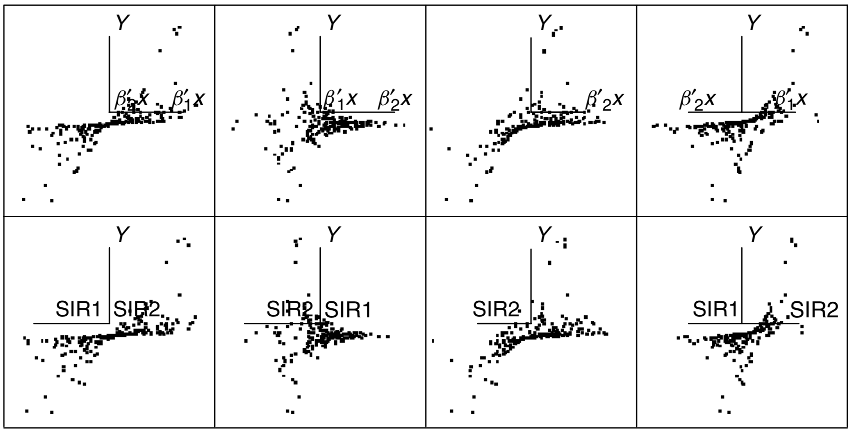


Figure 5 True view (upper panel) and SIR view (lower panel) of model (2)

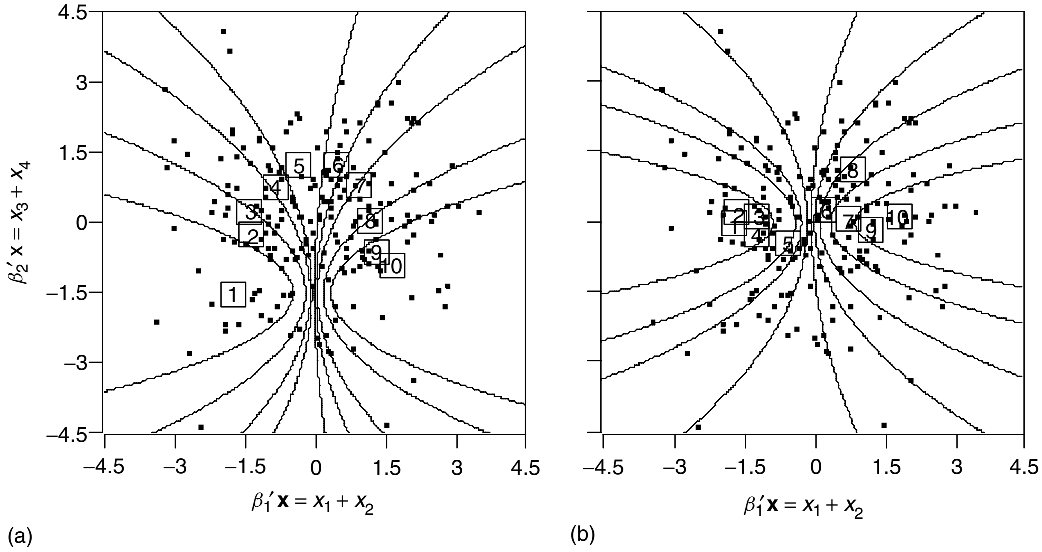


Figure 6 The scatter plot of $\beta_1'\mathbf{x}$ and $\beta_2'\mathbf{x}$ with the contours of Y_is. (a) Y_is generated from model (2). (b) Y_is generated from model (4)

produce a crude estimate of the inverse regression curve $\eta(y) = \mathrm{E}(\mathbf{x}|Y = y)$ through step functions from $\hat{m}_h, h = 1, \ldots, H$. The SIR theorem (Theorem 3.1, [12]) states that under the dimension reduction assumption in model (1), the centered inverse regression curve $\eta(y) - E(\mathbf{x})$ is contained in the linear subspace spanned by $\Sigma_{\mathbf{x}}\beta_i, i = 1, \ldots, k$, provided a linear design condition (Condition 3.1, [12]) on the distribution of $\mathbf{x}$ holds. When this is the case, the covariance matrix of $\eta(Y)$ can be written as a linear combination of $\Sigma_{\mathbf{x}}\beta_i\beta_i'\Sigma_{\mathbf{x}}, i = 1, \ldots, k$. Thus, any eigenvector b_i with nonzero eigenvalue λ_i from the eigenvalue decomposition

$$\mathrm{cov}[\eta(Y)]b_i = \lambda_i \Sigma_{\mathbf{x}} b_i \qquad (6)$$

must fall into the e.d.r. space. Now, because the covariance matrix of the slice average, $\hat{\Sigma}_m$, gives an estimate of $\mathrm{cov}[\eta(Y)]$, the fourth step of the SIR algorithm is just a sample version of (6). It is noteworthy that more sophisticated nonparametric regression methods, such as kernel, nearest neighbor, or smoothing splines can be used to yield a better estimate of the inverse regression curve.

On the basis of the theorem, SIR estimates have been shown to be root-n consistent. They are not sensitive to the number of slices used. Significance tests are available for determining the dimensionality. Further discussions on the theoretical foundation of SIR can be found in Brillinger [1], Chen and Li [2], Cook and Weisberg [5], Cook and Wetzel [6], Duan and Li [7], Hall and Li [9], Hsing and Carroll [10], Li [12, 13], Li and Duan [16], Schott [17], Zhu and Fang [18], and Zhu and Ng [19].

Extensions

Regression Graphics and Graphical Regression

One possible problem, but not necessary a disadvantage, for SIR is that it does not attempt to directly formulate (estimate) the function form of g in model (1). Instead, advocates of SIR argue that users can gain better insights about the relationship structure after visualizing how the response variable Y is associated with reduced input variables. This data analysis strategy is a reversal of the standard practice, which relies on model specification. In a high-dimensional situation, without informative graphical input, formal model specification is seldom efficient. See Cook [4] for more detailed discussion on graphical regression.

Limitations and Generalizations of SIR

SIR successfully finds the two true directions of model (2). But, sometimes, it may not work out as well as expected. To investigate the reason behind, let us change (2) to the following:

$$\begin{aligned} y &= g(\beta_1'\mathbf{x}, \beta_2'\mathbf{x}, \varepsilon) \\ &= \frac{\beta_1'\mathbf{x}}{0.5 + (\beta_2'\mathbf{x})^2} + 0 \cdot \varepsilon. \qquad (7) \end{aligned}$$

We generate 200 Y_is according to (7) while keeping the same set of input variables $\mathbf{x}$ used earlier. The same scatter plot of $\beta_1'\mathbf{x}$ and $\beta_2'\mathbf{x}$ is overlaid with the contours of the new Y_is in Figure 6(b). The symmetric center of each contour is now shifted up to the horizontal axis, resulting in a symmetric contour structure along the second component $\beta_2'\mathbf{x}$. This symmetrical pattern causes the slice means to spread only along the first component and SIR fails in identifying the second component. However, SIR can still find the first component.

The above argument holds for any symmetric function form and the inverse regression curve may not span the entire e.d.r. space. One possible remedy to this problem is to use statistics other than the mean in each slice. For example, the covariance matrix from each slice can be computed and compared with each other. From the contour plot in Figure 6(b), we see that the magnitude of variances within each slice does vary along the second direction. This suggests that slicing the covariance matrix may be able to help. Unfortunately, this second moment–based strategy is not as effective as the first moment–based SIR in finding the first component because the slice variances do not change much along this direction. This interesting phenomenon suggests a possible hybrid technique. That is to combine the directions identified by the first moment SIR and second moment SIR in order to form the complete e.d.r. space. There are several variants of SIR-related dimension reduction strategy such as SAVE ([5]) and SIRII ([12]). These procedures are related to the method of principal Hessian directions ([14]).

References

[1] Brillinger, D.R. (1991). Discussion of "Sliced inverse regression for dimension reduction", *Journal of the American Statistical Association* **86**, 333.

[2] Chen, C.H. & Li, K.C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica* **8**, 289–316.

[3] Chen, C.H., Hwu, H.G., Jang, W.J., Kao, C.H., Tien, Y.J., Tzeng, S. & Wu, H.M. (2004). Matrix visualization and information mining, *Proceedings in Computational Statistics 2004 (Compstat 2004)*, Physika Verlag, Heidelberg.

[4] Cook, R.D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*, Wiley, New York.

[5] Cook, R.D. & Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction", *Journal of the American Statistical Association* **86**, 328–333.

[6] Cook, R.D. & Wetzel, N. (1994). Exploring regression structure with graphics, (with discussion), *Test* **2**, 33–100.

[7] Duan, N. & Li, K.C. (1991). Slicing regression: a link-free regression method, *The Annals of Statistics* **19**, 505–530.

[8] Friedman, J. & Stuetzle, W. (1981). Projection pursuit regression, *Journal of the American Statistical Association* **76**, 817–823.

[9] Hall, P. & Li, K.C. (1993). On almost linearity of low dimensional projection from high dimensional data, *Annals of Statistics* **21**, 867–889.

[10] Hsing, T. & Carroll, R.J. (1992). An asymptotic theory for sliced inverse regression, *Annals of Statistics* **20**, 1040–1061.

[11] Huber, P. (1985). Projection pursuit, with discussion, *Annals of Statistics* **13**, 435–526.

[12] Li, K.C. (1991). Sliced inverse regression for dimension reduction, with discussion, *Journal of the American Statistical Association* **86**, 316–412.

[13] Li, K.C. (1992a). Uncertainty analysis for mathematical models with SIR, in *Probability and Statistics*, Z.P. Jiang, S.H. Yan, P. Cheng & R. Wu, eds, World Scientific, Singapore, pp. 138–162.

[14] Li, K.C. (1992b). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma, *Journal of the American Statistical Association* **87**, 1025–1039.

[15] Li, K.C. (1997). Sliced inverse regression, in *Encyclopedia of Statistical Sciences*, Vol. 1, (update), S. Kotz, C. Read & D. Banks, eds, John Wiley, New York, pp. 497–499.

[16] Li, K.C. & Duan, N. (1989). Regression analysis under link violation, *Annals of Statistics* **17**, 1009–1052.

[17] Schott, J.R. (1994). Determining the dimensionality in sliced inverse regression, *Journal of the American Statistical Association* **89**, 141–148.

[18] Zhu, L.X. & Fang, K.T. (1996). Asymptotics for kernel estimate of sliced inverse regression, *Annals of Statistics* **24**, 1053–1068.

[19] Zhu, L.X. & Ng, K.W. (1995). Asymptotics of sliced inverse regression, *Statistica Sinica* **5**, 727–736.

CHUN-HOUH CHEN

Snedecor, George Waddell

David C. Howell

Volume 4, pp. 1863–1864

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Snedecor, George Waddell

Born: October 20, 1881, in Memphis, USA.
Died: February 20, 1974, in Amherst, USA.

George Snedecor was born in Tennessee in 1882 and graduated from the University of Alabama in 1905 with a B.S. degree in mathematics and physics. He then moved to Michigan State, where he obtained a Master's degree in physics. In 1913, he accepted a position at Iowa State to teach algebra, and soon persuaded his department to allow him to introduce courses in the relatively new field of statistics. He remained at Iowa State until his retirement.

While George Snedecor did not make many major contributions to the theory of statistics, he was one of the field's most influential pioneers. In 1924, he paired with Henry Wallace, later to become vice president under Roosevelt, to jointly publish a manual on machine methods for computational and statistical methods [4]. Three years later, Iowa State formed the Mathematical Statistics Service headed by Snedecor and A. E. Brandt. Then, in 1935, the Iowa Agriculture Experiment Station formed a Statistical Section that later became the Department of Statistics at Iowa State. This was the first department of statistics in the United States. Again, George Snedecor was its head.

Beginning in the early 1930s, Snedecor invited eminent statisticians from Europe to spend summers at Iowa. **R. A. Fisher** was one of the first to come, and he came for several years. His influence on Snedecor's interest in experimental design and the **analysis of variance** was significant, and in 1937, Snedecor published the first of seven editions of his famous *Statistical Methods* [3]. (This work was later written jointly by Snedecor and **W. G. Cochran**, and is still in press.)

As is well-known, R. A. Fisher developed the analysis of variance, and in his 1924 book, included an early table for evaluating the test statistic. In 1934, Snedecor published his own table of $F = \widehat{\sigma}^2_{\text{Treatment}}/\widehat{\sigma}^2_{\text{Error}}$, which derives directly from the calculations of the analysis of variance [2]. He named this statistic F in honor of Fisher, and it retains that name to this day [1].

Snedecor's department included many eminent and influential statisticians of the time, among whom were **Gertrude Cox** and William Cochran, both of whom he personally recruited to Iowa. He was president of the American Statistical Association in 1948, and made an Honorary Fellow of the British Royal Statistical Society (1954). Like many other major figures in statistics at the time (e.g., Egon Pearson and Gertrude Cox), he apparently never earned a Ph. D. However, he was awarded honorary D. Sc. degrees from both North Carolina State University (1956) and Iowa State University (1958). In 1976, the Committee of Presidents of Statistical Societies established the George W. Snedecor Award. This honors individuals who were instrumental in the development of statistical theory in biometry.

References

[1] Kirk, R.E. (2003). Experimental design, in J. Schinka & W.F. Velicer, eds, *Handbook of Psychology*, *Research Methods in Psychology*, Vol. 2. (I.B. Weiner, Editor-in-Chief). New York, Wiley, Available at `http://media.wiley.com/product_data/excerpt/31/04713851/0471385131.pdf`.

[2] Snedecor, G.W. (1934). *Analysis of Variance and Covariance*, Iowa State University Press, Ames.

[3] Snedecor, G.W. (1937). *Statistical Methods*, Iowa State University Press, Ames.

[4] Wallace, H.A. &. Snedecor, G.W. (1925). *Correlation and Machine Calculation*, Iowa State College Press, Ames.

DAVID C. HOWELL

Social Interaction Models

John K. Hewitt

Volume 4, pp. 1864–1865

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Social Interaction Models

Social interaction is modeled in behavior genetics as the influence of one individual's **phenotype** on another, usually within the same family [1, 2]. The principle feature of social interaction in such models is that the phenotype of a given individual (P_1) has an additional source of influence besides the more usually considered additive genetic (A_1), shared family environmental (C_1), and nonshared environmental (E_1) influences (*see* **ACE Model**). This additional source of influence is the phenotype of another individual (P_2), who is often a sibling and, in the traditional twin study, a monozygotic or dizygotic twin.

Thus, the linear path model changes from: $P_1 = aA_1 + cC_1 + eE_1$ to $P_1 = sP_2 + aA_1 + cC_1 + eE_1$. Of course, in a sibling pair, there is usually no reason to suppose asymmetry of influence (although allowance can be made for such asymmetry in the case of, for example, parents and offspring, or siblings of different ages, or of different sex.) In the symmetrical case, $P_2 = sP_1 + aA_2 + cC_2 + eE_2$. If s is positive, then we would be modeling *cooperative social interactions or imitation*; if s is negative, then we would be modeling *competitive social interactions or contrast.*

We have

$$\begin{aligned} P_1 &= sP_2 + aA_1 + cC_1 + eE_1 \\ P_2 &= sP_1 + aA_2 + cC_2 + eE_2 \end{aligned} \qquad (1)$$

or

$$\begin{aligned} P_1 - sP_2 &= aA_1 + cC_1 + eE_1 \\ P_2 - sP_1 &= aA_2 + cC_2 + eE_2 \end{aligned} \qquad (2)$$

or

$$(\mathbf{I} - \mathbf{B})\mathbf{P} = a\mathbf{A} + c\mathbf{C} + e\mathbf{E}, \qquad (3)$$

where **I** is a 2 by 2 identity matrix, **B** is a 2 by 2 matrix with zeros on the leading diagonal and s in each of the off diagonal positions, and **P**, **A**, **C**, and **E** are each 2 by 1 vectors of the corresponding phenotypes, and genetic and environmental influences.

With a little rearrangement, we see that

$$\mathbf{P} = (\mathbf{I} - \mathbf{B})^{-1}(a\mathbf{A} + c\mathbf{C} + e\mathbf{E}), \qquad (4)$$

and then the expected covariance matrix (*see* **Covariance/variance/correlation**) of the phenotypes is the usual expectation for the given type of pair of relatives, premultiplied by $(\mathbf{I} - \mathbf{B})^{-1}$ and postmultiplied by its transpose.

As a consequence, both the expected phenotypic variances of individuals, and the expected covariances between pairs of individuals, are changed by social interaction *and the extent of those changes depends on the a priori correlation of the interacting pair.* Thus, in the presence of cooperative social interaction, the phenotypic variance will be increased, but more so for monozygotic twin pairs than dizygotic twin pairs or full siblings and, in turn, more so for these individuals than for pairs of adopted (not biologically related) siblings. The family covariance will also be increased in the same general pattern, but the proportional increase will be greater for the less closely related pairs. Thus, the overall effect of cooperative social interaction will be similar to that of a shared family environment (which in a sense it is), but will differ in its telltale differential impact on the phenotypic variances of individuals from different types of interacting pairs. For categorical traits, such as disease diagnoses, different **prevalences** in different types of interacting pairs may be detected [2].

In the presence of competitive social interactions, the consequences depend on the details of the parameters of the model but the signature characteristic is that, in addition to phenotypic variances differing for pairs of different relationship, typically in the opposite pattern than for cooperative interactions, the pattern of pair resemblance suggests nonadditive genetic influences. The model predicts negative family correlations under strong competition. These have sometimes been reported for such traits as toddler temperament [4] or childhood hyperactivity, but they may result from the rater contrasting one child with another, especially when a single reporter, such as a parent, is rating both members of a pair of siblings [5].

Although social interactions are of considerable theoretical interest, convincing evidence for the effects of such interactions is hard to find in the behavior genetic literature. The prediction that phenotypic variances are dependent on what kind of pair is being considered is implicitly tested whenever a standard genetic and environmental model is fitted to empirical data, and it has rarely been found wanting. ‘For IQ, educational attainment, psychometric assessments of personality, social attitudes, body mass index, heart rate reactivity, and so on, the behavior genetic literature is replete with evidence for the *absence* of the effects of social interactions.’ ([3], p. 209.) Thus, the take home message from behavior genetics may be that social interactions, at least those occurring within the twin or sibling context, appear to have surprisingly little effect on individual differences in the traits routinely studied by psychologists.

References

[1] Carey, G. (1986). Sibling imitation and contrast effects, *Behavior Genetics* **16**, 319–341.

[2] Carey, G. (1992). Twin imitation for antisocial behavior: implications for genetic and family environment research, *Journal of Abnormal Psychology* **101**, 18–25.

[3] Neale, M.C. & Cardon, L. (1992). *Methodology for Genetic Studies of Twins and Families*, Kluwer Academic Publishers, Boston.

[4] Saudino, K.J., Cherny, S.S. & Plomin, R. (2000). Parent ratings of temperament in twins: explaining the ‘too low’ DZ correlations, *Twin Research* **3**, 224–233.

[5] Simonoff, E., Pickles, A., Hervas, A., Silberg, J., Rutter, M. & Eaves, L.J. (1998). Genetic influences on childhood hyperactivity: contrast effects imply parental rating bias, not sibling interaction, *Psychological Medicine* **28**, 825–837.

JOHN K. HEWITT

Social Networks

Stanley Wasserman, Philippa Pattison and Douglas Steinley

Volume 4, pp. 1866–1871

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Social Networks

Network analysis is the interdisciplinary study of social relations and has roots in anthropology, sociology, psychology, and applied mathematics. It conceives of social structure in relational terms, and its most fundamental construct is that of a *social network,* comprising at the most basic level a set of *social actors* and a set of *relational ties* connecting pairs of these actors. A primary assumption is that social actors are interdependent, and that the relational ties among them have important consequences for each social actor as well as for the larger social groupings that they comprise.

The *nodes* or *members* of the network can be groups or organizations as well as people. Network analysis involves a combination of theorizing, model building, and empirical research, including (possibly) sophisticated data analysis. The goal is to study network structure, often analyzed using such concepts as density, centrality, prestige, mutuality, and role. Social network data sets are occasionally multidimensional and/or longitudinal, and they often include information about *actor attribute*s, such as actor age, gender, ethnicity, attitudes, and beliefs.

A basic premise of the social network paradigm is that knowledge about the structure of social relationships enriches explanations based on knowledge about the attributes of the actors alone. Whenever the social context of individual actors under study is relevant, relational information can be gathered and studied. Network analysis goes beyond measurements taken on individuals to analyze data on patterns of relational ties and to examine how the existence and functioning of such ties are constrained by the social networks in which individual actors are embedded. For example, one might measure the relations 'communicate with', 'live near', 'feel hostility toward', and 'go to for social support' on a group of workers. Some network analyses are longitudinal, viewing changing social structure as an outcome of underlying processes. Others link individuals to events (affiliation networks), such as a set of individuals participating in a set of community activities.

Network structure can be studied at many different levels: the dyad, triad, subgroup, or even the entire network. Furthermore, network theories can be postulated at a variety of different levels. Although this multilevel aspect of network analysis allows different structural questions to be posed and studied simultaneously, it usually requires the use of methods that go beyond the standard approach of treating each individual as an independent unit of analysis. This is especially true for studying a *complete* or whole network: a census of a well-defined population of social actors in which all ties, of various types, among all the actors are measured. Such analyses might study structural balance in small groups, transitive flows of information through indirect ties, structural equivalence in organizations, or patterns of relations in a set of organizations.

For example, network analysis allows a researcher to model the interdependencies of organization members. The paradigm provides concepts, theories, and methods to investigate how informal organizational structures intersect with formal bureaucratic structures in the unfolding flow of work-related actions of organizational members and in their evolving sets of knowledge and beliefs. Hence, it has informed many of the topics of organizational behavior, such as leadership, attitudes, work roles, turnover, and computer-supported cooperative work.

Historical Background

Network analysis has developed out of several research traditions, including (a) the birth of sociometry in the 1930s spawned by the work of the psychiatrist Jacob L. Moreno; (b) ethnographic efforts in the 1950s and 1960s to understand migrations from tribal villages to polyglot cities, especially the research of A. R. Radcliffe–Brown; (c) survey research since the 1950s to describe the nature of personal communities, social support, and social mobilization; and (d) archival analysis to understand the structure of interorganizational and international ties. Also noteworthy is the work of Claude Lévi-Strauss, who was the first to introduce formal notions of kinship, thereby leading to a mathematical algebraic theory of relations, and the work of Anatol Rapoport, perhaps the first to propose an elaborate statistical model of relational ties and flow through various nodes.

Highlights of the field include the adoption of sophisticated mathematical models, especially discrete mathematics and graph theory, in the 1940s and 1950s. Concepts such as transitivity, structural equivalence, the strength of weak ties, and centrality arose from network research by James A. Davis, Samuel Leinhardt, Paul Holland, Harrison White, Mark Granovetter, and Linton Freeman in the 1960s and 1970s. Despite the separateness of these many research beginnings, the field grew and was drawn together in the 1970s by formulations in graph theory and advances in computing. Network analysis, as a distinct research endeavor, was born in the early 1970s. Noteworthy in its birth is the pioneering text by Harary, Norman, and Cartwright [4]; the appearance in the late 1970s of network analysis software, much of it arising at the University of California, Irvine; and annual conferences of network analysts, now sponsored by the International Network for Social Network Analysis. These well-known 'Sunbelt' Social Network Conferences now draw as many as 400 international participants. A number of fields, such as organizational science, have experienced rapid growth through the adoption of a network perspective.

Over the years, the social network analytic perspective has been used to gain increased understanding of many diverse phenomena in the social and behavioral sciences, including (taken from [8])

- Occupational mobility
- Urbanization
- World political and economic systems
- Community elite decision making
- Social support
- Community psychology
- Group problem solving
- Diffusion and adoption of information
- Corporate interlocking
- Belief systems
- Social cognition
- Markets
- Sociology of science
- Exchange and power
- Consensus and social influence
- Coalition formation

In addition, it offers the potential to understand many contemporary issues, including (see [1])

- The Internet
- Knowledge and distributed intelligence
- Computer-mediated communication
- Terrorism
- Metabolic systems
- Health, illness, and epidemiology, especially of HIV

Before a discussion of the details of various network research methods, we mention in passing a number of important measurement approaches.

Measurement

Complete Networks

In complete network studies, a census of network ties is taken for all members of a prespecified population of network members. A variety of methods may be used to observe the network ties (e.g., survey, archival, participant observation), and observations may be made on a number of different types of network tie. Studies of complete networks are often appropriate when it is desirable to understand the action of network members in terms of their location in a broader social system (e.g., their centrality in the network, or more generally in terms of their patterns of connections to other network members). Likewise, it may be necessary to observe a complete network when properties of the network as a whole are of interest (e.g., its degree of centralization, fragmentation, or connectedness).

Ego-centered Networks

The size and scope of complete networks generally preclude the study of all the ties and possibly all the nodes in a large, possibly unbounded population. To study such phenomena, researchers often use survey research to study a sample of personal networks (often called *ego-centered* or *local* networks). These smaller networks consist of the set of specified ties that links *focal persons* (or *egos*) at the centers of these networks to a set of close 'associates' or *alters.* Such studies focus on an ego's ties and on ties among ego's alters. Ego-centered networks can include relations such as kinship, weak ties, frequent contact, and provision of emotional or instrumental aid. These relations can be characterized by their variety, content, strength, and structure. Thus, analysts might study network member *composition* (such as the percentage of women providing social or emotional support, for example, or basic actor attributes more generally); network characteristics (e.g., percentage of dyads that are mutual); measures of *relational association* (do strong ties with immediate kin also imply supportive relationships?); and network structure (how densely knit are various relations? do actors cluster in any meaningful way?).

Snowball Sampling and Link Tracing Studies

Another possibility, to study large networks, is simply to sample nodes or ties. Sampling theory for networks contains a small number of important results (e.g., estimation of subgraphs or subcomponents; many originated with Ove Frank) as well as a number of unique techniques or strategies such as snowball sampling, in which a number of nodes are sampled, then those linked to this original sample are sampled, and so forth, in a multistage process. In a link-tracing sampling design, emphasis is on the links rather than the actors – a set of social links is followed from one respondent to another. For hard-to-access or hidden populations, such designs are considered the most practical way to obtain a sample of nodes. Related are recent techniques that obtain samples, and based on knowledge of certain characteristics in the population and the structure of the sampled network, make inferences about the population as a whole (see [5, 6]).

Cognitive Social Structures

Social network studies of *social cognition* investigate how individual network actors perceive the ties of others and the social structures in which they are contained. Such studies often involve the measurement of multiple perspectives on a network, for instance, by observing each network member's view of who is tied to whom in the network. David Krackhardt referred to the resulting data arrays as *cognitive social structures.* Research has focused on clarifying the various ways in which social cognition may be related to network locations: (a) People's positions in social structures may determine the specific information to which they are exposed, and hence, their perception; (b) structural position may be related to characteristic patterns of social interactions; (c) structural position may frame social cognitions by affecting people's perceptions of their social locales.

Methods

Social network analysts have developed methods and tools for the study of relational data. The techniques include graph theoretic methods developed by mathematicians (many of which involve counting various types of subgraphs); algebraic models popularized by mathematical sociologists and psychologists; and statistical models, which include the social relations model from social psychology and the recent family of random graphs first introduced into the network literature by Ove Frank and David Strauss. Software packages to fit these models are widely available.

Exciting recent developments in network methods have occurred in the statistical arena and reflect the increasing theoretical focus in the social and behavioral sciences on the interdependence of social actors in dynamic, network-based social settings. Therefore, a growing importance has been accorded the problem of constructing theoretically and empirically plausible parametric models for structural network phenomena and their changes over time. Substantial advances in statistical computing are now allowing researchers to more easily fit these more complex models to data.

Some Notation

In the simplest case, network studies involve a single type of directed or nondirected tie measured for all pairs of a node set $N = \{1, 2, \ldots, n\}$ of individual actors. The observed tie linking node i to node j ($i, j \in N$) can be denoted by x_{ij} and is often defined to take the value 1 if the tie is observed to be present and 0 otherwise. The network may be either *directed* (in which case x_{ij} and x_{ji} are distinguished and may take different values) or *nondirected* (in which case x_{ij} and x_{ji} are not distinguished and are necessarily equal in value). Other cases of interest include the following:

1. *Valued* networks, where x_{ij} takes values in the set $\{0, 1, \ldots, C - 1\}$.
2. *Time-dependent* networks, where x_{ijt} represents the tie from node i to node j at time t.
3. *Multiple relational* or *multivariate* networks, where x_{ijk} represents the tie of type k from node i to node j (with $k \in R = \{1, 2, \ldots, r\}$, a fixed set of *types* of tie).

In most of the statistical literature on network methods, the set N is regarded as fixed and the network ties are assumed to be random. In this case, the tie linking node i to node j may be denoted by the random variable X_{ij} and the $n \times n$ array $\boldsymbol{X} = [X_{ij}]$

of random variables can be regarded as the adjacency matrix of a *random (directed) graph* on N.

Graph Theoretic Techniques

Graph theory has played a critical role in the development of network analysis. Graph theoretical techniques underlie approaches to understanding cohesiveness, connectedness, and fragmentation in networks. Fundamental measures of a network include its *density* (the proportion of possible ties in the network that are actually observed) and the degree sequence of its nodes. In a nondirected network, the *degree* d_i of node i is the number of distinct nodes to which node i is connected. Methods for characterizing and identifying cohesive subsets in a network have depended on the notion of a *clique* (a subgraph of network nodes, every pair of which is connected) as well as on a variety of generalizations (including k-clique, k-plex, k-core, LS-set, and k-connected subgraph).

Our understanding of connectedness, connectivity, and centralization is also informed by the distribution of path lengths in a network. A *path* of length k from one node i to another node j is defined by a sequence $i = i_1, i_2, \ldots, i_{k+1} = j$ of distinct nodes such that i_h and i_{h+1} are connected by a network tie. If there is no path from i to j of length $n - 1$ or less, then j is not reachable from i and the distance from i to j is said to be infinite; otherwise, the distance from i to j is the length of the shortest path from i to j. A directed network is *strongly connected* if each node is reachable from each other node; it is *weakly connected* if, for every pair of nodes, at least one of the pair is reachable from the other. For nondirected networks, a network is connected if each node is reachable from each other node, and the *connectivity,* κ, is the least number of nodes whose removal results in a disconnected (or trivial) subgraph.

Graphs that contain many cohesive subsets as well as short paths, on average, are often termed *small world* networks, following early work by Stanley Milgram, and more recent work by Duncan Watts. Characterizations of the centrality of each actor in the network are typically based on the actor's degree (*degree* centrality), on the lengths of paths from the actor to all other actors (*closeness* centrality), or on the extent to which the shortest paths between other actors pass through the given actor (*betweenness* centrality). Measures of network *centralization* signify the extent of heterogeneity among actors in these different forms of centrality.

Algebraic Techniques

Closely related to graph theoretic approaches is a collection of algebraic techniques that has been developed to understand social roles and structural regularities in networks. Characterizations of role have developed in terms of mappings on networks, and descriptions of structural regularities have been facilitated by the construction of algebras among labeled network walks. An important proposition about what it means for two actors to have the same social role is embedded in the notion of *structural equivalence:* Two actors are said to be *structurally equivalent* if they are relate to and are related to by every other network actor in exactly the same way (thus, nodes i and j are structurally equivalent if, for all $k \in N, x_{ik} = x_{jk}$ and $x_{ki} = x_{kj}$). Generalizations to automorphic and regular equivalence are based on more general mappings on N and capture the notion that similarly positioned network nodes are related to *similar* others in the same way.

Statistical Techniques

A simple statistical model for a (directed) graph assumes a Bernoulli distribution (*see* **Catalogue of Probability Density Functions**), in which each edge, or tie, is statistically independent of all others and governed by a theoretical probability P_{ij}. In addition to edge independence, simplified versions also assume equal probabilities across ties; other versions allow the probabilities to depend on structural parameters. These distributions often have been used as models for at least 40 years, but are of questionable utility because of the independence assumption.

Dyadic Structure in Networks

Statistical models for social network phenomena have been developed from their edge-independent beginnings in a number of major ways. The p_1 model recognized the theoretical and empirical importance of dyadic structure in social networks, that is, of the interdependence of the variables X_{ij} and X_{ji}.

This class of Bernoulli dyad distributions and their generalization to valued, multivariate, and time-dependent forms gave parametric expression to ideas of reciprocity and exchange in dyads and their development over time. The model assumes that each dyad (X_{ij}, X_{ji}) is independent of every other, resulting in a **log-linear model** that is easily fit. Generalizations of this model are numerous, and include *stochastic block models,* representing hypotheses about the interdependence of social positions and the patterning of network ties; mixed models, such as p_2; and *latent space models* for networks.

Null Models for Networks

The assumption of dyadic independence is questionable. Thus, another series of developments has been motivated by the problem of assessing the degree and nature of departures from simple structural assumptions like dyadic independence. A number of *conditional uniform* random graph distributions were introduced as null models for exploring the structural features of social networks. These distributions, denoted by U|Q, are defined over subsets Q of the state space Ω_n of directed graphs and assign equal probability to each member of Q. The subset Q is usually chosen to have some specified set of properties (e.g., a fixed number of mutual, asymmetric, and null dyads). When Q is equal to Ω_n, the distribution is referred to as the *uniform* (di)graph distribution, and is equivalent to a Bernoulli distribution with homogeneous tie probabilities. Enumeration of the members of Q and simulation of U|Q are often straightforward, although certain cases, such as the distribution that is conditional on the indegree and outdegree of each node i in the network, require more complicated approaches.

A typical application of these distributions is to assess whether the occurrence of certain higher-order (e.g., triadic) features in an observed network is unusual, given the assumption that the data arose from a uniform distribution that is conditional on plausible lower-order (e.g., dyadic) features. This general approach has also been developed for the analysis of multiple networks. The best known example is probably Frank Baker and Larry Hubert's quadratic assignment procedure (QAP) for networks. In this case, the association between two graphs defined on the same set of nodes is assessed using a uniform multigraph distribution that is conditional on the unlabeled graph structure of each constituent graph.

Extradyadic Local Structure in Networks

A significant step in the development of parametric statistical models for social networks was taken by Frank and Strauss [3] with the introduction of the class of *Markov random graphs*, denoted as p^* by later researchers. This class of models permitted the parameterization of extradyadic local structural forms, allowing a more explicit link between some important theoretical propositions and statistical network models. These models are based on the fact that the Hammersley–Clifford theorem provides a general probability distribution for $\boldsymbol{X}$ from a specification of which pairs (X_{ij}, X_{kl}) of tie random variables are conditionally dependent, given the values of all other random variables.

These random graph models permit the parameterization of many important ideas about local structure in univariate social networks, including transitivity, local clustering, degree variability, and centralization. Valued, multiple, and temporal generalizations also lead to parameterizations of substantively interesting multirelational concepts, such as those associated with balance and clusterability, generalized transitivity and exchange, and the strength of weak ties. Pseudomaximum likelihood estimation is easy; **maximum likelihood estimation** is difficult, but not impossible.

Dynamic Models

A significant challenge is to develop models for the emergence of network phenomena, including the evolution of networks and the unfolding of individual actions (e.g., voting, attitude change, decision making) and interpersonal transactions (e.g., patterns of communication or interpersonal exchange) in the context of long-standing relational ties. Early attempts to model the evolution of networks in either discrete or continuous time assumed dyad independence and Markov processes in time. A step towards continuous time **Markov chain** models for network evolution that relaxes the assumption of dyad independence has been taken by Tom Snijders and colleagues. This approach also illustrates the potentially valuable role of simulation techniques for models that make empirically plausible assumptions; clearly, such methods provide a promising focus for future development. Computational models based on simulations are becoming increasingly popular in network analysis; however, the development of associated model evaluation approaches poses a significant challenge.

Current research, including future challenges, such as statistical estimation of complex model parameters, model evaluation, and dynamic statistical models for longitudinal data, can be found in [2]. Applications of the techniques and definitions mentioned here can be found in [7] and [8].

References

[1] Breiger, R., Carley, K. & Pattison, P., eds (2003). *Dynamic Social Network Modeling and Analysis*, The National Academies Press, Washington.

[2] Carrington, P.J., Scott, J. & Wasserman, S., eds (2004). *Models and Methods in Social Network Analysis*, Cambridge University Press, New York.

[3] Frank, O. & Strauss, D. (1986). Markov graphs, *Journal of the American Statistical Association* **81**, 832–842.

[4] Harary, F., Norman, D. & Cartwright, D. (1965). *Structural Models for Directed Graphs*, Free Press, New York.

[5] Killworth, P.D., McCarty, C., Bernard, H.R., Shelley, G.A. & Johnsen, E.C. (1998). Estimation of seroprevalence, rape, and homelessness in the U.S. using a social network approach, *Evaluation Review* **22**, 289–308.

[6] McCarty, C., Killworth, P.D., Bernard, H.R., Johnsen, E.C. & Shelley, G.A. (2001). Comparing two methods for estimating network size, *Human Organization* **60**, 28–39.

[7] Scott, J. (1992). *Social Network Analysis*, Sage Publications, London.

[8] Wasserman, S. & Faust, K. (1994). *Social Network Analysis: Methods and Applications*, Cambridge University Press, New York.

Further Reading

Boyd, J.P. (1990). *Social Semigroups: A Unified Theory of Scaling and Bockmodeling as Applied to Social Networks*, George Mason University Press, Fairfax.

Friedkin, N. (1998). *A Structural Theory of Social Influence*, Cambridge University Press, New York.

Monge, P. & Contractor, N. (2003). *Theories of Communication Networks*, Oxford University Press, New York.

Pattison, P.E. (1993). *Algebraic Models for Social Networks*, Cambridge University Press, New York.

Wasserman, S. & Galaskiewicz, J., eds (1994). *Advances in Social Network Analysis*, Sage Publications, Thousand Oaks.

Watts, D. (1999). *Small Worlds: The Dynamics of Networks Between Order and Randomness*, Princeton University Press, Princeton.

Wellman, B. & Berkowitz, S.D., eds (1997). *Social Structures: A Network Approach*, (updated Edition), JAI, Greenwich.

Stanley Wasserman, Philippa Pattison
and Douglas Steinley

Social Psychology

Charles M. Judd and Dominique Muller

Volume 4, pp. 1871–1874

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Social Psychology

Social psychologists are interested in the ways behavior is affected by the social situations in which people find themselves. Recurring topics of research interest in this subdiscipline include (a) stereotyping, prejudice, and intergroup behavior; (b) interpersonal attraction and close relationships; (c) behavior in small groups and group decision making; (d) social influence, conformity, and social norms; (e) attitudes and affective responses to social stimuli; and (f) dyadic interaction and communication. These topics do not exhaust the list of social psychological interests, but they are representative of the broad array of social behaviors that have attracted research attention.

As an empirical discipline, social psychology has taken pride in its use of systematic observation methods to reach conclusions about the causes, processes, and consequences of social behavior. In this endeavor, it has historically adopted an experimental perspective by simulating social environments in the laboratory and other settings, and gathering data on social behavior in those settings. The impetus for this approach came from the seminal influence of Kurt Lewin (and his students) who showed how important social phenomena can be captured in simulated experimental environments. For example, to study the effects of different leadership styles in small groups, Lewin and colleagues [7] examined the consequences of democratic and autocratic leadership styles in simulated small group interactions.

In this empirical approach, participants are typically randomly assigned to various experimental conditions (*see* **Randomization**) (e.g., a democratic or autocratic leader) and differences in their responses (e.g., observed behaviors, self-reports on questionnaires) are calculated. Accordingly, statistical procedures in experimental social psychology have been dominated by examining and testing mean differences using *t* Tests and analysis of variance procedures. A typical report of experimental results in social psychological journals gives condition means (and standard deviations) and the inferential statistic that shows that the observed mean differences are significantly different from zero. Accordingly, traditional null hypothesis testing has been the dominant approach, with attention given only recently to reporting **effect size estimates** or **confidence intervals** for observed means (and for their differences).

Experimental designs in social psychology have become increasingly complex over the years, incorporating multiple (crossed) experimental factors (*see* **Factorial Designs**), as researchers have hypothesized that the effects of some manipulations depend on other circumstances or events. Accordingly, **analysis of variance** models with multiple factors are routinely used, with considerable attention focused on statistical interactions (*see* **Interaction Effects**): Does the effect of one experimental factor depend upon the level of another? In fact, it is not unusual for social psychological experiments to cross three or more experimental factors, with resulting higher order interactions that researchers attempt to meaningfully interpret. One of our colleagues jokingly suggests that a criterion for publication in the leading social psychological journals is a significant (at least two-way) interaction.

For some independent variables of interest to social psychologists, manipulations are done 'within participants', exposing each participant in turn to multiple levels of a factor, and measuring responses under each level (*see* **Repeated Measures Analysis of Variance**). This has seemed particularly appropriate where the effects of those manipulations are short-lived, and where one can justify assumptions about the lack of **carry-over effects**. Standard analysis of variance courses taught to social psychologists have typically included procedures to analyze data from within-participant designs, which are routinely used in experimental and cognitive psychological research as well. Thus, beyond multifactorial analyses of variance, standard analytic tools have included **repeated measures analysis of variance** models and split-plot designs.

While social psychologists have traditionally been very concerned about the integrity of their experimental manipulations and the need to randomly assign participants to experimental conditions (e.g., [3]), they have historically been much less concerned about sampling issues (*see* **Experimental Design**). In fact, the discipline has occasionally been subjected to substantial criticism for its tendency to rely on undergraduate research participants, recruited from psychology courses where 'research participation' is a course requirement (see [9]). The question raised is about the degree to which results from social psychological experiments can be generalized to a public broader than college undergraduates. From a statistical point of view, the use of convenience samples such as undergraduates who happen to enroll in psychology classes means that standard inferential statistical procedures are difficult to interpret since there is no known population from which one has sampled. Rather than concluding that a statistically significant difference suggests a true difference in some known population, one is left with hypothetical populations and unclear inferences such as 'If I were to repeat this study over and over again with samples randomly chosen from the unknown hypothetical population from which I sampled, I would expect to continue to find a difference.'

A classic interaction perspective in social psychology, again owing its heritage to Kurt Lewin, is that situational influences on social behavior depend on the motives, aspirations, and attributes of the participants: Behavior results from the interaction of the personality of the actor and the characteristics of the situation. Accordingly, many social psychological studies involve at least one measured, rather than manipulated, independent variable, which is typically a variable that characterizes a difference among participants. Because many social psychologists received training exclusively in classic analysis of variance procedures, the analysis of designs, where one or more measured independent variables are crossed with other factors, has routinely been accomplished by dividing the measured variable into discrete categories, by, for example, a median split (*see* **Categorizing Data**), and including this variable in a factorial analysis of variance. This practice leads to a loss of statistical **power** and, occasionally, to biased estimates [8]. Recently, however, because of the pioneering effort of **Jacob Cohen** [4] and others, social psychologists began to appreciate that analysis of variance is a particular instantiation of the **generalized linear model**. Accordingly, both categorical and more continuously measured independent variables (and their interactions) can be readily included in models that are estimated by standard ordinary least squares regression programs. As social psychologists have become more familiar with the wide array of models estimable under the general linear model, they have increasingly strayed from the laboratory, measuring and manipulating independent variables and using longitudinal designs to assess naturally occurring changes in behavior overtime and in different situations.

With these developments, social psychologists have been forced to confront the limitations inherent in their standard analytic toolbox, specifically assumptions concerning the independence of errors or residuals in both standard analysis of variance and ordinary least squares regression procedures (*see* **Regression Models**). Admittedly, they use specific techniques for dealing with nonindependence in repeated measures designs, in which participants are repeatedly measured under different levels of one or more independent variables. But the assumptions underlying **repeated measures analysis of variance** models, specifically the assumption of equal **covariances** among all repeated observations, seems too restrictive for many research designs where data are collected overtime from participants in naturally occurring environments.

In many situations of interest to social psychologists, data are likely to exhibit nested structures, wherein observations come from dyads, small groups, individuals who are located in families, and so forth. These nested structures mean that observations within groupings are likely to be more similar to each other than observations that are from different groupings. For instance, in collecting data on close relationships, it is typical to measure both members of a number of couples, asking, for instance, about levels of marital satisfaction. Unsurprisingly, observations are likely to show resemblances due to couples. Or, in small group research, we may ask for the opinions of group members who come from multiple groups. Again, it is likely that group members agree with each other to some extent. Or, on a larger scale, if we are interested in behavior in larger organizations, we may sample individuals who are each located in different organizations.

Additional complications arise when individuals or dyads or groups are measured overtime, with, perhaps, both the interval of observations and the frequency of observations varying. For instance, a social psychologist might ask each member of a couple to indicate emotional responses each time an interaction between them lasted more than five minutes. What might be of interest here is the consistency of the emotional responses of one member of the couple compared to the other's and the degree to which that consistency varied with other known factors about the couple. Clearly, standard analysis of variance and ordinary least squares regression procedures cannot be used in such situations. Accordingly, social psychologists are increasingly making use of more general models for nested and dependent data, widely known as multilevel modeling procedures (*see* **Linear Multilevel Models**) or random regression models [2, 5]. Although the use of such analytic approaches to data is still not widespread in social psychology, their utility for the analysis of complex data structures involving dependent observations suggests that they will become standard analytic tools in social psychology in the near future.

In addition to these advances in analytic practices, social psychologists have also become more sophisticated in their use of analytic models for dichotomous-dependent variables. The use **of logistic regression** procedures, for instance, is now fairly widespread in the discipline. Such procedures have significantly extended the range of questions that can be asked about dichotomous variables, permitting, for instance, tests of higher order interactions.

Additionally, while the focus has historically been on the assessment of the overall impact of one or more independent variables on the dependent variable, social psychologists have become increasingly interested in process questions such as what is the mediating process by which the impact is produced? Thus, social psychologists have increasingly made use of analytic procedures designed to assess **mediation**. Given this and the long-standing interest in the discipline in statistical interactions, it is no accident that the classic article on procedures for assessing mediation and **moderation** was written by two social psychologists and published in a leading social psychological outlet [1].

All of the procedures discussed up to this point involve the estimation of relationships between variables thought to assess different theoretical constructs such as the effect of a manipulated independent variable on an outcome variable. Additionally, statistical procedures are used in social psychology to support measurement claims, such as that a particular variable successfully measures a construct of theoretical interest [6]. For these measurement claims, both exploratory and, more recently, confirmatory factor analysis (*see* **Factor Analysis: Confirmatory**) procedures are used. Exploratory factor analysis (*see* **Factor Analysis: Exploratory**) is routinely used when one develops a set of self-report questions, designed, for instance, to measure a particular attitude, and one wants to verify that the questions exhibit a pattern of correlations that suggests they have a single underlying factor in common. In this regard, social psychologists are likely to conduct a principal factoring (*see* **Principal Component Analysis**) or components analysis to demonstrate that the first factor or component explains a substantial amount of the variance in all of the items. Occasionally, researchers put together items that they suspect may measure a number of different latent factors (*see* **Latent Class Analysis**; **Latent Variable**) or constructs that they are interested in 'uncovering'. This sort of approach has a long history in intelligence testing, where researchers attempted to uncover the 'true' dimensions of intelligence. However, as that literature suggests, this use of factor analysis is filled with pitfalls. Not surprisingly, if items that tap a given factor are not included, then that factor cannot be 'uncovered'. Additionally, a factor analytic solution is indeterminant, with different rotations yielding different definitions of underlying dimensions.

Recently, confirmatory factor analytic models have become the approach of choice to demonstrate that items measure a hypothesized underlying construct. In this approach, one hypothesizes a latent factor structure that involves one or more factors, and then examines whether the item covariances are consistent with that structure. Additionally, **structural equation modeling** procedures are sometimes used to model both the relationships between the latent constructs and the measured variables, and also the linear structural relations among those constructs. This has the advantage of estimating the relationships among constructs potentially unbiased by measurement error, because those errors of measurement are modeled in the confirmatory factor analysis part of the estimation. The use of structural equation modeling procedures will remain limited in social psychology, however, for they are not efficient at examining interactive and nonlinear predictions.

In summary, social psychologists are abundant users of statistical and data analytic tools. They pride themselves on the fact that theory evaluation in their discipline ultimately rests on gathering data through systematic observation that can be used to either bolster theoretical conjectures or argue against them. As a function of being usually trained in psychology departments, their standard analytic tools have been those taught in experimental design courses. However, social psychologists often collect data that demand other data analytic approaches. Gradually, social psychologists are becoming sophisticated users of these more flexible approaches. In fact, the statistical demands of some of the data routinely collected by social psychologists means that many new developments in statistical tools for behavioral and social scientists are being developed by social psychologists. And it is no accident that in many psychology departments, the quantitative and analytic courses are now being taught by social psychologists, considerably expanding the traditional analysis of variance and experimental design emphases of such courses.

References

[1] Baron, R.M. & Kenny, D.A. (1986). The moderator – mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations, *Journal of Personality and Social Psychology* **51**, 1173–1182.

[2] Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*, Sage Publications, Newbury Park.

[3] Campbell, D.T. & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research, in *Handbook of Research on Teaching*, N.L. Gage, ed., Rand McNally, Chicago, pp. 171–246.

[4] Cohen, J. & Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 2nd Edition, Erlbaum, Hillsdale.

[5] Goldstein, H. (1995). *Multilevel Statistical Models*, Halstead Press, New York.

[6] Judd, C.M. & McClelland, G.H. (1998). Measurement, in *The Handbook of Social Psychology*, Vol. 1, 4th Edition, D.T. Gilbert, S.T. Fiske & G. Lindzey, eds, McGraw-Hill, Boston, pp. 180–232.

[7] Lewin, K., Lippitt, R. & White, R. (1939). Patterns of aggressive behavior in experimental created "social climates.", *Journal of Social Psychology* **10**, 271–299.

[8] Maxwell, S.E. & Delaney, H.D. (1993). Bivariate mediation splits and spurious statistical significance, *Psychological Bulletin* **113**, 181–190.

[9] Sears, D.O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature, *Journal of Personality and Social Psychology* **51**, 515–530.

Charles M. Judd and Dominique Muller

Social Validity

Alan E. Kazdin

Volume 4, pp. 1875–1876

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Social Validity

Social validity is a concept that is used in intervention research in which the focus is on treatment, prevention, rehabilitation, education, and, indeed, any area in which the goal is to produce change in human behavior and adaptive functioning (*see* **Clinical Trials and Intervention Studies**). The concept of 'social' in the term emphasizes the views and perspectives of individuals who are stakeholders, consumers, or recipients of the intervention. Social validity raises three questions in the context of designing, implementing, and evaluating interventions: (a) Are the *goals* of the intervention relevant to everyday life? (b) Are the *intervention procedures* acceptable to consumers and to the community at large? (c) Are the *outcomes* of the intervention important; that is, do the changes make a difference in the everyday lives of individuals to whom the intervention was directed or those who are in contact with them? The focus is on whether consumers of the interventions find the goals, intervention, and outcomes, reasonable, acceptable, and useful.

Background

Social validation grew out of work in *applied behavior analysis*, an area of behavior modification within psychology [1, 2]. Applied behavior analysis draws on operant conditioning, a type of learning elaborated by B. F. Skinner [7] and that focuses on antecedents, behaviors, and consequences. In the late 1950s and early 1960s, the principles, methods, and techniques of operant conditioning, developed in animal and human laboratory research, were extended to problems of treatment, education, and rehabilitation in applied settings. The applications have included a rather broad array of populations (infants, children adolescents, adults), settings (e.g., medical hospitals, psychiatric hospitals, schools [preschool through college], nursing homes), and contexts (e.g., professional sports, business and industry, the armed forces) [3].

Social validation was initially developed by Montrose Wolf [8], a pioneer in applied behavior analysis, to consider the extent to which the intervention was addressing key concerns of individuals in everyday life. In applying interventions to help people, he reasoned, it was invariably important to ensure that the interventions, their goals, and the effects that were attained were in keeping with the interests of individuals affected by them.

An Example

Consider as an example the application of operant conditioning principles to reduce self-injurious behavior in seriously disturbed children. Many children with pervasive developmental disorder or severe mental retardation hit or bite themselves with such frequency and intensity that physical damage can result. Using techniques to alter antecedents (e.g., prompts, cues, and events presented before the behavior occurs) and consequences (e.g., carefully arranged incentives and activities), these behaviors have been reduced and eliminated [6].

As a hypothetical example, assume for a moment that we have several institutionalized children who engage in severe self-injury such as head banging (banging head against a wall or pounding one's head) or biting (biting one's hands or arms sufficiently to draw blood). We wish to intervene to reduce self-injury. The initial social validity question asks if the goals are relevant or important to everyday life. Clearly they are. Children with self-injury cannot function very well even in special settings, often must be restrained, and are kept from a range of interactions because of the risk to themselves and to others. The next question is whether the interventions are acceptable to consumers (e.g., children and their families, professionals who administer the procedures). Aversive procedures (e.g., punishment, physical restraint, or isolation) might be used, but they generally are unacceptable to most professionals and consumers. Fortunately, effective procedures are available that rely on variations of reinforcement and are usually quite acceptable to consumers [3, 6].

Finally, let us say we apply the intervention and the children show a reduction of self-injury. Let us go further and say that before the intervention, the children of this example hit themselves a mean of 100 times per hour, as observed directly in a special classroom of a hospital facility. Let us say further that treatment reduced this to a mean of 50 times. Does the change make a difference in everyday life? To be sure, a 50% reduction is large, but still not likely to improve adjustment and functioning of the individuals in everyday life. Much larger reductions, indeed, elimination of the behaviors, are needed to have clear social impact.

Extensions

The primary application of social validity has evolved into a related concept, 'clinical significance'that focuses on the outcomes of treatment in the context of psychotherapy for children, adolescents, and adults. The key question of clinical significance is the third one of social validity, namely, does treatment make a difference to the lives of those treated? **Clinical trials** of therapy (e.g., for depression, anxiety, marital discord) often compare various treatment conditions or treatment and control conditions. At the end of the study, statistical tests are usually used to determine whether the group differences and whether the changes from pre- to post-treatment are statistically significant. Statistical significance is not intended to reflect important effects in relation to the functioning of individuals. For example, a group of obese individuals (e.g., >90 kilograms overweight) who receive treatment may lose a mean of nine kilograms, and this change could be statistically significant in comparison to a control group that did not receive treatment. Yet, the amount of weight lost is not very important or relevant from the standpoint of clinical significance. Health (morbidity and mortality) and adaptive functioning (e.g., activities in everyday life) are not likely to be materially improved with such small changes (*see* **Effect Size Measures**).

Treatment evaluation increasingly supplements statistical significance with indices of clinical significance to evaluate whether the changes are actually important to the patients or clients, and those with whom they interact (e.g., spouses, coworkers). There are several indices currently in use such as evaluating whether the level of symptoms at the end of treatment falls within a normative range of individuals functioning well in everyday life, whether the condition that served as the basis for treatment (e.g., depression, panic attacks) is no longer present, and whether the changes made by the individual are especially large [4, 5]. Social validity and its related but more focused concept of clinical significance have fostered increased attention to whether treatment outcomes actually help people in everyday life. The concept has not replaced other ways of evaluating treatment, for example, statistical significance, magnitude of change, but has expanded the criteria by which to judge intervention effects.

References

[1] Baer, D.M., Wolf, M.M. & Risley, T.R. (1968). Some current dimensions of applied behavior analysis, *Journal of Applied Behavior Analysis* **1**, 91–97.

[2] Baer, D.M., Wolf, M.M. & Risley, T.R. (1987). Some still-current dimensions of applied behavior analysis, *Journal of Applied Behavior Analysis* **20**, 313–328.

[3] Kazdin, A.E. (2001). *Behavior Modification in Applied Settings*, 6th Edition, Wadsworth, Belmont.

[4] Kazdin, A.E. (2003). *Research Design in Clinical Psychology*, 4th Edition, Allyn & Bacon, Needham Heights.

[5] Kendall, P.C., ed. (1999). Special section: clinical significance. *Journal of Consulting and Clinical Psychology* **67**, 283–339.

[6] Repp, A.C. & Singh, N.N., eds, (1990). *Perspectives on the Use of Nonaversive and Aversive Interventions for Persons with Developmental Disabilities*, Sycamore Publishing, Sycamore.

[7] Skinner, B.F. (1938). *The Behavior of Organisms: An Experimental Analysis*, Appleton-Century, New York.

[8] Wolf, M.M. (1978). Social validity: the case for subjective measurement, or how applied behavior analysis is finding its heart, *Journal of Applied Behavior Analysis* **11**, 203–214.

Alan E. Kazdin

Software for Behavioral Genetics

MICHAEL C. NEALE

Volume 4, pp. 1876–1880

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Software for Behavioral Genetics

Historical Background

The term software was coined in the 1950s by the eminent statistician **John Tukey** (1915–2000). It usually refers to the program and algorithms used to control the electronic machinery (hardware) of a computer, and may include the documentation. Typically, software consists of source code, which is then compiled into machine-executable code which the end user applies to a specific problem. This general scenario applies to software for genetically informative studies, and might be considered to have existed before Professor Tukey invented the term in the mid twentieth Century. Algorithms are at the heart of software, and this term dates back to the ninth century Iranian mathematician, Al-Khawarizmi. Although formal analysis of data collected from twins did not begin until the 1920s, it was, nevertheless, algorithmic in form. A heuristic estimate of **heritability**, such as twice the difference between the MZ and the DZ correlations, may be implemented using mental arithmetic, the back of an envelope, or on a supercomputer. In all cases the algorithm constitutes software; it is only the hardware that differs.

Software for Model-fitting

Much current behavior genetic analysis is built upon the statistical framework of **maximum likelihood**, attributed to **Ronald Fisher** [4]. As its name suggests, maximum likelihood requires an algorithm for *optimization*, of which there are many: some general and some specific to particular applications. All such methods use *input data* whose likelihood is computed under a particular statistical model. The values of the parameters of this model are not known, but it is often possible to obtain the set of values that maximize the likelihood. These *maximum likelihood estimates* have two especially desirable statistical properties; they are asymptotically unbiased, and have minimum variance of all asymptotically unbiased estimates. Therefore, in the analysis of both genetic linkage (*see* **Linkage Analysis**) using genetic markers, and of **twin studies** to estimate variance components, there was motivation to pursue these more complex methods. This section focuses on twin studies and their extensions.

Before the advent of high-speed computers, maximum likelihood estimation would typically involve: (a) writing out the formula for the likelihood; (b) finding the first and second derivatives of this function with respect to the parameters of the model; and (c) solving the (often nonlinear) simultaneous equations to find those values of the parameters that maximize the likelihood, that is, where the first derivatives are zero and the second derivatives are negative. The first of these steps is often relatively simple, as it typically involves writing out the formula for the probability density function (pdf) (*see* **Catalogue of Probability Density Functions**) of the parameters of the model. In many cases, however, the second and third steps can prove to be challenging or intractable. Therefore, the past 25 years has seen the advent of software designed to estimate parameters under increasingly general conditions.

Early applications of software for numerical optimization (*see* **Optimization Methods**) to behavior genetic data primarily consisted of purpose-built computer programs which were usually written in the high-level language FORTRAN, originally developed in the 1950s by John Backus. From the 1960s to the 1980s this was very much the language of choice, mainly because a large library of numerical algorithms had been developed with it. The availability of these libraries saved behavior geneticists from having to write quite complex code for optimization themselves. Two widely used libraries were MINUIT from the (Centre Européen de Recherche Nucléaire) (CERN) and certain routines from the E04 library of the Numerical Algorithms group (NAg). The latter were developed by Professor Murray and colleagues in the Systems Optimization Laboratory at Stanford University. A key advantage of these routines was that they incorporated methods to obtain numerical estimates of the first and second derivatives, rather than requiring the user to provide them. Alleviated of the burden of finding algebraic expressions for the derivatives, behavior geneticists in the 1970s and 1980s were able to tackle a wider variety of both statistical and substantive problems [3, 7].

Nevertheless, some problems remained which curtailed the widespread adoption of model-fitting by maximum likelihood. Not least of these was that the geneticist had to learn to use FORTRAN or a similar programming language in order to fit models to their data, particularly if they wished to fit models for which no suitable software was already available. Those skilled in programing were able to assemble loose collections of programs, but these typically involved idiosyncratic formats for data input, program control and interpretation of output. These limitations in turn made it difficult to communicate use of the software to other users, difficult to modify the code for alternative types of data or pedigree structure, and difficult to fit alternative statistical models. Fortunately, the development by Karl Jöreskog and Dag Sörbom of a more general program for maximum likelihood estimation, called LISREL, alleviated many of these problems [1, 8]. Although other programs, such as COSAN, developed by C. Fraser & R. P. McDonald [5] existed, these proved to be less popular with the behavior genetic research community. In part, this was because they did not facilitate the simultaneous analysis of data collected from multiple groups, such as from MZ and DZ twin pairs, which is a prerequisite for estimating heritability and other components of variance. The heart of LISREL's flexibility was its matrix algebra formula for the specification of what are now usually called **structural equation models** In essence, early versions of the program allowed the user to specify the elements of matrices in the formula:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Lambda}_y\mathbf{A}(\mathbf{\Gamma\Phi\Gamma}' + \mathbf{\Psi})\mathbf{A}'\mathbf{\Lambda}_y' + \mathbf{\Theta}_\epsilon & \begin{matrix}\mathbf{\Lambda}_y\mathbf{A\Gamma\Phi\Lambda}_x' \\ +\mathbf{\Theta}_{\delta\epsilon}'\end{matrix} \\ \mathbf{\Lambda}_x\mathbf{\Phi\Gamma}'\mathbf{A}'\mathbf{\Lambda}_y' + \mathbf{\Theta}_{\delta\epsilon} & \mathbf{\Lambda}_x\mathbf{\Phi\Lambda}_x' + \mathbf{\Theta}_\delta \end{pmatrix}, \tag{1}$$

where $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$. The somewhat cumbersome expression (1) is the predicted covariance within a set of dependent y variables (upper left), within a set of independent variables, x (lower right) and between these two sets (lower left and upper right). Using this framework, a wide variety of models may be specified. The program was used in many of the early (1987–1993) workshops on Methodology for Genetic Studies of Twins and Families, and was used in the Neale and Cardon [12] text. What was particularly remarkable about LISREL, COSAN, and similar products that emerged in the 1980s was that they formed a bridge between a completely general programming language such as FORTRAN or C, and a purpose-built piece of software that was limited to one or two models or types of input data. These programs allowed the genetic epidemiologist to fit, by maximum likelihood, a vast number of models to several types of summary statistic, primarily means and correlation and covariance matrices, and all without the need to write or compile FORTRAN.

Although quite general, some problems could not be tackled easily within the LISREL framework, and others appeared insurmountable. These problems included the following:

1. A common complication in behavior genetic studies is that human families vary in size and configuration, whereas the covariance matrices used as input data assumed an identical structure for each family.
2. Data collected from large surveys and interviews are often incomplete, lacking responses on one or more items from one or more relatives.
3. Genetic models typically specify many linear constraints among the parameters; for example, in the **ACE model** the impact of genetic and environmental factors on the phenotypes is expected to be the same for twin 1 and twin 2 within a pair, and also for MZ and DZ pairs.
4. Certain models – such as those involving mixed genetic and cultural transmission from parent to child [12] – require nonlinear constraints among the parameters.
5. Some models use a likelihood framework that is not based on normal theory.
6. Methods to handle, for example, imperfect diagnosis of identity-by-descent in sib pairs, or zygosity in twin pairs, require the specification of the likelihood as a finite mixture distribution.
7. Tests for **genotype X environment** (or age or sex) interactions may involve continuous or discrete moderator variables.
8. Model specification using matrices is not straightforward especially for the novice.

These issues led to the development, by the author of this article and his colleagues, of the Mx software [10, 11]. Many of the limitations encountered in the version of LISREL available in the early 1990s have been lifted in recent versions (and in LISREL's competitors in the marketplace such as EQS, AMOS and MPLUS). However, at the time of writing, Mx seems to be the most popular program for the analysis of data from twins and adoptees. In part, this may be due to economic factors, since Mx is freely distributed while the commercial programs cost up to $900 per user.

Mx was initially developed in 1990, using FORTRAN and the NPSOL numerical optimizer from Professor Walter Murray's group [6]. Fortunately, compiled FORTRAN programs still generate some of the fastest executable programs of any high-level programming language. An early goal of Mx was to liberate the user from the requirement to use the single (albeit quite general) matrix algebra formula that LISREL provided. Therefore, the software included a matrix algebra interpreter, which addressed problem three because large numbers of equality constraints could be expressed in matrix form. It also permitted the analysis of raw data with missing values, by maximizing the likelihood of the observed data, instead of the likelihood of summary statistics, which simultaneously addressed problems one and two. Facilities for matrix specification of linear and nonlinear constraints addressed problem four.

Problem five was partially addressed by the advent of user-defined fit functions, which permit parameter estimation under a wider variety of models and statistical theory. In 1994, raw data analysis was extended by permitting the specification of matrix elements to contain variables from the raw data to overcome problem seven. This year also saw the development of a graphical interface to draw path diagrams and fit models directly to the data (problem eight) and the following year mixture distributions were added to address problem six. More recently, developments have focused on the analysis of binary and ordinal data; these permit a variety of Item Response Theory (*see* **Item Response Theory (IRT) Models for Dichotomous Data**) and **Latent Class models** to be fitted relatively efficiently, while retaining the genetic information in studies of twins and other relatives. These developments are especially important for behavior genetic studies, since conclusions about sex-limitation and genotype-environment interaction may be biased by inconsistent measurement [9].

The development of both commercial and noncommercial software continues today. Many of the features developed in Mx have been adopted by the commercial packages, particularly the analysis of raw data. There is also some progress in the development of a package for structural equation model-fitting package written in the R language (`http://r-project.org`). Being an open-source project, this development is should prove readily extensible. Overall, the feature sets of these programs overlap, so that each program has some unique features and some that are in common with some of the others.

Several new methods, most notably Bayesian approaches involving Monte Carlo Markov Chain (MCMC) (*see* **Markov Chain Monte Carlo and Bayesian Statistics**) algorithms permit greater flexibility in model specification, and in some instances have more desirable statistical properties [2]. For example, estimation of (genetic or environmental or phenotypic) factor scores is an area where the MCMC approach has some clear advantages. Bayesian factor score estimates will incorporate the error inherent in the estimates of the factor loadings, whereas traditional methods will assume that the factor loadings are known without error and are thus artificially precise. Future developments in this area seem highly likely.

Acknowledgment

Michael C. Neale is primarily supported from PHS grants MH-65322 and DA018673.

References

[1] Baker, L.A. (1986). Estimating genetic correlations among discontinuous phenotypes: an analysis of criminal convictions and psychiatric-hospital diagnoses in Danish adoptees, *Behavior Genetics* **16**, 127–142.

[2] Eaves, L., & Erkanli, A. (2003). Markov chain Monte Carlo approaches to analysis of genetic and environmental components of human developmental change and g × e interaction, *Behavior Genetics* **33**(3), 279–299.

[3] Eaves, L.J., Last, K., Young, P.A., & Martin, N.G. (1978). Model fitting approaches to the analysis of human behavior, *Heredity* **41**, 249–320.

[4] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society of London, Series A* **222**, 309–368.

[5] Fraser, C. (1988). *Cosan User's Guide. Unpublished Documentation*, Centre for Behavioural Studies in Education, University of England, Armidale Unpublished documentation, p. 2351.

[6] Gill, P.E., Murray, W., & Wright, M.H. (1991). *Numerical Linear Algebra and Optimization*, Vol. 1, Addison-Wesley, New York.

[7] Jinks, J.L., & Fulker, D.W. (1970). Comparison of the biometrical genetical, MAVA, and classical approaches

to the analysis of human behavior, *Psychological Bulletin* **73**, 311–349.

[8] Jöreskog, K.G., & Sörbom, D. (1986). *LISREL: Analysis of Linear Structural Relationships by the Method of Maximum Likelihood*, National Educational Resources, Chicago.

[9] Lubke, G.H., Dolan, C.V., & Neale, M.C. (2004). Implications of absence of measurement invariance for detecting sex limitation and genotype by environment interaction, *Twin Research* **7**(3), 292–298.

[10] Neale, M.C. (1991). *Mx: Statistical Modeling*, Department of Human Genetics, Virginia Commonwealth University, Richmond, VA.

[11] Neale, M., Boker, S., Xie, G., & Maes, H. (2003). *Mx: Statistical Modeling*, 6th Edition, Department of Psychiatry, Virginia Commonwealth University, Richmond, VA.

[12] Neale, M.C. & Cardon, L.R. (1992). *Methodology for Genetic Studies of Twins and Families*, Kluwer Academic Publishers.

(*See also* **Structural Equation Modeling: Software**)

MICHAEL C. NEALE

Software for Statistical Analyses

N. Clayton Silver

Volume 4, pp. 1880–1885

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Software for Statistical Analyses

Introduction

One of the first statistical packages on the market was the Biomedical Statistical Software Package (BMDP), developed in 1960 at the University of California, Los Angeles. One of the reasons for its popularity was that it was written in FORTRAN, which was a major computer language in the 1950s and 1960s.

In 1968, three individuals from Stanford University, Norman Nie, a social scientist, 'Tex' Hull, a programmer with an MBA, and Dale Bent, an operations researcher, developed the Statistical Package for the Social Sciences (SPSS) for analyzing a multitude of social science research data. McGraw-Hill published the first manual for SPSS in 1970. In the mid-1980s, SPSS was first sold for personal computer use [18].

Statistical Analysis Systems (SAS) software was developed in the early 1970s at North Carolina State University by a number of students who discovered that there was no software for managing and analyzing agricultural data. These students wrote the software for a variety of student projects, which provided the impetus for SAS [16].

MINITAB was developed by Barbara Ryan, Thomas Ryan, and Brian Joiner in 1972 in an attempt to make statistics more interesting to students. Because SPSS, SAS, and BMDP were difficult for undergraduates, these innovators constructed a software program that could be learned in about one hour of class time [13]. For an overview of the history of the major statistical software companies, see the statistics site at George Mason University [6].

Modern Statistical Software

The eight general purpose software packages listed in Table 1 perform descriptive statistics, **multiple linear regression**, **analysis of variance (ANOVA)**, **analysis of covariance (ANCOVA)**, **multivariate analysis**, and nonparametric methods (*see* **Distribution-free Inference, an Overview**). For each package, a website and system compatibility are shown.

The hierarchical linear modeling program described in Table 2 estimates multivariate linear models from incomplete data and imports data from other statistical packages. It computes **latent variable** analysis, ordinal and multinomial regression for two-level data (*see* **Logistic Regression**), and **generalized estimating equations** with robust standard errors.

Table 3 lists six programs that perform meta-analyses. Biostat has a variety of **meta-analysis**

Table 1 General statistical packages

Name	Website	Compatibility
BMDP	`http://www.statsol.ie/bmdp/bmdp.htm`	Windows 95, 98, 2000, NT
JMP	`www.jmpdiscovery.com`	Windows 95, 98, 2000, NT; Macintosh OS 9.1 or higher
MINITAB	`www.minitab.com`	Windows 98, 2000, NT, Me, XP
SAS	`http://www.sas.com`	Windows 95, 98, 2000, NT, XP
SPSS	`http://www.spss.com`	Windows 95, 98, 2000, NT, XP
STATA	`http://www.stata.com`	Windows 98, 2000, NT, XP; Macintosh OS X 10.1; UNIX
STATISTICA	`http://www.statsoftinc.com`	Windows 95, 98, 2000, NT, Me, XP; Macintosh
SYSTAT	`http://www.systat.com`	Windows 95, 98, 2000, NT, XP

Table 2 Hierarchical linear modeling

Name	Website	Compatibility
HLM-5	`http://www.ssicentral.com/hlm/hlm.htm`	Windows 95, 98, 2000, NT, XP; UNIX

Table 3 Meta-analysis

Name	Website	Compatibility
Biostat	http://www.meta-analysis.com	Windows 95, 98, 2000, NT
Meta	http://users.rcn.com/dakenny/meta.htm	DOS
Meta-analysis	http://www.lyonsmorris.com/MetaA/links.htm	Windows 95, 98, NT, Me; MacIntosh OS
Meta-analysis 5.3	http://www.fu.berlin.de/gesund_engl/meta_e.htm	IBM compatibles (DOS 3.3 or higher)
MetaWin 2.0	http://www.metawinsoft.com	Windows 95, 98, NT
WeasyMA	http://www.weasyma.com	Windows 95, 98, 2000, NT, XP

Table 4 Power analysis

Name	Website	Compatibility
G*Power	http://www.psych.uni-duesseldorf.de/aap/projects/gpower	DOS or MacIntosh OS systems
Power analysis	http://www.math.yorku.ca/SCS/online/power/	Program on website
PASS-2002	http://www.ncss.com	Windows 95, 98, 2000, NT, Me, XP
Java applets for power	http://www.stat.uiowa.edu/~rlenth/power/index.html	Program on website
Power and precision	http://www.power-analysis.com	Windows 95, 98, 2000, NT
Power on	http://www.macupdate.com/info.php/id/7624	MacIntosh OS X 10.1 or later
StudySize 1.0	http://www.studysize.com/index.htm	Windows 95, 98, 2000, NT, XP

algorithms. It provides **effect size** indices, moderator variables (*see* **Moderation**), and forest plots. Meta computes and pools effect sizes, and tests whether the average effect size differs from zero. Meta-analysis performs the Hunter-Schmidt method. It computes effect sizes and corrects for range restriction and sampling error. Meta-analysis 5.3 has algorithms utilizing exact probabilities and effect sizes (d or r). The program also provides **cluster analysis** and **stem-and-leaf displays** of correlation coefficients. MetaWin 2.0 computes six different effect sizes and performs cumulative and graphical analyses. It reads text, EXCEL, and Lotus files. WeasyMA performs cumulative analyses and it provides funnel, radial, and forest plots.

Rothstein, McDaniel, and Borenstein [15] provided a summary and a brief evaluative statement (e.g., user friendliness) of eight meta-analytic software programs. They found the programs Meta, True EPISTAT, and DSTAT to be user friendly.

The seven programs listed in Table 4 compute **power analyses**. G*Power, PASS-2002, and Power and Precision compute power, sample size, and effect sizes for t Tests, ANOVA, regression, and chi-square. G*power is downloadable freeware. Power Analysis by Michael Friendly and the Java Applet for Power by Russ Lenth are interactive programs found on their websites. Power Analysis computes power and sample size for one effect in a factorial ANOVA design, whereas the Java Applet for Power computes power for the t Test, ANOVA, and proportions. Power On computes the power and sample sizes needed for t Tests. StudySize 1.0 computes power and sample size for t Tests, ANOVA, and chi-square. It also computes **confidence intervals** and performs **Monte Carlo simulations**.

The website http://www.insp.mx/dinf/stat_list.html lists a number of additional power programs. Yu [21] compared the Power and Precision package [2], PASS, G*Power, and a SAS

Table 5 Qualitative data analysis packages

Name	Website	Compatibility
AtlasT.ti	`http://www.atlasti.de/features.shtml`	Windows 95, 98, 2000, NT, XP; MacIntosh; SUN
NUD*IST – N6	`http://www.qsr.com`	Windows 2000, Me, XP
Nvivo 2.0	`http://www.qsr.com`	Windows 2000, Me, XP

Table 6 Structural equation modeling

Name	Website	Compatibility
AMOS 5	`http://www.smallwaters.com/amos/features.html`	Windows 98, Me, NT4, 2000, XP
EQS 6.0	`http://www.mvsoft.com/eqs60.htm`	Windows 95, 98, 2000, NT, XP; UNIX
LISREL 8.5	`http://www.ssicentral.com/lisrel/mainlis.htm`	Windows 95, 98, 2000, NT, XP; MacIntosh OS 9

Macro developed by Friendly [5]. Yu recommended the Power and Precision package (which is marketed by SPSS, Inc. under the name Sample Power) because of its user-friendliness and versatility. In a review of 29 different power programs [20], the programs nQuery advisor, PASS, Power and Precision, Statistical Power Analysis, Stat Power, and True EPISTAT were rated highest with regard to ease of learning. PASS received the highest mark for ease of use.

Three packages that analyze qualitative data (*see* **Qualitative Research**) are listed in Table 5. Atlas.ti generates PROLOG code for building knowledge based systems and performs semiautomatic coding with multistring text search and pattern matching. It integrates all relevant material into Hermeneutic Units, and creates and transfers knowledge networks between projects. NUD*IST – N6 provides rapid handling of text records, automated data processing, and the integration of qualitative and quantitative data. It codes questionnaires or focus group data. NVivo 2.0 performs qualitative modeling and integrated searches for qualitative questioning. It provides immediate access to interpretations and insights, and tools that show, shape, filter, and assay data.

Barry [1] compared Atlas.ti and NUD*IST across project complexity, interconnected versus sequential structure, and software design. According to Barry, Atlas.ti's strengths were a well-designed interface that was visually attractive and creative, and its handling of simple sample, one timepoint projects. She believed that NUD*IST had a better searching structure and was more suitable for complex projects, although it was not as visually appealing as Atlas.ti.

Structural equation modeling packages are listed in Table 6. AMOS 5 fits multiple models into a single analysis. It performs **missing data** modeling (via casewise maximum likelihood), **bootstrap** simulation, **outlier detection**, and multiple fit statistics such as Bentler–Bonnet and Tucker–Lewis indices. EQS 6.0 performs EM-type missing data methods, heterogeneous kurtosis methods, subject weighting methods, multilevel methods, and resampling and simulation methods and statistics. LISREL 8.5 performs structural equation modeling with incomplete data, multilevel structural equation modeling, formal inference based recursive modeling, and multiple imputation and nonlinear multilevel regression modeling.

Kline [8] analyzed the features of AMOS, EQS, and LISREL and concluded that AMOS had the most user-friendly graphical interface; EQS had numerous test statistics and was useful for nonnormal data; LISREL had flexibility in displaying results under a variety of graphical views. He concluded that all three programs were capable of handling many SEM situations (*see* **Structural Equation Modeling: Software**).

The package in Table 7 not only computes **confidence intervals** for **effect sizes** but it also performs six different simulations that could be used for teaching concepts such as meta-analysis and power. The program runs under Microsoft Excel97 or Excel2000.

Table 7 Confidence intervals for effect sizes

Name	Website	Compatibility
ESCI	`http://www.latrobe.edu.au/psy/esci/`	Windows 95, 98, 2000, NT, Me, XP

Evaluating Statistical Software: A Consumer Viewpoint

Articles comparing statistical software often focus on the accuracy of the statistical calculations and tests of random number generators (e.g., [12]). But, Kuonen and Roehrl [9] list characteristics that may be of more utility to the behavioral researcher. These characteristics are performance (speed and memory); scalability (maximizing computing power); predictability (computational time); compatibility (generalizable code for statistical packages or computers); user-friendliness (easy to learn interfaces and commands); extensibility ('rich, high-level, object-oriented, extensive, and open language' (p. 10)); intelligent agents (understandable error messages); and good presentation of results (easy-to-understand, orderly output). All information should be labeled and presented in an easy to read font type and relatively large font size. Moreover, it should fit neatly and appropriately paged on standard paper sizes.

Kuonen and Roehrl [9] evaluated nine statistical software packages on seven of these characteristics (excluding predictability). They found that no major statistical software package stands out from the rest and that many of them have poor performance, scalability, intelligent agents, and compatibility. For additional reviews of a wider range of major statistical software packages, see `www.stats.gla.ac.uk/cti/activities/reviews/alphabet.html`.

Problems with Statistical Software Packages

Researchers often blindly follow the statistical software output when reporting results. Many assume that the output is appropriately labeled, perfectly computed, and is based on up-to-date procedures. Indeed, each subsequent version of a statistical software package generally has more cutting edge procedures and is also more user friendly. Unfortunately, there are still a number of difficulties that these packages have not solved. The examples that follow are certainly not exhaustive, but they demonstrate that researchers should be more cognizant of the theory and concepts surrounding a statistical technique rather than simply parroting output. As one example, behavioral science researchers often correlate numerous measures. Many of these measures contain individual factors that are also contained in the correlation matrix. It is common to have, for example, a 15×15 correlation matrix and associated probability values of tests of the null hypothesis that $\rho = 0$ for each correlation. These probabilities are associated with the standard F, t, or z Tests. Numerous researchers (e.g., [4]) have suggested that multiple F, t, or z tests for testing the null hypothesis that $\rho = 0$, may lead to an inflation of Type I error (*see* **Multiple Comparison Procedures**). In some cases, the largest correlation in the matrix may have a Type I error rate that is above. 40! To guard against this Type I error rate inflation, procedures such as the multistage Bonferroni [10], step up Bonferroni [14] and step down Bonferroni [7], and the rank order method [19] have been proposed. In standard statistical software packages, these and other options are not available, leaving the researcher to resort to either a stand-alone software program or to ignore the issue completely.

As a second example, Levine and Hullett [11] reported that in SPSS for Windows 9.0 (1998), the measure of effect size in the **generalized linear models** procedure was labeled as eta squared, but it should have been labeled as partial eta squared. According to Levine and Hullett, the measure of effect size was correctly labeled in the documentation, but not in the actual printouts. Hence, researchers were more likely to misreport effect sizes for two-way or larger ANOVAs. In some cases, they noted that the effect sizes summed to more than 1.0. Fortunately, in later editions of SPSS for Windows, the label was changed to partial eta squared, so effect size reporting errors should be reduced for output from this and future versions of SPSS.

Finding Other Behavioral Statistical Software

Although many statistical software packages perform a wide variety of statistics, there are times when software packages may not provide a specific hypothesis test. For example, suppose an industrial psychologist correlated salary with job performance at time 1, and after providing a 5% raise, correlated salary and job performance six months later. This constitutes testing the null hypothesis of no statistically significant difference between dependent population correlations with zero elements in common ($\rho_{12} = \rho_{34}$). As an additional example, suppose an educational psychologist was interested in determining if the correlation between grade point average and scholastic aptitude test scores are different for freshmen, sophomores, juniors, and seniors. This tests the null hypothesis that there are no statistically significant differences among independent population correlations ($\rho_1 = \rho_2 = \rho_3 = \rho_4$). Finally, suppose one wanted to orthogonally break a large chi-square contingency table into individual 2×2 contingency tables using the Bresnahan and Shapiro [3] methodology. In all three cases, researchers either need to perform the computational procedures by hand, which is extremely cumbersome, or ask their local statistician to program them. Fortunately, during the past 25 years, many behavioral statisticians published stand-alone programs that augment the standard statistical software packages. In many cases, the programs are available free or a nominal cost. One source of statistical software is a book by Silver and Hittner [17], a compendium of more than 400 peer-reviewed stand-alone statistical programs. They provided a description of each program, its source, compatibility and memory requirements, and information for obtaining the program.

A number of journals publish statistical software. *Applied Psychological Measurement* sporadically includes the Computer Program Exchange that features one-page abstracts of statistical software. *Behavior Research Methods, Instruments, and Computers* is a journal that is devoted entirely to computer applications. The *Journal of Statistical Software* is an Internet peer-reviewed journal that publishes and reviews statistical software, manuals, and user's guides.

A number of website links provide free or nominal cost statistical software. The website (`http://members.aol.com/johnp71/index.html`), developed by John C. Pezzullo, guides users to free statistical software, some of which are downloads available on a 30-day free trial basis. This extremely well-documented, highly encompassing website lists a wide variety of general packages, subset packages, survey, testing, and measurement software, programming languages, curve fitting and modeling software, and links to other free software. The websites `http://www.statserv.com/softwares.html` and `http://www.statistics.com/content/commsoft/fulllist.php3` provide lists of statistical software packages that range from general purpose programs such as SPSS to more specific purpose programs such as BILOG3 for Item Response Theory. These websites provide brief descriptions and system compatibilities for more than 100 statistical programs. The website `http://sociology.ca/sociologycalinks.html` has links to a variety of statistical software package tutorials.

The Future of Statistical Software

Although it is difficult to prognosticate the future, there have been trends over the past few years that may continue. First, the major statistical software packages will continue to be upgraded in terms of user-friendly interface and general statistical techniques. This upgrading may include asking questions of the user, similar to that of income tax software, to assure that the design and analysis are appropriate. Help files and error messages will also be more user friendly. Moreover, widely used individual statistical techniques (e.g., meta-analysis and reliability generalization) will continue to be provided as separate programs offered by statistical software companies or by individual statisticians. The programming of less widely used techniques (e.g., testing the difference between two independent **intraclass correlations**), will still be performed by individual statisticians, although, there may be fewer outlets for their publication. Hopefully, there will be additional print or online statistical software journals that will describe computer programs understandably for the applied researcher. Without additional peer-reviews of statistical software for seldom-used techniques, these statistical programs may not meet acceptable standards for quality and quantity.

References

[1] Barry, C.A. (1998). Choosing qualitative data analysis software: Atlas/ti and NUD*IST compared. *Sociological Research Online*, *3*. Retrieved August 18, 2003 from http://www.socresonline.org.uk/socresonline/3/3/4.html

[2] Borenstein, M., Cohen, J. & Rothstein, H. (1997). Power and precision. [Computer software]. Retrieved from http://www.dataxiom.com

[3] Bresnahan, J.L. & Shapiro, M.M. (1966). A general equation and technique for partitioning of chi-square contingency tables, *Psychological Bulletin* **66**, 252–262.

[4] Collis, B.A. & Rosenblood, L.K. (1984). The problem of inflated significance when testing individual correlations from a correlation matrix, *Journal of Research in Mathematics Education* **16**, 52–55.

[5] Friendly, M. (1991). SAS Macro programs: mpower [Computer software]. Retrieved from http://www.math.yorku.ca/SCS/sasmac/mpower/html

[6] George Mason University (n.d.). A guide to statistical software. Retrieved October 22, 2003 from www.galaxy.gmu.edu/papers/astr1.html

[7] Holm, S. (1979). A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* **6**, 65–70.

[8] Kline, R.B. (1998). Software programs for structural equation modeling: AMOS, EQS, and LISREL, *Journal of Psychoeducational Assessment* **16**, 343–364.

[9] Kuonen, D. & Roehrl, A.S.A. (n.d.). Identification of key steps to evaluate statistical software. Retrieved October 20, 2003 from http://interstat.stat.vt.edu/InterStat/articles/1998/abstracts

[10] Larzelere, R.E. & Mulaik, S.A. (1977). Single-sample tests for many correlations, *Psychological Bulletin* **84**, 557–569.

[11] Levine, T.R. & Hullett, C.R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research, *Human Communication Research* **28**, 612–625.

[12] McCullough, B.D. (1999). Assessing the reliability of statistical software: part II, *The American Statistician* **53**, 149–159.

[13] MINITAB (n.d.). Retrieved September 17, 2003, from http://www.minitab.com/company/identity/*History*.htm

[14] Rom, D.M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality, *Biometrika* **77**, 663–665.

[15] Rothstein, H.R., McDaniel, M.A. & Borenstein, M. (2002). Meta-analysis: a review of quantitative cumulation methods, in *Measuring and Analyzing Behavior in Organizations: Advances in Measurement and Data-Analysis*, F. Drasgow & N. Schmitt, eds, Jossey-Bass, San Francisco, pp. 534–570.

[16] SAS (n.d.). Retrieved September 17, 2003, from http://www.sas.com/offices/europe/uk/academic/history/

[17] Silver, N.C. & Hittner, J.B. (1998). *A Guidebook of Statistical Software for the Social and Behavioral Sciences*, Allyn & Bacon, Boston.

[18] SPSS (n.d.). Retrieved September 15, 2003, from http://www.spss.com/corpinfo/history.htm

[19] Stavig, G.R. & Acock, A.C. (1976). Evaluating the degree of dependence for a set of correlations, *Psychological Bulletin* **83**, 236–241.

[20] Thomas, L. & Krebs, C.J. (1997). A review of statistical power analysis software, *Bulletin of the Ecological Society of America* **78**, 126–139.

[21] Yu, C. (n.d.). Power analysis. Retrieved September 16, 2003 from http://seamonkey.ed.asu.edu/~alex/teaching/WBI/power_es.html

N. Clayton Silver

Spearman, Charles Edward

Pat Lovie

Volume 4, pp. 1886–1887

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Spearman, Charles Edward

Born: September 10, 1863, in London, England.
Died: September 17, 1945, in London, England.

Charles Spearman's entry into academic life was not by any conventional route. On leaving school, he served as an Army officer, mainly in India, for almost 15 years. It seems, however, that military duties were mixed with reading about philosophy and psychology [10]. In 1897, having just completed a two-year course at the Army Staff College, Spearman resigned his commission and, though entirely self-taught, set out for Germany to study experimental psychology in Wundt's laboratory in Leipzig.

Although Spearman eventually obtained a PhD in 1906, his studies had been interrupted by a recall in 1900 to serve as a staff officer during the South African war. And it was during the few months between his release from these duties early in 1902 and returning to Leipzig that he carried out some mental testing studies on schoolchildren, which laid the foundations for his 'Two-Factor Theory' of human ability as well as his pioneering work on correlation and **factor analysis**. These investigations were published in the *American Journal of Psycholog*y in 1904 ([6] and [7]) – a remarkable achievement for someone with no academic qualifications other than 'passed staff college'.

Spearman's system hypothesized an underlying factor common to all intellectual activity and a second factor specific to the task; later on, these became known as g and s. Furthermore, whilst individuals were assumed to possess g (and s) to different degrees, g would be invoked to different degrees by different tasks. In Spearman's view, once the effects of superfluous variables and observational errors had been reduced to a minimum, the hierarchical pattern of intercorrelations of measurements on a 'hotchpotch' of abilities gave ample evidence of a common factor with which any intellectual activity was 'saturated' to a specific, measurable degree. In obtaining numerical values for these 'saturations', Spearman had carried out a rudimentary factor analysis, which, at last, promised a way of measuring general intelligence, sought after for so long by those in the field of psychological testing.

However, Spearman's two-factor theory was by no means universally acclaimed. Indeed, for the best part of the next 30 years, Spearman would engage in very public battles with critics on both sides of the Atlantic. Some, such as Edward Thorndike and **Louis Thurstone**, doubted that human ability or intelligence could be captured so neatly. Others, especially **Karl Pearson** (whose correlational methods Spearman had adapted), **Godfrey Thomson**, **William Brown** (early in his career), and E.B. Wilson saw grave faults in Spearman's mathematical and statistical arguments (see [4]). Spearman's Herculean effort to establish the two-factor theory as the preeminent model of human intelligence reached its peak in 1927 with the publication of *The Abilities of Man* [9]. But, more sophisticated, multiple factor theories would gradually overshadow this elegantly simple system.

In 1907, Spearman had returned to England to his first post at University College, London (UCL) as part-time Reader in Experimental Psychology. Four years later, he was appointed as the Grote Professor of Mind and Logic and head of psychology and, finally, in 1928, he became Professor of Psychology until his retirement as Emeritus Professor in 1931. He was elected a Fellow of the Royal Society in 1924 and received numerous other honours.

Perhaps Spearman's chief legacy was to put British psychology on the international map by creating the first significant centre of psychological research in the country. The 'London School', as it became known, was renowned for its rigorous pursuit of the scientific and statistical method for studying human abilities – an approach entirely consonant with principles advocated by **Francis Galton** in the previous century.

Nowadays, however, Spearman is remembered almost solely for his correlational work, especially the rank correlation coefficient (*see* **Spearman's Rho**) (although it is not entirely clear that the version we know today is in fact what Spearman developed [2]), and the so-called Spearman–Brown reliability formula. Although for a long time there were doubts and heated debates (mainly because of claims and stories put about by **Cyril Burt**, a former protégé and his successor as Professor of Psychology at UCL) about who exactly was the originator of factor analysis, Spearman's status as its creator is now firmly established (see [1] and [3]).

And yet, paradoxically, Spearman himself regarded this psychometric and statistical work as secondary to his far more ambitious mission – establishing fundamental laws of psychology, which would encompass not just the processes inherent in the two-factor theory, but all cognitive activity (see [8]). In spite of his own hopes and claims, Spearman never succeeded in developing this work much beyond an embryonic system. Ironically, though, some of his key ideas have recently reemerged within cognitive psychology.

After retirement, Spearman had continued publishing journal articles and books as well as travelling widely. By the early 1940s, however, he was in failing health and poor spirits. His only son had been killed during the evacuation of Crete in 1941 and he was suffering from blackouts which made working, and life in general, a trial. A bad fall during such an episode in the late summer of 1945 led to a bout of pneumonia. He was admitted to University College Hospital where he took his own life by jumping from a fourth floor window. Spearman believed in the right of individuals to decide when their lives should cease.

Further material about Charles Spearman's life and work can be found in [5] and [10].

References

[1] Bartholomew, D.J. (1995). Spearman and the origin and development of factor analysis, *British Journal of Mathematical and Statistical Psychology* **48**, 211–220.

[2] Lovie, A.D. (1995). Who discovered Spearman's rank correlation? *British Journal of Mathematical and Statistical Psychology* **48**, 255–269.

[3] Lovie, A.D. & Lovie, P. (1993). Charles Spearman, Cyril Burt, and the origins of factor analysis, *Journal of the History of the Behavioral Sciences* **29**, 308–321.

[4] Lovie, P. & Lovie, A.D. (1995). The cold equations: Spearman and Wilson on factor indeterminacy, *British Journal of Mathematical and Statistical Psychology* **48**, 237–253.

[5] Lovie, P. & Lovie, A.D. (1996). Charles Edward Spearman, F.R.S. (1863–1945), *Notes and Records of the Royal Society of London* **50**, 1–14.

[6] Spearman, C. (1904a). The proof and measurement of association between two things, *American Journal of Psychology* **15**, 72–101.

[7] Spearman, C. (1904b). "General intelligence" objectively determined and measured, *American Journal of Psychology* **15**, 202–293.

[8] Spearman, C. (1923). *The Nature of 'Intelligence' and the Principles of Cognition*, Macmillan Publishing, London.

[9] Spearman, C. (1927). *The Abilities of Man, their Nature and Measurement*, Macmillan Publishing, London.

[10] Spearman, C. (1930). C. Spearman, in *A History of Psychology in Autobiography*, Vol. 1, C. Murchison, ed., Clark University Press, Worcester, pp. 299–331.

PAT LOVIE

Spearman's Rho

Diana Kornbrot

Volume 4, pp. 1887–1888

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Spearman's Rho

Spearman's rho, r_s, is a measure of correlation based on ranks (*see* **Rank Based Inference**). It is useful when the raw data are ranks, as for example job applicants, or when data are ordinal. Examples of ordinal data include common rating scales based on responses ranging from 'strongly disagree' to 'strongly agree'. Figure 1 shows examples of metric data where r_s, is useful. Panel B demonstrates a nonlinear monotonic relation. Panel C demonstrates the effect of an 'outlier'.

Spearman's rho is simply the normal **Pearson product moment correlation** r computed on the ranks of the data rather than the raw data.

Calculation

In order to calculate r_s, it is first necessary to rank both the X and Y variable as shown in Table 1. Then for each pair of ranked values the difference between the ranks, D, is calculated, so that the simple formula in (1) [1] can be used to calculate r_s

$$r_s = 1 - \frac{6\sum D^2}{N(N^2-1)}. \qquad (1)$$

Equation (1) overestimates r_s if there are ties. Equations for adjusting for ties exist, but are cumbersome. The simplest method [1] is to use averaged ranks, where tied values are all given the average rank of their positions. Thus, a data set (1, 2, 2, 4) would have ranks (1, 2.5, 2.5, 4).

Table 1 Ranking of data to calculate r_s

X	2	3	5	7	8	40
Y	8	4	5	6	7	1
Rank X	6	5	4	3	2	1
Rank Y	1	5	4	3	2	6
D^2	25	0	0	0	0	25

Hypothesis Testing

For the null hypothesis of no association, that is, $r_s = 0$, and $N > 10$, (2) gives a statistic that is t-distributed with $N - 2$ degrees of freedom [1]

$$t = \frac{r_s\sqrt{N-2}}{\sqrt{1-r_s{}^2}}. \qquad (2)$$

Accurate Tables for $N \le 10$ are provided by Kendall & Gibbons [2].

Confidence limits and hypotheses about values of r_s other than 0 can be obtained by noting that the Fisher transformation gives a statistic z_r that is normally distributed with variance $1/(N-3)$

$$z_r = \frac{1}{2}\ln\left[\frac{1+r_s}{1-r_s}\right]. \qquad (3)$$

Comparison with Pearson's r and Kendall's τ

Figure 1 illustrates comparisons of r, r_s and **Kendall's tau**, τ. For normally distributed data with a linear relation, parametric tests based on r are usually more powerful than rank tests based on either r_s or τ. So in panel A: $r = 0.46$, $p = 0.044$; $r_s = 0.48$,

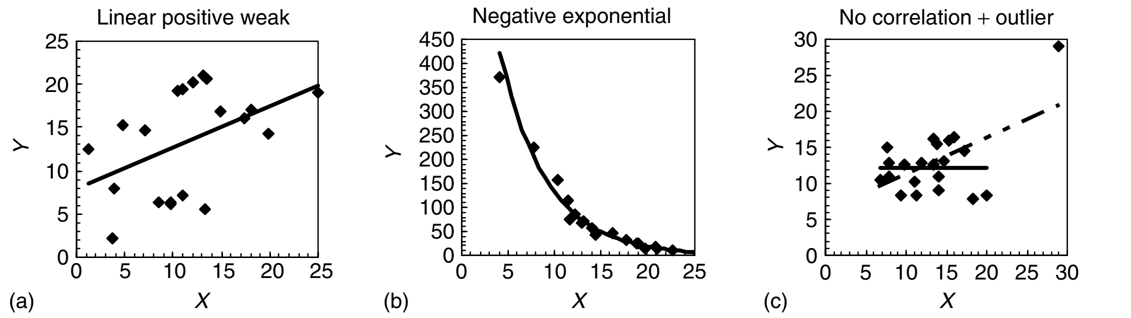


Figure 1 Three relations between Y and X to demonstrate comparisons between r, r_s, and τ

$p = 0.032$; $\tau = 0.32$, $p = 0.052$. Panel B shows a nonlinear relation, with the rank coefficients better at detecting a trend: $r = -0.82$; $r_s = -0.99$; $\tau = -0.96$, all of course highly significant. Panel C shows that rank coefficients provide better protection against outliers: $r = 0.56$, $p = 0.010$; $r_s = 0.18$, $p = 0.443$; $\tau = 0.15$, $p = 0.364$. The outlier point (29, 29) no longer causes a spurious significant correlation. Experts [1, 2] recommend τ over r_s as the best rank based procedure, but r_s is far easier to calculate if a computer package is not available.

References

[1] Howell, D.C. (2004). *Fundamental Statistics for the Behavioral Sciences*, 5th Edition, Duxbury Press, Pacific Grove.

[2] Kendall, M.G. & Gibbons, J.D. (1990). *Rank Correlation Methods*, 5th Edition, Edward Arnold, London & Oxford.

DIANA KORNBROT

Sphericity Test

H.J. Keselman

Volume 4, pp. 1888–1890

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Sphericity Test

Repeated measures (*see* **Repeated Measures Analysis of Variance**) or longitudinal designs (*see* **Longitudinal Data Analysis**) are used frequently by researchers in the behavioral sciences where **analysis of variance** F tests are typically used to assess treatment effects. However, these tests are sensitive to violations of the assumptions on which they are based, particularly when the design is unbalanced (i.e., group sizes are unequal) [6].

To set the stage for a description of sphericity, consider a hypothetical study described by Maxwell and Delaney [10, p. 534] where 12 subjects are observed in each of four conditions (factor K), for example, at 30, 36, 42, and 48 months of age, and the dependent measure (Y) 'is the child's age-normed general cognitive score on the McCarthy Scales of Children's Abilities'. In this simple repeated measures design, the validity of the within-subjects main effects F test of factor K rests on the assumptions of normality, independence of errors, and homogeneity of the treatment-difference variances (i.e., circularity or sphericity) [3, 14, 15]. Homogeneity of treatment-difference variances means that for all possible differences in scores among the levels of the repeated measures variable [i.e., $Y(30) - Y(36), Y(30) - Y(42), \ldots, Y(42) - Y(48)$], the population variances of these differences are equal. For designs including a between-subjects grouping factor (J), the validity of the within-subjects main and interaction tests ($F_{(K)}$ and $F_{(JK)}$) rest on two assumptions, in addition to those of normality and independence of errors. First, for each level of the between-subjects factor J, the population treatment-difference variances among the levels of K must be equal. Second, the population covariance matrices (*see* **Correlation and Covariance Matrices**) at each level of J must be equal. Since the data obtained in many disciplines rarely conform to these requirements, researchers using these traditional procedures will erroneously claim treatment effects when none are present, thus filling their literature with false positive claims.

McCall and Appelbaum [12] provide an illustration as to why in many areas of behavioral science research (e.g., developmental psychology, learning psychology), the covariances between the levels of the repeated measures variable will not conform to the required covariance pattern for a valid univariate F test. They use an example from developmental psychology to illustrate this point. Specifically, adjacent-age assessments typically correlate more highly than developmentally distant assessments (e.g., 'IQ at age three correlates 0.83 with IQ at age four but 0.46 with IQ at age 12'); this type of correlational structure does not correspond to a circular (spherical) covariance structure. That is, for many applications, successive or adjacent measurement occasions are more highly correlated than nonadjacent measurement occasions, with the correlation between these measurements decreasing the farther apart the measurements are in the series. Indeed, as McCall and Appelbaum note 'Most longitudinal studies using age or time as a factor cannot meet these assumptions' (p. 403). McCall and Appelbaum also indicate that the covariance pattern found in learning experiments would also not likely conform to a spherical pattern. As they note, 'experiments in which change in some behavior over short periods of time is compared under several different treatments often cannot meet covariance requirements' (p. 403).

When the assumptions of the conventional F tests have been satisfied, the tests will be valid and will be uniformly most powerful for detecting treatment effects when they are present. These conventional tests are easily obtained with the major statistical packages (e.g., SAS [16] and SPSS [13]; *see* **Software for Statistical Analyses**). Thus, when assumptions are known to be satisfied, behavioral science researchers can adopt the conventional procedures and report the associated P values, since, under these conditions, these values are an accurate reflection of the probability of observing an F value as extreme or more than the observed F statistic.

The result of applying the conventional tests of significance to data that do not conform to the assumptions of (multisample) sphericity will be that too many null hypotheses will be falsely rejected [14]. Furthermore, as the degree of nonsphericity increases, the conventional repeated measures F tests become increasingly liberal [14].

When sphericity/circularity does not exist, the Greenhouse and Geisser [2] and Huynh and Feldt [4] tests are robust alternatives to the traditional tests, provided that the design is balanced or that the covariance matrices across levels of J are equal (*see* **Repeated Measures Analysis of Variance**). The empirical literature indicates that the Greenhouse and Geisser and Huynh and Feldt adjusted degrees of freedom tests are robust to violations of multisample sphericity as long as group sizes are equal [6]. The P values associated with these statistics will provide an accurate reflection of the probability of obtaining these adjusted statistics by chance under the null hypotheses of no treatment effects. The major statistical packages [SAS, SPSS] provide Greenhouse and Geisser and Huynh and Feldt adjusted P values. However, the Greenhouse and Geisser and Huynh and Feldt tests are not robust when the design is unbalanced [6].

In addition to the Geisser and Greenhouse [2] and Huynh and Feldt [4] corrected degrees of freedom univariate tests, other univariate, multivariate, and hybrid analyses are available to circumvent the restrictive assumption of (multisample) sphericity/circularity. In particular, Johansen's [5] procedure has been found to be robust to violations of multisample sphericity in unbalanced repeated measures designs (see [8]). I refer the reader to [6], [7], and [9].

I conclude by noting that one can assess the (multisample) sphericity/circularity assumption with formal test statistics. For completely within-subjects repeated measures designs, sphericity can be checked with Mauchly's [11] W-test. If the design also contains between-subjects grouping variables, then multisample sphericity is checked in two stages (see [3]). Specifically, one can test whether the population covariance matrices are equal across the between-subjects grouping variable(s) with Box's modified criterion M (see [3]), and if this hypothesis is not rejected, whether sphericity exists (with Mauchly's W-test). However, these tests have been found to be problematic; that is, according to Keselman et al. [8] 'These tests indicate that even when data is obtained from normal populations, the tests for circularity (the M and W criteria) are sensitive to all but the most minute departures from their respective null hypotheses, and consequently the circularity hypothesis is not likely to be found tenable' (p. 481). Thus, it is recommended that researchers adopt alternative procedures, as previously noted, for assessing the effects of repeated measures/longitudinal variables. Lastly, it should be noted that Boik [1] discusses multisample sphericity for repeated measures designs containing multiple dependent variables.

References

[1] Boik, R.J. (1991). The mixed model for multivariate repeated measures: validity conditions and an approximate test, *Psychometrika* **53**, 469–486.

[2] Greenhouse, S.W. & Geisser, S. (1959). On methods in the analysis of profile data, *Psychometrika* **24**, 95–112.

[3] Huynh, H. & Feldt, L. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F distributions, *Journal of the American Statistical Association* **65**, 1582–1589.

[4] Huynh, H. & Feldt, L.S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs, *Journal of Educational Statistics* **1**, 69–82.

[5] Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression, *Biometrika* **67**, 85–92.

[6] Keselman, H.J., Algina, A., Boik, R.J. & Wilcox, R.R. (1999). New approaches to the analysis of repeated measurements, in *Advances in Social Science Methodology*, Vol. 5, B. Thompson, ed., JAI Press 251–268.

[7] Keselman, H.J., Carriere, K.C. & Lix, L.M. (1993). Testing repeated measures hypotheses when covariance matrices are heterogeneous, *Journal of Educational Statistics* **18**, 305–319.

[8] Keselman, H.J., Rogan, J.C., Mendoza, J.L. & Breen, L.J. (1980). Testing the validity conditions of repeated measures F tests, *Psychological Bulletin* **87**(3), 479–481.

[9] Keselman, H.J., Wilcox, R.R. & Lix, L.M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs, *Psychophysiology* **40**, 586–596.

[10] Maxwell, S.E. & Delaney, H.D. (2004). *Designing Experiments and Analyzing Data. A Model Comparison Perspective*, 2nd Edition, Lawrence Erlbaum, Mahwah.

[11] Mauchly, J.L. (1940). Significance test for sphericity of a normal n-variate distribution, *The Annals of Mathematical Statistics* **29**, 204–209.

[12] McCall, R.B. & Appelbaum, M.I. (1973). Bias in the analysis of repeated-measures designs: some alternative approaches, *Child Development* **44**, 401–415.

[13] Norusis, M.J. (2004). *SPSS 12.0 Statistical Procedures Companion*, SPSS Inc., Prentice Hall, Upper Saddle River, NJ.

[14] Rogan, J.C., Keselman, H.J. & Mendoza, J.L. (1979). Analysis of repeated measurements, *British Journal of Mathematical and Statistical Psychology* **32**, 269–286.

[15] Rouanet, H. & Lepine, D. (1970). Comparison between treatments in a repeated measures design: ANOVA and multivariate methods, *British Journal of Mathematical and Statistical Psychology* **23**, 147–163.

[16] SAS Institute Inc. (1990). *SAS/STAT User's Guide, Version 8*, 3rd Edition, SAS Institute, Cary.

(*See also* **Linear Multilevel Models**)

H.J. KESELMAN

Standard Deviation

David Clark-Carter

Volume 4, pp. 1891–1891

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Standard Deviation

The standard deviation (SD) is a measure of spread or dispersion. It is defined as the (positive) square root of the variance. Thus, we can find the SD of a population, or of a sample of data, or estimate the SD of a population, by taking the square root of the appropriate version of the calculated variance. The standard deviation for a sample is often represented as S, while for the population, it is denoted by σ.

The standard deviation has an advantage over the variance because it is expressed in the same units as the original measure, whereas the variance is in squared units of the original measure. However, it is still affected by extreme scores.

The SD is a way of putting the mean of a set of values in context. It also facilitates comparison of the distributions of several samples by showing their relative spread. Moreover, the standard deviation of the distribution of various statistics is also called the **standard error**. The standard error of the mean (SEM), for instance, is important in inferential procedures such as the t Test.

Finally, if a set of data has a normal distribution, then approximately 68% of the population will have a score within the range of one standard deviation below the mean to one standard deviation above the mean. Thus, if IQ were normally distributed and the mean in the population were 100 and the standard deviation were 15, then approximately 68% of people from that population would have an IQ of between 85 and 115 points.

David Clark-Carter

Standard Error

David Clark-Carter

Volume 4, pp. 1891–1892

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Standard Error

Any statistic is a random variable and, thus, has its own distribution, called a **sampling distribution**. The standard error is the standard deviation of the sampling distribution of a statistic. The most commonly encountered standard error is the standard error of the mean (SEM), which is a measure of the spread of means of samples of the same size from a specific population. Imagine that a sample of a given size is taken randomly from a population and the mean for that sample is calculated and this process is repeated an infinite number of times from the same population. The standard deviation of the distribution of these means is the standard error of the mean. It is found by dividing the standard deviation (SD) for the population by the square root of the sample size (n):

$$\text{SEM} = \frac{\text{SD}}{\sqrt{n}}. \tag{1}$$

Suppose that the population standard deviation of people's recall of words is known to be 4.7 (though usually, of course, we do not know the population SD and must estimate it from the sample), and that we have a sample of six participants, then the standard error of the mean number of words recalled would be $4.7/\sqrt{6} = 1.92$.

The standard error of the mean is a basic element of parametric hypothesis tests on means, such as the z-test and the t Test, and of confidence intervals for means.

DAVID CLARK-CARTER

Standardized Regression Coefficients

Ronald S. Landis

Volume 4, pp. 1892–1892

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Standardized Regression Coefficients

Standardized regression coefficients, commonly referred to as beta weights (β), convey the extent to which two standardized variables are linearly related in regression analyses (*see* **Multiple Linear Regression**). Mathematically, the relationship between unstandardized (b) weights and standardized (β) weights is

$$b = \beta\frac{\sigma_y}{\sigma_x} \quad \text{or} \quad \beta = b\frac{\sigma_x}{\sigma_y} \tag{1}$$

where σ_x and σ_y are the standard deviations of the predictor and criterion variables respectively. Because standardized coefficients reflect the relationship between a predictor and criterion variable after converting both the z-scores, beta weights vary between -1.00 and $+1.00$.

In the simple regression case, a standardized regression coefficient is equal to the correlation (r_{xy}) between the predictor and criterion variable. With multiple predictors, however, the predictor intercorrelations must be controlled for when computing standardized regression coefficients. For example, in situations with two predictor variables, the standardized coefficients (β_1 and β_2) are computed as

$$\beta_1 = \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}$$

$$\beta_2 = \frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2} \tag{2}$$

where r_{y1} and r_{y2} are the zero-order correlations between each predictor and the criterion and r_{12}^2 is the squared correlation between the two predictor variables.

Like unstandardized coefficients, standardized coefficients reflect the degree of change in the criterion variable associated with a unit change in the predictor. Since the standard deviation of a standardized variable is 1, this coefficient is interpreted as the associated standard deviation change in the criterion.

Standardized regression coefficients are useful when a researcher's interest is the estimation of predictor–criterion relationships, independent of the original units of measure. For example, consider two researchers studying the extent to which cognitive ability and conscientiousness accurately predict academic performance. The first researcher measures cognitive ability with a 50-item, multiple-choice test; conscientiousness with a 15-item, self-report measure; and academic performance with college grade point average (GPA). By contrast, the second researcher measures cognitive ability with a battery of tests composed of hundreds of items, conscientiousness through a single-item peer rating, and academic performance through teacher ratings of student performance. Even if the correlations between the three variables are identical across these two situations, the unstandardized regression coefficients will differ, given the variety of measures used by the two researchers. As a result, direct comparisons between the unstandardized coefficients associated with each of the predictors across the two studies cannot be made because of scaling differences. Standardized regression weights, on the other hand, are independent of the original units of measure. Thus, a direct comparison of relationships across the two studies is facilitated by standardized regression weights, much like correlations facilitate generalizations better than covariances. This feature of standardized regression weights is particularly appealing to social scientists who (a) frequently cannot attach substantive meaning to scale scores and (b) wish to compare results across studies that have used different scales to measure specific variables.

RONALD S. LANDIS

Stanine Scores

David Clark-Carter

Volume 4, pp. 1893–1893

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Stanine Scores

A stanine score is a type of standardized score. Instead of standardizing the original scores to have a mean of 0 and standard deviation (SD) of 1, as is the case of ***z*-scores**, the scores are transformed into a nine-point scale; hence, the name stanine as an abbreviation for standard nine. The original scores are generally assumed to be normally distributed or to have been 'normalized' by a normalizing transformation. The transformation to stanine scores produces a distribution with a mean of 5 and a standard deviation of 1.96.

Table 1 Percentages of the distribution for each stanine score

Stanine score	1	2	3	4	5	6	7	8	9
Percentage	4	7	12	17	20	17	12	7	4

The percentages of a distribution falling into each stanine score are shown in Table 1.

We see from the table that, for example, 11% of the distribution will have a stanine score of 2 or less; in other words, 11% of a group of people taking the test would achieve a stanine score of either 1 or 2. As with any standardized scoring system, stanine scores allow comparisons of the scores of different people on the same measure or of the same person over different measures.

The development of stanine scoring is usually attributed to the US Air Force during the Second World War [1].

Reference

[1] Anastasi, A. & Urbina, S. (1997). *Psychological Testing*, 7th Edition, Prentice Hall, Upper Saddle River.

DAVID CLARK-CARTER

Star and Profile Plots

Brian S. Everitt

Volume 4, pp. 1893–1894

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Star and Profile Plots

Plotting multivariate data is often a useful step in gathering insights into their possible structure; these insights may be useful in directing later, more formal analyses (*see* **Multivariate Analysis: Overview**). An excellent review of the many possibilities is given in [1]. For comparing the relative variable values of the observations in small- or moderate-sized multivariate data sets, two similar approaches, the star plot and the profile plot, can, on occasions, be helpful.

In the star plot, each multivariate observation (suitably scaled) is represented by a 'star' consisting of a sequence of equiangular spokes called *radii*, with each spoke representing one of the variables. The length of a spoke is proportional to the variable value it represents relative to the maximum magnitude of the variable across all the observations in the sample. A line is drawn connecting the data values for each spoke.

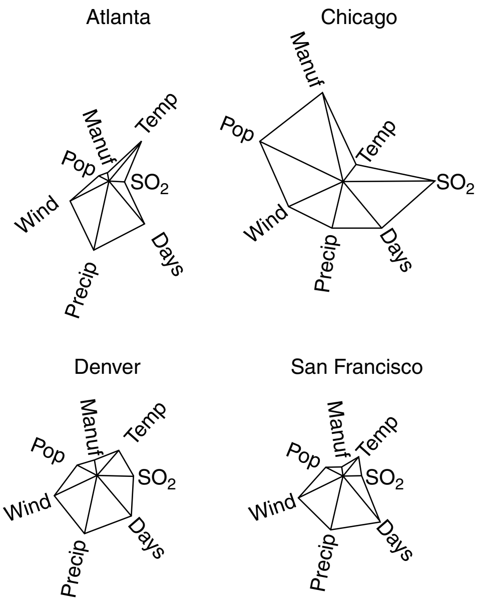


Figure 1 Star plots for air pollution data from four cities in the United States

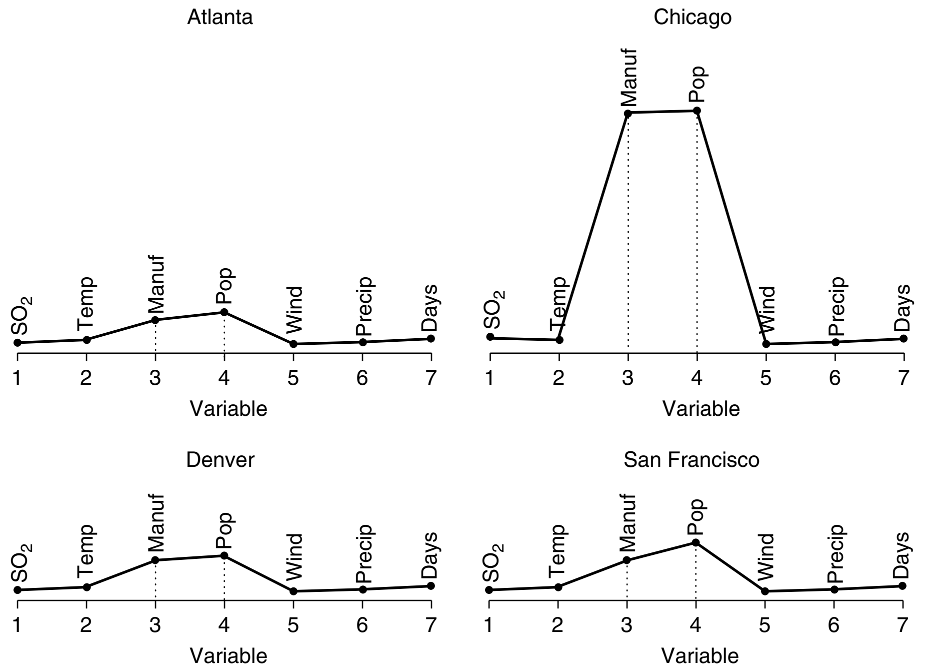


Figure 2 Profile plots for air pollution data from four cities in the United States

Table 1 Air pollution data for four cities in the United States

	SO_2 (μg/m^3)	Temp (°F)	Manuf	Pop	Wind (miles/h)	Inches (miles/h)	Days
Atlanta	24	61.5	368	497	9.1	48.34	115
Chicago	110	50.6	3344	3369	10.4	34.44	122
Denver	17	51.9	454	515	9.0	12.95	86
San Francisco	12	56.7	453	71.6	8.7	20.66	67

SO_2: Sulphur dioxide content of air,
Temp: Average annual temperature,
Manuf: Number of manufacturing enterprises employing 20 or more workers,
Pop: Population size,
Wind: Average annual wind speed,
Precip: Average annual precipitation,
Days: Average number of days with precipitation per year.

In a profile plot, a sequence of equispaced vertical spikes is used, with each spike representing one of the variables. Again, the length of a given spike is proportional to the magnitude of the variable it represents relative to the maximum magnitude of the variable across all observations.

As an example, consider the data in Table 1 showing the level of air pollution in four cities in the United States along with a number of other climatic and human ecologic variables.

The star plots of the four cities are shown in Figure 1 and the profile plots in Figure 2. In both diagrams, Chicago is clearly identified as being very different from the other cities. In the profile plot, the remaining three cities appear very similar, but in the star plot, Atlanta is identified as having somewhat different characteristics form the other two.

Star plots are available in some software packages, for example, *S-PLUS*, and profile plots are easily constructed using the command line language of the same package.

Reference

[1] Carr, D.B. (1998). Multivariate graphics, in *Encyclopedia of Biostatistics*, P. Armitage & T. Colton, eds, Wiley, Chichester.

Brian S. Everitt

State Dependence

Anders Skrondal and Sophia Rabe-Hesketh

Volume 4, pp. 1895–1895

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

State Dependencezy

It is often observed in the behavorial sciences that the current outcome of a dynamic process depends on prior outcomes, even after controlling or adjusting for covariates. The outcome is often categorical with different categories corresponding to different 'states', giving rise to the term *state dependence*.

Examples of state dependence include (a) the elevated risk of committing a crime among previous offenders and (b) the increased probability of experiencing future unemployment among those currently unemployed.

Let y_{it} be the state for unit or subject i at time or occasion t and $\mathbf{x}_{it}$ a vector of observed covariates. For simplicity, we consider dichotomous states ($y_{it} = 1$ or $y_{it} = 0$) and two occasions $t = 1, 2$. State dependence then occurs if

$$\Pr(y_{i2}|\mathbf{x}_{i2}, y_{i1}) \neq \Pr(y_{i2}|\mathbf{x}_{i2}). \qquad (1)$$

Statistical models including state dependence are often called *autoregressive models*, *Markov models* (*see* **Markov Chains**) [4] or *transition models* [1].

James J. Heckman, [2, 3] among others, has stressed the importance of distinguishing between true and spurious state dependence in social science. In the case of *true state dependence*, the increased probability of future unemployment is interpreted as 'causal'. For instance, a subject having experienced unemployment may be less attractive employers than an identical subject not having experienced unemployment. Alternatively, state dependence may be apparent, which is called *spurious state*to *dependence*. In this case, past unemployment has no 'causal' effect on future unemployment. It is rather unobserved characteristics of the subject (unobserved heterogeneity) not captured by the observed covariates $\mathbf{x}_{it}$ that produce the dependence over time. Some subjects are just more prone to experience unemployment than others, perhaps because they are not 'suitable' for the labor market, regardless of their prior unemployment record and observed covariates.

Letting ζ_i denote unobserved heterogeneity for subject i, spurious state dependence occurs if there is state dependence as in the first equation, but the dependence on the previous state disappears when we condition on ζ_i,

$$\Pr(y_{i2}|\mathbf{x}_{i2}, y_{i1}, \zeta_i) = \Pr(y_{i2}|\mathbf{x}_{i2}, \zeta_i). \qquad (2)$$

Referenceszy

[1] Diggle, P.J., Heagerty, P.J., Liang, K.-Y. & Zeger, S.L. (2002). *Analysis of Longitudinal Data*, Oxford University Press, Oxford.

[2] Heckman, J.J. (1981). Heterogeneity and state dependence, in *Studies in Labor Markets*, S. Rosen ed., Chicago University Press, Chicago, pp. 91–139.

[3] Heckman, J.J. (1991). Identifying the hand of past: distinguishing state dependence from heterogeneity, *American Economic Review* **81**, 75–79.

[4] Lindsey, J.K. (1999). *Models for Repeated Measurements*, 2nd Edition, Oxford University Press, Oxford.

Anders Skrondal and
Sophia Rabe-Hesketh

Statistical Models

David A. Kenny

Volume 4, pp. 1895–1897

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Statistical Models

Statistical analysis, like breathing, is so routinely performed that we typically do not understand very well exactly what it is that we are doing. Statistical modeling begins with measurements that are attached to units (*see* **Measurement: Overview**). Very often the units are persons, but they may be other animals, objects, or events. It is important to understand that a measurement may refer not to a single entity, but possibly to multiple entities. For example, a measure of aggression refers to at least two people, the victim and the perpetrator, not just one person. Sometimes a measurement refers to multiple levels. So, for instance, children can be in classrooms and classrooms can be in schools. Critical then in statistical analysis is determining the appropriate units and levels.

An essential part of statistical analysis is the development of a statistical model. The model is one or more equations. In each equation, a variable or set of variables are explained. These variables are commonly called *dependent variables* in experimental studies or *outcome variables* in nonexperimental studies. For these variables, a different set of variables are used to explain them and, hence, are called *explanatory variables*.

Some models are causal models (*see* **Linear Statistical Models for Causation: A Critical Review**) for which there is the belief that a change in an explanatory variable changes the outcome variable. Other models are predictive in the sense that only a statistical association between variables is claimed.

Normally, a statistical model has two different parts. In one part, the outcome variables are explained by explanatory variables. Thus, some of the variation of a variable is systematically explained. The second part of the model deals with what is not explained and is the random piece of the model. When these two pieces are placed together, the statistical model can be viewed as [1]

$$\text{DATA} = \text{FIT} + \text{ERROR}. \tag{1}$$

The 'fit' part can take on several different functional forms, but very often the set of explanatory variables is assumed to have additive effects. In deciding whether to have a variable in a model, a judgment of the improvement fit and, correspondingly, reduction of error must be made.

A fundamental difficulty in statistical analysis is how much complexity to add to the model. Some models have too many terms and are too complex, whereas other models are too simple and do not contain enough variables. A related problem is how to build such a model. One can step up and start with no variables and keep adding variables until 'enough' variables are added. Alternatively, one can step down and start with a large model and strip out of the model unnecessary variables. In some cases, the most complicated model can be stated. Such a model is commonly referred to as the *saturated model.*

Not all explanatory variables in statistical analysis are the same. Some variables are the essential variables of interest and the central question of the research is their effect. Other variables, commonly called *covariates*, are not of central theoretical interest. Rather, they are included in the model for two fundamentally different reasons. Sometimes they are included because they are presumed to reduce error. At other times, they are included because they are correlated with a key explanatory variable and so their effects need to be controlled.

Decisions about adding and removing a term from a model involves model comparison; that is, the fit of two models are compared: Two models are fit, one of which includes the explanatory variable and the other does not. If the variable is needed in the model, the fit of the latter model would be better than that of the former.

Another issue in statistical analysis is the scaling of variables in statistical models. Sometimes variables have an arbitrary scaling. For instance, gender might be dummy coded ($0 =$ male and $1 =$ female) or effect coded ($-1 =$ male and $1 =$ female) coding. The final model should be the very same model, regardless of the coding method used.

The errors in the model are typically assumed to have some sort of distribution. One commonly assumed distribution that is assumed is the normal distribution (*see* **Catalogue of Probability Density Functions**). Very typically, assumptions are made that units are randomly and independently sampled from that distribution.

Classically, in statistical analysis, a distinction is made concerning descriptive versus inferential statistics. Basically, descriptive statistics concerns the estimation of the model and its parameters. For instance, if a **multiple regression** equation were estimated, the descriptive part of the model concerns the coefficients, the intercept and slopes, and the error variance. Inferential statistics focuses on decision making: Should a variable be kept in the model or are the assumptions made in the model correct?

Model building can either be confirmatory or exploratory [2] (*see* **Exploratory Data Analysis**). In confirmatory analyses, the steps are planned in advance. For instance, if there are four predictor variables in a model, it might be assumed that three of the variables were important and one was not. So, we might first see if the fourth variable is needed, and then test that the three that were specified are in fact important. In exploratory analyses, researchers go where the data take them. Normally, statistical analyses are a mix of exploratory and confirmatory analyses. While it is often helpful to ask the data a preset set of questions, it is just as important to let the data provide answers to questions that were not asked.

A critical feature of statistical analysis is an understanding of how much the data can tell us. One obvious feature is sample size. More complex models can be estimated with larger sample sizes. However, other features are important. For instance, the amount of variation, the design of research, and the precision of the measuring instruments are important to understand. All too often, researchers fail to ask enough of their data and so perform too limited statistical analyses. Alternatively, other researchers ask way too much of their data and attempt to estimate models that are too complex for the data. Finding the right balance can be a difficult challenge.

References

[1] Judd, C.M. & McClelland, G.H. (1989). *Data Analysis: A Model-Comparison Approach*, Harcourt Brace Jovanovich, San Diego.

[2] Tukey, J.W. (1969). Analyzing data: sanctification or detective work? *American Psychologist* **24**, 83–91.

(*See also* **Generalized Linear Mixed Models**; **Generalized Linear Models (GLM)**; **Linear Multilevel Models**)

DAVID A. KENNY

Stem and Leaf Plot

Sandy Lovie

Volume 4, pp. 1897–1898

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Stem and Leaf Plot

Stem and leaf plots, together with **box plots**, form important parts of the graphical portfolio bequeathed us by **John Tukey**, the inventor of **exploratory data analysis** (EDA). These plots trade in on the unavoidable redundancy found in Western number systems in order to produce a more compact display. They can also be thought of as *value-added* **histograms** in that while both displays yield the same distributional information, stem and leaf plots allow the original sample to be easily reconstituted, thus enabling the analyst to quickly read off such useful measures as the median and the upper and lower **quartiles**.

The stem in the plot is the most significant part of a number, that is, the largest or most left-hand value, while the leaf is the increasingly less significant right-hand parts of a number. So in the number 237, 2 could be the stem, thus making 3 and 7 the leaves. In most instances, a sample will contain relatively few *differently* valued stems, and hence many redundant stems, while the same sample is likely to have much more variation in the value of the leaves, and hence less redundancy in the leaf values. What Tukey did to reduce this redundancy was to create a display made up of a single vertical column of numerically ordered stems, with the appropriate leaves attached to each stem in the form of a horizontal, ordered row. An example using the first 20 readings of 'Pulse After Exercise' (variable Pulse2) from Minitab's *Pulse* dataset (see Figure 1) will illustrate these notions. (Notice that Minitab has decided on a stem width of 5 units, so there are two stems valued 6, two 7s, two 8s, and so on, which are used to represent the numbers 60 to 64 and 65 to 69, 70 to 74 and 75 to 79, 80 to 84 and 85 to 89 respectively).

```
 1   5 | 8
 1   6 |
 1   6 |
 4   7 | 022
10   7 | 566668
10   8 | 002444
 4   8 | 8
 3   9 | 4
 2   9 | 6
 1  10 |
 1  10 |
 1  11 |
 1  11 | 8
```

Figure 1 Stem and leaf plot of 'Pulse After Exercise' data (first 20 readings on variable Pulse2). Column 1 shows the up-and-down cumulated COUNTS; column 2 contains the STEM values, and column 3 the LEAVES

Reading from the left, the display in Figure 1 consists of three columns, the first containing the cumulative count up to the middle from the lowest value (58), and down to the middle from the highest value (118); the second lists the ordered stems, and the final column the ordered leaves. Thus the very first row of the plot (1 5|8) represents the number 58, which is the smallest number in the sample and hence has a cumulated value of 1 in the first column. The sample contains nothing in the 60s, so there are no leaves for either stem, and the cumulative total remains 1 for these stems. However, there are three readings in the low 70s (70, 72 and 72), hence the fourth row reads as 7|022. This row also has a cumulative total upward from the lowest reading of $(1 + 3) = 4$, thus giving a complete row of (4 7|022). The next row has six numbers (75, four 76s, and a 78), with the row reading 7|566668, that is, a stem of 7 and six leaves. The cumulative total *upward* from the smallest number of 58 in this row is therefore $(1 + 3 + 6) = 10$, which is exactly half the sample size of 20 readings, and hence is the stopping point for this particular count. Finally, the next row reads 10 8|002444, which is the compact version of the numbers 80, 80, 82, 84, 84 and 84, while the 10 in the first position in the row represents the cumulative count from the largest pulse of 118 *down* to the middle of the sample.

Although the major role of the stem and leaf plot is to display the distributional properties of a single sample, it also easy to extract robust measures from it by simply counting up from the lowest value or down from the highest using the cumulative figures in column 1. Thus, in our sample of 20, the lower quartile is the interpolated value lying between the fifth and sixth numbers from the bottom, that is, 75.25, while the upper quartile is a similarly interpolated value of 84, with the median lying midway between the two middle numbers of 78 and 80, that is, 79. Overall, the data looks single peaked and somewhat biased toward low pulse values, a tendency that shows up even more strongly with the full 92 observations.

Stem and leaf plots are, however, less useful for comparing several samples because of their complexity, with the useful upper limit being two samples

```
       | 4 | 8
       | 5 | 44
     8 | 5 | 88
012224 | 6 | 000222222444
 66888 | 6 | 66688888888
    22 | 7 | 000000222244444
668888 | 7 | 6668
002244 | 8 | 0244
  6788 | 8 | 8
    04 | 9 | 00022
    66 | 9 |
     0 |10 |
```

Figure 2 Back-to-back stem and leaf plots of the 'Pulse after Exercise' data: males on the left, females on the right

for which the *back-to-back stem and leaf plot* was invented. This consists of the stem and leaf plot for one sample staying orientated as above, with the second one rotated through 180° in the plane and then butted up against the first. The example in Figure 2 again draws again on the Minitab *Pulse* dataset, this time using all 92 'pulse before exercise' readings (variable Pulse1); the left-hand readings are for males, the right for females – note that the counts columns have been omitted to reduce the chart clutter. Although the female data seems little more peaked than the male, the spreads and medians do not appear to differ much.

Further information on stem and leaf plots can be found in [1–3].

References

[1] Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. (1983). *Graphical Methods for Data Analysis*, Duxbury Press, Boston.

[2] Hoaglin, D.C., Mosteller, F. & Tukey, J.W., eds (1983). *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, New York.

[3] Velleman, P.F. & Hoaglin, D.C. (1981). *Applications, Basics and Computing of Exploratory Data Analysis*, Duxbury Press, Boston.

SANDY LOVIE

Stephenson, William

James Good

Volume 4, pp. 1898–1900

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Stephenson, William

Born: May 14 1902, Chopwell, Co. Durham, UK.
Died: June 14 1989, Columbia, MO.

Initially trained in Physics at the University of Durham (BSc, 1923; MSc, 1925; PhD, 1927), Stephenson also completed his Diploma in the Theory and Practice of Teaching there which brought him into contact with **Godfrey Thomson**, one of the pioneers of **factor analysis**. Inspired by his encounter with Thomson to explore the application of factor analysis in the study of mind, Stephenson moved in 1926 to University College, London to study psychophysics with **Charles Spearman**, and work as Research Assistant to Spearman and to **Cyril Burt**. He also became interested in psychoanalysis and in 1935 began analysis with Melanie Klein.

In 1936, Stephenson accepted an appointment as Assistant Director of the newly established Oxford Institute of Experimental Psychology. War service interrupted his career and he served as a Consultant to the British Armed Forces, rising to the rank of Brigadier General. He became Reader in Experimental Psychology in 1942 and successor to William Brown as Director of the Institute of Experimental Psychology in 1945. Failing to secure the first Oxford Chair in Psychology (filled by George Humphrey in 1947), Stephenson emigrated to the United States, first to the University of Chicago as a Visiting Professor of Psychology and then in 1955, when a permanent academic post at Chicago was not forthcoming, to Greenwich, Connecticut as Research Director of a leading market research firm, Nowland & Co. In 1958, he became Distinguished Research Professor of Advertising at the School of Journalism, University of Missouri, Columbia, where he remained until his retirement in 1972.

Spearman once described Stephenson as the foremost 'creative statistician' of the psychologists of his generation, a view that was endorsed by Egon Brunswik when he wrote that 'Q-technique [was] the most important development in psychological statistics since Spearman's introduction of factor analysis' [1]. Stephenson was a central figure in the development of, and debates about psychometrics and factor analysis. Although the idea of correlating persons rather than traits or test items had been proposed as early as 1915 by **Cyril Burt** [2], it was Stephenson who saw the potential of this procedure for psychological analysis. He first put forward his ideas about Q-methodology in a letter to *Nature* in 1935 [4]. A detailed exposition together with a challenge to psychology 'to put its house in scientific order' did not appear until 1953 [5]. In his writings, Stephenson employs a distinction (first put forward by Godfrey Thomson) between correlating persons (Q-methodology) and the traditional use of factor analysis in psychometrics to correlate traits or test items (R-methodology). Q-methodology applies to a population of tests or traits, with persons as variables; R-methodology to a population of persons, with tests or traits as variables. Q-methodology provides a technique for assessing a person's subjectivity or point of view, especially concerning matters of value and preference. Stephenson rejected the 'reciprocity principle' promoted by Burt and Cattell that Q and R are simply reciprocal solutions (by rows or by columns) of a single data matrix of scores from objective tests. Q-methodology was seen by Stephenson as involving two separate data matrices, one containing objective scores (R), the other containing data of a subjective kind reflecting perceived representativeness or significance (Q). This was a matter about which Burt and Stephenson eventually agreed to differ in a jointly published paper [3] (*see* **R & Q Analysis**).

When used with multiple participants, Q-methodology identifies the views that participants have in common and is therefore a technique for the assessment of shared meaning. Stephenson also developed Q for use with a single participant with multiple conditions of instruction [5, 7]. The single case use of Q affords a means of exploring the structure and content of the views that individuals hold about their worlds (for example, the interconnections between a person's view of self, of ideal self, and of self as they imagine they are seen by a variety of significant others).

In developing his ideas about Q-methodology, Stephenson eschewed Cartesian mind-body dualism, thus reflecting an important influence on his thinking of the transactionalism of John Dewey and Arthur Bentley, and the interbehaviorism of Jacob Kantor. His functional and processual theory of self was heavily influenced by Kurt Koffka and Erving Goffman [9]. Building on the work of Johan Huizinga and Wilbur Schramm, Stephenson also developed a theory of communication that focused on the social and pleasurable aspects of communication as opposed to the exchange of information [6, 8]. Following his retirement, Stephenson devoted much of his time to writing a series of papers on what had been one of his earliest preoccupations, the exploration of the links between quantum theory and subjectivity [10]. Many of Stephenson's central notions are succinctly brought together in a posthumously published monograph [11].

References

[1] Brunswik, E. (1952). *Letter to Alexander Morin*, Associate Editor, University of Chicago Press, May 20, Western Historical Manuscript Collection, Ellis Library, University of Missouri-Columbia, Columbia, 1.

[2] Burt, C. (1915). General and specific factors underlying the emotions, *British Association Annual Report* **84**, 694–696.

[3] Burt, C. & Stephenson, W. (1939). Alternative views on correlations between persons, *Psychometrika* **4**(4), 269–281.

[4] Stephenson, W. (1935). Technique of factor analysis, *Nature* **136**, 297.

[5] Stephenson, W. (1953). *The Study of Behavior: Q-Technique and its Methodology*, University of Chicago Press, Chicago.

[6] Stephenson, W. (1967). *The Play Theory of Mass Communication*, University of Chicago Press, Chicago.

[7] Stephenson, W. (1974). Methodology of single case studies, *Journal of Operational Psychiatry* **5**(2), 3–16.

[8] Stephenson, W. (1978). Concourse theory of communication, *Communication* **3**, 21–40.

[9] Stephenson, W. (1979). The communicability and operantcy of the self, *Operant Subjectivity* **3**(1), 2–14.

[10] Stephenson, W. (1988). Quantum theory of subjectivity, *Integrative Psychiatry* **6**, 180–195.

[11] Stephenson, W. (1994). *Quantum Theory of Advertising*, School of Journalism, University of Missouri-Columbia, Columbia.

JAMES GOOD

Stevens, S S

Robert B. Faux

Volume 4, pp. 1900–1902

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Stevens, S S

Born: November 4, 1906, in Ogden Utah, USA.
Died: January 18, 1973, in Vail, USA.

S.S. Stevens was a twentieth-century American experimental psychologist who conducted foundational research on sensation and perception, principally in psychoacoustics. However, it is the critical role Stevens played in the development of **measurement** and operationism for which he is probably best known by psychologists and social scientists in general [1, 2].

Upon completion of high school in 1924, Stevens' went to Belgium and France as a Mormon missionary. In 1927, his missionary work completed, Stevens entered the University of Utah where he took advanced courses in the humanities and social sciences and made a failed attempt at algebra [6]. After two years, he transferred to Stanford University where he took a wide variety courses without ever declaring a major, threatening his graduation from that institution. He did graduate in 1931 and was accepted into Harvard Medical School. A $50.00 fee and the requirement to take organic chemistry during the summer persuaded Stevens that medical school was not for him. He enrolled in Harvard's School of Education, reasoning that the tuition of $300.00 was the cheapest way to take advantage of the school's resources. He found only one course in education that looked interesting: an advanced statistic course taught by T.L. Kelley. At this time Stevens also took a course in physiology with W.J. Crozier. While exploring Crozier's laboratory one day Stevens encountered B.F. Skinner plotting data. Skinner explained that he was plotting eating curves for rats and that they could be described by power functions. Stevens admitted that he did know what power functions were to which Skinner replied that the best way for him to overcome such inferiority in mathematics was to learn it; advice Stevens was to take seriously [6, 9].

Among the most critical and far-reaching experiences for Stevens at this time was his intellectual relationship with E. G. Boring, the sole professor of psychology at Harvard at this time, as psychology was still affiliated with the philosophy department. It was while he was conducting an experiment for Boring on color perception Stevens recounts that his scientific career began: He discovered a law-like relationship between color combinations, distance, and perception. This resulted in his first published experiment. Eventually transferring from education to psychology, Stevens defended his dissertation on tonal attributes in May of 1933. Stevens was to remain at Harvard, first as an instructor of psychology, then as Professor of Psychophysics, a title conferred in 1962, until his death [9].

Stevens also audited courses in mathematics as well as in physics, becoming a Research Fellow in Physics for some time. In 1936 he settled on psychology as his profession [6, 9]. After this he spent a year with Hallowell Davis studying electrophysiology at Harvard Medical School. This proved to be another fruitful intellectual relationship, culminating in the book *Hearing* in 1938 [7]. This book was for many years considered a foundational text in psychoacoustics. In addition to this work, Stevens did research on the localization of auditory function. In 1940 he established the Psycho-Acoustic Laboratory at Harvard. Stevens gathered a group of distinguished colleagues to help in this work, among them were G. A. Miller and G. von Békésy. It was during his tenure in Stevens' laboratory that von Békésy won the Nobel Prize for his work on the ear [6, 9]. Interestingly, Stevens was quite uncomfortable in his role as classroom instructor. The only teaching for which Stevens felt affinity was the give and take of laboratory work with apprentices and editorial work with authors. This is reflected in Stevens' ability to attract a group of gifted collaborators.

For many, Stevens' most important achievement was the discovery of the Psychophysical Power Law or Stevens' Law [4]. This law describes the link between the strength of a stimulus, for example, a tone, and the corresponding sensory sensation, in this case loudness. In 1953 Stevens began research in psychophysics that would upend a longstanding psychophysical law that stated that as the strength of a stimulus grew geometrically (as a constant ratio), sensation of that stimulus grew arithmetically – a logarithmic function. This is known as the Weber-Fechner Law. This view of things held sway for many years despite a lack of empirical support. Finding accurate ways to measure experience was a fundamental question in psychophysics since its inception in the nineteenth century. Stevens set about finding a more accurate measure of this relationship. He was able to demonstrate that the relationship between the subjective intensity of a stimulus and its physical intensity itself was not a logarithmic function but was, rather, a power function. He did this by having observers assign numbers to their subjective impressions of a stimulus, in this case a tone [8, 9]. The results of Stevens' research were better explained as power functions than as logarithmic functions. That is, equal ratios in the stimulus corresponded to equal ratios in one's subjective impressions of the stimulus. Stevens was also able to demonstrate that such a relationship held for different modalities [4, 9].

Stevens' interest in the nature of measurement and operationism stemmed from his research in psychoacoustics. As noted previously, throughout the history of psychophysics more accurate measures of experience were continually being sought. It was Stevens' belief that the nature of measurement needed to be clarified if quantification of sensory attributes was to take place [6]. Throughout the 1930s Stevens set about trying to elucidate the nature of measurement. Harvard University at this time was a hotbed of intellectual debate concerning the nature of science. Stevens was exposed to the ideas of philosopher A.N. Whitehead, physicist P. Bridgman, and mathematician G.D. Birkhoff. Near the end of the decade there was an influx of European philosophers, many from the Vienna Circle, among whom was Rudolf Carnap. At Carnap's suggestion, Stevens organized a club to discuss the Science of Science. Invitations were sent and in late October 1940 the inaugural meeting took place with P.W. Bridgman discussing operationism. Bridgman argued that measurement should be based upon the operations that created it rather than on what was being measured. Throughout the latter half of the 1930s Stevens published several papers on the concept of operationism [6]. For Stevens, as for many psychologists, operationism was a way to reintroduce rigor in the formulation of concepts [6]. Measurement and operationism proved to be quite alluring to psychologists, as it was believed that if psychology was to be taken seriously as a positivistic science it required rigorous measurement procedures akin to those of physics [2, 9].

Stevens recounts that his attempt to explicate the nature of measurement and describe various scales at a Congress for the Unity of Science Meeting in 1939 was unsuccessful. Initially, Stevens identified three scales: ordinal, intensive, and extensive. From feedback given at the Congress Stevens set about forming various scales and describing the operations he used to form them [6]. Finally, in a 1946 paper Stevens presented his taxonomy of measurement scales: nominal, ordinal, interval, and ratio (*see* **Scales of Measurement**). Following Bridgman, Stevens defined each of his scales based upon the operations used to create it, leaving the form of the scale invariant, rather than upon what was being measured. Stevens further argued that these operations maintained a hierarchical relationship with each other [3]. Nominal scales have no order, they are used simply to distinguish among entities. For instance: 1 = Tall, 2 = Short, 3 = Large, 4 = Small. Neither arithmetical nor logical operations can be performed on nominal data. Ordinal scales are comprised of rank orderings of events. For example, students relative rankings in a classroom: Student A has achieved a rank of 100; Student B, a rank of 97; Student C, a rank of 83; and so on. Because the intervals between ranks are variable arithmetical operations cannot be carried out; however, logical operations such as 'more than' and 'less than' are possible. Interval scales maintain order and have equal intervals, in other words they have constant units of measurement, as in scales of temperature. The arithmetical operations of addition and subtraction are permitted, as are logical operations. Ratio scales also maintain constant units of measurement and have a true zero point, thus allowing values to be expressed as ratios.

Stevens argued that each scale type was characterized by an allowable transformation that would leave the scale type invariant. For example, nominal scales allow one-to-one substitutions of numbers as they only identify some variable [9]. Stevens used the property of invariance to relate measurement scales to certain allowable statistical procedures. For instance, the correlation coefficient r will retain its value under a linear transformation [5]. This view of measurement scales could be used as a guide to choosing appropriate statistical techniques was challenged from the time of its appearance and continues to be [2]. However, despite its many challenges, Stevens' views on measurement were quickly and widely disseminated to the psychological community. This occurred most notably through Stevens' membership in the Psychological Round Table. This group of experimental psychologists met yearly from 1936 until 1946 to discuss the latest advancements in the discipline [1].

To this day Stevens' views on measurement maintain their influence in psychology [2, 9]. Most students of psychology in their statistics classes become familiar with Stevens' four measurement scales, often without any reference to Stevens himself. Almost from its inception as a distinct academic discipline, the nature of psychology has been questioned. Is it, indeed, a science? Can it be a science? The attraction of Stevens' scales of measurement was that they offered a degree of rigor that lent legitimacy to psychology's claim to be a science. In many ways Stevens continued the tradition begun by **Gustav Fechner** and Wilhelm Wundt, among others in the nineteenth century, of applying the rigors of mathematics and science to psychological questions such as the relationship between a stimulus and the concomitant subjective experience of that stimulus [9].

References

[1] Benjamin, L.T. (1977). The psychology roundtable: revolution of 1936, *American Psychologist* **32**, 542–549.

[2] Michell, J. (1999). *Measurement in Psychology: A Critical History of a Methodological Concept*, Cambridge University Press, Cambridge.

[3] Stevens, S.S. (1946). On the theory of scales of measurement, *Science* **103**, 677–680.

[4] Stevens, S.S. (1957). On the psychophysical law, *Psychological Review* **64**, 153–181.

[5] Stevens, S.S. (1968). Measurement, statistics, and the schemapiric view, *Science* **161**, 849–856.

[6] Stevens, S.S. (1974). S.S. Stevens, in *A History of Psychology in Autobiography*, Vol. VI, G. Lindzey, ed., Prentice Hall, Englewood Cliffs, pp. 393–420.

[7] Stevens, S.S. & Davis, H. (1938). *Hearing: Its Psychology and Physiology*, Wiley, New York.

[8] Still, A. (1997). Stevens, Stanley Smith, in *Biographical Dictionary of Psychology*, N. Sheehy, A.J. Chapman & W.A. Conroy, eds, Routledge, London, pp. 540–543.

[9] Teghtsoonian, R. (2001). Stevens, Stanley, Smith (1906–73), *International Encyclopedia of the Social & Behavioral Sciences*, Elsevier Science, New York.

ROBERT B. FAUX

Stratification

VANCE W. BERGER AND JIALU ZHANG

Volume 4, pp. 1902–1905

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Stratification

The word 'stratify' means 'to make layers' in Latin. When a population is composed of several subpopulations and the subpopulations vary considerably on the factors in which we are interested, stratification can reduce variability of the estimates by dividing the population into relatively homogeneous subgroups and treating each subgroup separately. This concept can be applied for stratified sampling (sampling from within each stratum) and stratified random allocation of subjects that need not constitute a random sample, stratified or not (*see* **Survey Sampling Procedures**). The purpose of stratified sampling is to obtain more precise estimates of variables of interest. Stratified sampling divides the whole population into more homogeneous subpopulations (also called *strata*) in such a way that every observation in the population belongs to one and only one stratum (a partition). From each stratum, a sample is selected independent of other strata. The simplest method is to use a simple random sample in each stratum.

The population estimate is then computed by pooling the information from each stratum together. For example, one may conduct a survey to find out the average home price in the United States. The simplest method may be a simple random sample, by which each home in the United States has an equal chance to be selected into the sample. Then the estimated average home price in the United States would be the average home price in the sample. However, home prices around metropolitan areas, such as New York and Washington, DC, tend to be much higher than those in rural areas. In fact, the variable of interest, rural versus urban, and the influence it exerts on home prices, is not dichotomous, but rather more continuous. One can exploit this knowledge by defining three strata based on the size of the city (rural, intermediate, metropolitan).

The average home price would then be computed in each stratum, and the overall estimated home price in United States would be obtained by pooling the three average home prices with some weight function. This would result in the same estimate as that obtained from simple random sampling if the weight function were derived from the proportions of each type in the sample. That is, letting $X1/n1$, $X2/n2$, and $X3/n3$ denote the average (in the sample) home prices in rural, intermediate, and metropolitan areas, with $X = X1 + X2 + X3$ and $n = n1 + n2 + n3$, one finds that $X/n = (X1/n1)(n1/n) + (X2/n2)(n2/n) + (X3/n3)(n3/n)$. However, the key to stratified sampling is that the weight functions need not be the observed proportions $n1/n$, $n2/n$, and $n3/n$. In fact, one can use weights derived from external knowledge (such as the number of homes in each type of area), and then the sampling can constrain $n1$, $n2$, and $n3$ so that they are all equal to each other (an equal number of homes would be sampled from each type of area).

This approach, with making use of external weight functions (to reflect the proportion of each stratum in the population, instead of in the sample), results in an estimated average home price obtained with smaller variance compared to that obtained from simple random sampling. This is because each estimate is based on a more homogeneous sample than would otherwise be the case. The drawback of this stratified sampling design is that it adds complexity to the survey, and sometimes the improvement in the estimation, which may not be substantial in some cases, may not be worth the extra complexity that stratified sampling brings to the design [3].

Stratification is also used in random allocation (as opposed to the random sampling just discussed). In the context of a comparative trial (*see* **Clinical Trials and Intervention Studies**), the best comparative inference takes place when all key covariates are balanced across the treatment groups. When a false positive finding, or a Type I error occurs, this is because of differences across treatment groups in key predictors of the outcome. Such covariate imbalances can also cause Type II errors, or the masking of a true treatment effect.

Stratification according to prognostic factors, such as gender, age, smoking status, or number of children, can guarantee balance with respect to these factors. A separate randomization list, with balance built in, is prepared for each stratum. A consequence of this randomization within strata is that the numbers of patients receiving each treatment are similar not only in an overall sense, but also within each stratum. Generally, the randomization lists across the various strata are not only separate, but also independent. A notable exception is a study of etanercept for children with juvenile rheumatoid arthritis [4], which used blocks within each of two strata, and the corresponding blocks in the two strata were mirror images of each other [1].

The problem with this stratification method is that the number of strata increases quickly as the number of prognostic factors or as the level of a prognostic factor increases. Some strata may not have enough patients to be randomized. Other restricted randomization procedures (also called *stratified randomization procedures*) include, for example, the randomized block design, the maximal procedure [2], and minimization [6].

Within each block, one can consider a variety of techniques for the randomization. The maximal procedure [2] has certain optimality properties in terms of balancing chronological bias and selection bias, but generally, the random allocation rule [5] is used within each block within each stratum. This means that randomization actually occurs within blocks within strata and is conducted without any restrictions besides the blocking itself and the balance it entails at the end of each block. So, for example, if the only stratification factor were gender, then there would be two strata, male and female. This means that there would be one (balanced for treatment groups) randomization for males and another (balanced for treatment groups) randomization for females. If blocking were used within strata, with a fixed block size of four, then the only restriction within the strata would be that each block of four males and each block of four females would have two subjects allocated to each treatment group.

The first four males would constitute a block, as would the first four females, the next four females, the next four males, and so on. There is a limit to the number of strata that can be used to advantage [7]. Each binary stratification factor multiplies the existing number of strata by two; stratification factors with more than two levels multiply the existing number of strata by more than two. For example, gender is binary, and leads to two strata. Add in smoking history (never, ever, current) and there are now $2 \times 3 = 6$ strata. Add in age bracket (classified as 20–30, 30–40, 40–50, 50–60, 60–70) and this six gets multiplied by five, which is the number of age brackets. There are now 30 strata.

If a study of a behavioral intervention has only 100 subjects, then on average there would be slightly more than three subjects per stratum. This situation would defeat the purpose of stratification in that the treatment comparisons within the strata could not be considered robust. Minimization [6] can handle more stratification factors than can a stratified design. The idea behind minimization is that an imbalance function is minimized to determine the allocation, or at least the allocation that is more likely. That is, a subject to be enrolled is sequentially allocated (provisionally) to each treatment group, and for each such provisional allocation, the resulting imbalance is computed. The treatment group that results in the smallest imbalance will be selected as the favored one. In a deterministic version, the subject would be allocated to this treatment group. In a stochastic version, this treatment group would have the largest allocation probability.

As a simple example, suppose that the trial is underway and 46 subjects have already been enrolled, 23 to each group. Suppose further that the two strongest predictors of outcome are gender and age (over 40 and under 40). Finally, suppose that currently Treatment Group A has four males over 40, five females over 40, seven males under 40, and seven females under 40, while Treatment Group B has eight males over 40, four females over 40, six males under 40, and five females under 40. The 47th subject to be enrolled is a female under 40. Provisionally place this subject in Treatment Group A and compute the marginal female imbalance to be $(5 + 7 + 1 - 4 - 5) = 4$, the marginal age imbalance to be $(7 + 7 + 1 - 6 - 5) = 4$, and the joint female age imbalance to be $(7 + 1 - 5) = 3$.

Now provisionally place this subject in Treatment Group B and compute the marginal female imbalance to be $(5 + 7 - 4 - 5 - 1) = 2$, the marginal age (under 40) imbalance to be $(7 + 7 - 6 - 5 - 1) = 2$, and the joint female age imbalance to be $(7 - 5 - 1) = 1$. Using joint balancing, Treatment Group B would be preferred, as 1 is less than three. Again, the actual allocation may be deterministic, as in simply assign the subject to the group that leads to better balance, B in this case, or it may be stochastic, as in make this assignment with high probability. Using marginal balancing, this subject would still either be allocated to Treatment Group B or have a high probability of being so allocated, as 2 is less than 4. Either way, then, Treatment Group B is favored for this subject. One problem with minimization is that is leads to predictable allocations, and these predictions can lead to strategic subject selection to create an imbalance in a covariate that is not being considered by the imbalance function.

References

[1] Berger, V. *FDA Product Approval Information – Licensing Action: Statistical Review*, http://www.fda.gov/cber/products/etanimm052799.htm 1999, accessed 3/7/02.

[2] Berger, V.W., Ivanova, A. & Deloria-Knoll, M. (2003). Enhancing allocation concealment through less restrictive randomization procedures, *Statistics in Medicine* **22**(19), 3017–3028.

[3] Lohr, S. (1999). *Sampling: Design and Analysis*, Duxbury Press.

[4] Lovell, D.J., Giannini, E.H., Reiff, A., Cawkwell, G.D., Silverman, E.D., Nocton, J.J., Stein, L.D., Gedalia, A., Ilowite, N.T., Wallace, C.A., Whitmore, J. & Finck, B.K. (2000). Etanercept in children with polyarticular juvenile rheumatoid arthritis, *The New England Journal of Medicine* **342**(11), 763–769.

[5] Rosenberger, W.F. & Rukhin, A.L. (2003). Bias properties and nonparametric inference for truncated binomial randomization, *Nonparametric Statistics* **15**, 4–5, 455–465.

[6] Taves, D.R. (1974). Minimization: a new method of assigning patients to treatment and control groups, *Clinical Pharmacology Therapeutics* **15**, 443–453.

[7] Therneau, T.M. (1993). How many stratification factors are "too many" to use in a randomization plan? *Controlled Clinical Trials* **14**(2), 98–108.

Vance W. Berger and Jialu Zhang

Structural Equation Modeling: Categorical Variables

ANDERS SKRONDAL AND SOPHIA RABE-HESKETH

Volume 4, pp. 1905–1910

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Structural Equation Modeling: Categorical Variables

Introduction

Structural equation models (SEMs) comprise two components, a measurement model and a structural model. The measurement model relates observed responses or 'indicators' to **latent variables** and sometimes to observed covariates. The structural model then specifies relations among latent variables and regressions of latent variables on observed variables. When the indicators are categorical, we need to modify the conventional measurement model for continuous indicators. However, the structural model can remain essentially the same as in the continuous case.

We first describe a class of structural equation models also accommodating dichotomous and ordinal responses [5]. Here, a conventional measurement model is specified for multivariate normal 'latent responses' or 'underlying variables'. The latent responses are then linked to observed categorical responses via threshold models yielding probit measurement models.

We then extend the model to generalized latent variable models (e.g., [1], [13]) where, conditional on the latent variables, the measurement models are **generalized linear models** which can be used to model a much wider range of response types.

Next, we briefly discuss different approaches to estimation of the models since estimation is considerably more complex for these models than for conventional structural equation models. Finally, we illustrate the application of structural equation models for categorical data in a simple example.

SEMs for Latent Responses

Structural Model

The structural model can take the same form regardless of response type. Letting j index units or subjects, Muthén [5] specifies the structural model for latent variables $\boldsymbol{\eta}_j$ as

$$\boldsymbol{\eta}_j = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_j + \boldsymbol{\Gamma}\mathbf{x}_{1j} + \boldsymbol{\zeta}_j. \qquad (1)$$

Here, $\boldsymbol{\alpha}$ is an intercept vector, $\mathbf{B}$ a matrix of structural parameters governing the relations among the latent variables, $\boldsymbol{\Gamma}$ a regression parameter matrix for regressions of latent variables on observed explanatory variables $\mathbf{x}_{1j}$ and $\boldsymbol{\zeta}_j$ a vector of disturbances (typically multivariate normal with zero mean). Note that this model is defined conditional on the observed explanatory variables $\mathbf{x}_{1j}$. Unlike conventional SEMs where all observed variables are treated as responses, we need not make any distributional assumptions regarding $\mathbf{x}_{1j}$.

In the example considered later, there is a single latent variable η_j representing mathematical reasoning or 'ability'. This latent variable is regressed on observed covariates (gender, race and their interaction),

$$\eta_j = \alpha + \boldsymbol{\gamma}\mathbf{x}_{1j} + \zeta_j, \quad \zeta_j \sim \mathrm{N}(0, \psi), \qquad (2)$$

where $\boldsymbol{\gamma}$ is a row-vector of regression parameters.

Measurement Model

The distinguishing feature of the measurement model is that it is specified for *latent* continuous responses $\mathbf{y}_j^*$ in contrast to observed continuous responses $\mathbf{y}_j$ as in conventional SEMs,

$$\mathbf{y}_j^* = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_j + \mathbf{K}\mathbf{x}_{2j} + \boldsymbol{\epsilon}_j. \qquad (3)$$

Here $\boldsymbol{\nu}$ is a vector of intercepts, $\boldsymbol{\Lambda}$ a factor loading matrix and $\boldsymbol{\epsilon}_j$ a vector of unique factors or 'measurement errors'. Muthén and Muthén [7] extend the measurement model in Muthén [5] by including the term $\mathbf{K}\mathbf{x}_{2j}$ where $\mathbf{K}$ is a regression parameter matrix for the regression of $\mathbf{y}_j^*$ on observed explanatory variables $\mathbf{x}_{2j}$. As in the structural model, we condition on $\mathbf{x}_{2j}$.

When $\boldsymbol{\epsilon}_j$ is assumed to be multivariate normal (*see* **Catalogue of Probability Density Functions**), this model, combined with the threshold model described below, is a *probit* model (*see* **Probits**). The variances of the latent responses are not separately identified and some constraints are therefore imposed. Muthén sets the total variance of the latent responses (given the covariates) to 1.

Threshold Model

Each observed categorical response y_{ij} is related to a latent continuous response y_{ij}^* via a threshold model.

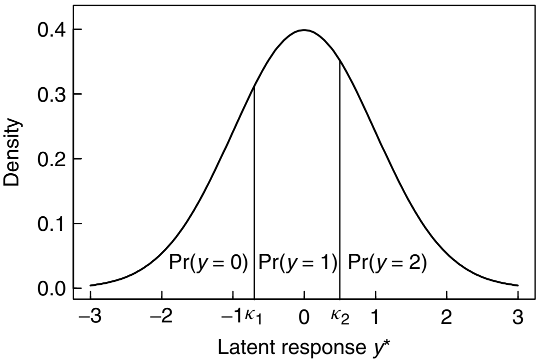


Figure 1 Threshold model for ordinal responses with three categories (This figure has been reproduced from [13] with permission from Chapman and Hall/CRC.)

For ordinal observed responses it is assumed that

$$y_{ij} = \begin{cases} 0 & \text{if} \quad -\infty < y^*_{ij} \le \kappa_{1i} \\ 1 & \text{if} \quad \kappa_{1i} < y^*_{ij} \le \kappa_{2i} \\ \vdots & \vdots \qquad\qquad \vdots \\ S & \text{if} \quad \kappa_{Si} < y^*_{ij} \le \infty. \end{cases} \tag{4}$$

This is illustrated for three categories ($S = 2$) in Figure 1 for normally distributed ϵ_i, where the areas under the curve are the probabilities of the observed responses.

Either the constants $\boldsymbol{\nu}$ or the thresholds κ_{1i} are typically set to 0 for identification. Dichotomous observed responses simply arise as the special case where $S = 1$.

Generalized Latent Variable Models

In generalized latent variable models, the measurement model is a generalized linear model of the form

$$\mathbf{g}(\boldsymbol{\mu}_j) = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_j + \mathbf{K}\mathbf{x}_{2j}, \tag{5}$$

where $\mathbf{g}(\cdot)$ is a vector of link functions which may be of different kinds handling mixed response types (for instance, both continuous and dichotomous observed responses or 'indicators'). $\boldsymbol{\mu}_j$ is a vector of conditional means of the responses given $\boldsymbol{\eta}_j$ and $\mathbf{x}_{2j}$ and the other quantities are defined as in (3). The conditional models for the observed responses given $\boldsymbol{\mu}_j$ are then distributions from the exponential family (*see* **Generalized Linear Models (GLM)**). Note that there are no explicit unique factors in the model because the variability of the responses for given values of $\boldsymbol{\eta}_j$ and $\mathbf{x}_{2j}$ is accommodated by the conditional response distributions. Also note that the responses are implicitly specified as conditionally independent given the latent variables $\boldsymbol{\eta}_j$ (*see* **Conditional Independence**).

In the example, we will consider a single latent variable measured by four dichotomous indicators or 'items' y_{ij}, $i = 1, \ldots, 4$, and use models of the form

$$\text{logit}(\mu_{ij}) \equiv \ln\left(\frac{\Pr(\mu_{ij})}{1 - \Pr(\mu_{ij})}\right) = \nu_i + \lambda_i \eta_j,$$
$$\nu_1 = 0, \lambda_1 = 1. \tag{6}$$

These models are known as two-parameter logistic item response models because two parameters (ν_i and λ_i) are used for each item i and the logit link is used (*see* **Item Response Theory (IRT) Models for Dichotomous Data**). Conditional on the latent variable, the responses are Bernoulli distributed (*see* **Catalogue of Probability Density Functions**) with expectations $\mu_{ij} = \Pr(y_{ij} = 1|\eta_j)$. Note that we have set $\nu_1 = 0$ and $\lambda_1 = 1$ for identification because the mean and variance of η_j are free parameters in (2). Using a probit link in the above model instead of the more commonly used logit would yield a model accommodated by the Muthén framework discussed in the previous section.

Models for counts can be specified using a log link and Poisson distribution (*see* **Catalogue of Probability Density Functions**). Importantly, many other response types can be handled including ordered and unordered categorical responses, rankings, durations, and mixed responses; see for example, [1, 2, 4, 9, 11, 12 and 13] for theory and applications. A recent book on generalized latent variable modeling [13] extends the models described here to 'generalized linear latent and mixed models' (GLLAMMs) [9] which can handle multilevel settings and discrete latent variables.

Estimation and Software

In contrast to the case of multinormally distributed continuous responses, **maximum likelihood estimation** cannot be based on sufficient statistics such as the empirical covariance matrix (and possibly mean vector) of the observed responses. Instead, the likelihood must be obtained by somehow 'integrating out' the latent variables $\boldsymbol{\eta}_j$. Approaches which work well but are computationally demanding include adaptive Gaussian quadrature [10] implemented in `gllamm` [8] and Markov Chain Monte Carlo methods (typically with noninformative priors) implemented in BUGS [14] (*see* **Markov Chain Monte Carlo and Bayesian Statistics**).

For the special case of models with multinormal latent responses (principally probit models), Muthén suggested a computationally efficient limited information estimation approach [6] implemented in Mplus [7]. For instance, consider a structural equation model with dichotomous responses and no observed explanatory variables. Estimation then proceeds by first estimating 'tetrachoric correlations' (pairwise correlations between the latent responses). Secondly, the asymptotic covariance matrix of the tetrachoric correlations is estimated. Finally, the parameters of the SEM are estimated using weighted least squares (*see* **Least Squares Estimation**), fitting model-implied to estimated **tetrachoric correlations**. Here, the inverse of the asymptotic covariance matrix of the tetrachoric correlations serves as weight matrix.

Skrondal and Rabe-Hesketh [13] provide an extensive overview of estimation methods for SEMs with noncontinuous responses and related models.

Example

Data

We will analyze data from the Profile of American Youth (US Department of Defense [15]), a survey of the aptitudes of a national probability sample of Americans aged 16 through 23. The responses (1: correct, 0: incorrect) for four items of the arithmetic reasoning test of the Armed Services Vocational Aptitude Battery (Form 8A) are shown in Table 1 for samples of white males and females and black males and females. These data were previously analyzed by Mislevy [3].

Model Specification

The most commonly used measurement model for ability is the two-parameter logistic model in (6) and (2) without covariates.

Item characteristic curves, plots of the probability of a correct response as a function of ability, are given by

$$\Pr(y_{ij} = 1|\eta_j) = \frac{\exp(\nu_i + \lambda_i\eta_j)}{1 + \exp(\nu_i + \lambda_i\eta_j)}. \qquad (7)$$

and shown for this model (using estimates under $\mathcal{M}_1$ in Table 2) in Figure 2.

We then specify a structural model for ability η_j. Considering the covariates

- [Female] F_j, a dummy variable for subject j being female
- [Black] B_j, a dummy variable for subject j being black

we allow the mean abilities to differ between the four groups,

$$\eta_j = \alpha + \gamma_1 F_j + \gamma_2 B_j + \gamma_3 F_j B_j + \zeta_j. \qquad (8)$$

This is a MIMIC model where the covariates affect the response via a latent variable only.

A path diagram of the structural equation model is shown in Figure 3. Here, observed variables are represented by rectangles whereas the latent variable is represented by a circle. Arrows represent regressions (not necessary linear) and short arrows residual variability (not necessarily an additive error term). All variables vary between subjects j and therefore the j subscripts are not shown.

We can also investigate if there are direct effects of the covariates on the responses, in addition to the indirect effects via the latent variable. This could be interpreted as 'item bias' or 'differential item

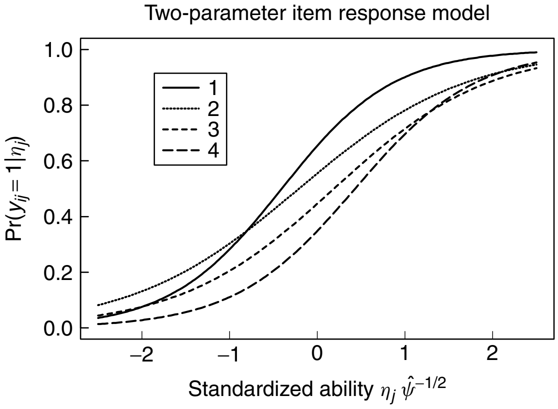


Figure 2 Item characteristic curves for items 1 to 4 (This figure has been reproduced from [13] with permission from Chapman and Hall/CRC.)

Table 1 Arithmetic reasoning data

Item Response							
1	2	3	4	White Males	White Females	Black Males	Black Females
0	0	0	0	23	20	27	29
0	0	0	1	5	8	5	8
0	0	1	0	12	14	15	7
0	0	1	1	2	2	3	3
0	1	0	0	16	20	16	14
0	1	0	1	3	5	5	5
0	1	1	0	6	11	4	6
0	1	1	1	1	7	3	0
1	0	0	0	22	23	15	14
1	0	0	1	6	8	10	10
1	0	1	0	7	9	8	11
1	0	1	1	19	6	1	2
1	1	0	0	21	18	7	19
1	1	0	1	11	15	9	5
1	1	1	0	23	20	10	8
1	1	1	1	86	42	2	4
			Total:	263	228	140	145

Source: Mislevy, R.J. Estimation of latent group effects (1985). *Journal of the American Statistical Association* **80**, 993–997 [3].

Table 2 Estimates for ability models

	$\mathcal{M}_1$		$\mathcal{M}_2$		$\mathcal{M}_3$	
Parameter	Est	(SE)	Est	(SE)	Est	(SE)
Intercepts						
ν_1 [Item1]	0	–	0	–	0	–
ν_2 [Item2]	−0.21	(0.12)	−0.22	(0.12)	−0.13	(0.13)
ν_3 [Item3]	−0.68	(0.14)	−0.73	(0.14)	−0.57	(0.15)
ν_4 [Item4]	−1.22	(0.19)	−1.16	(0.16)	−1.10	(0.18)
ν_5 [Item1] × [Black] × [Female]	0	–	0	–	−1.07	(0.69)
Factor loadings						
λ_1 [Item1]	1	–	1	–	1	–
λ_2 [Item2]	0.67	(0.16)	0.69	(0.15)	0.64	(0.17)
λ_3 [Item3]	0.73	(0.18)	0.80	(0.18)	0.65	(0.14)
λ_4 [Item4]	0.93	(0.23)	0.88	(0.18)	0.81	(0.17)
Structural model						
α [Cons]	0.64	(0.12)	1.41	(0.21)	1.46	(0.23)
γ_1 [Female]	0	–	−0.61	(0.20)	−0.67	(0.22)
γ_2 [Black]	0	–	−1.65	(0.31)	−1.80	(0.34)
γ_3 [Black] × [Female]	0	–	0.66	(0.32)	2.09	(0.86)
ψ	2.47	(0.84)	1.88	(0.59)	2.27	(0.74)
Log-likelihood		−2002.76		−1956.25		−1954.89
Deviance		204.69		111.68		108.96
Pearson X^2		190.15		102.69		100.00

Source: Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Chapman & Hall/CRC, Boca Raton [13].

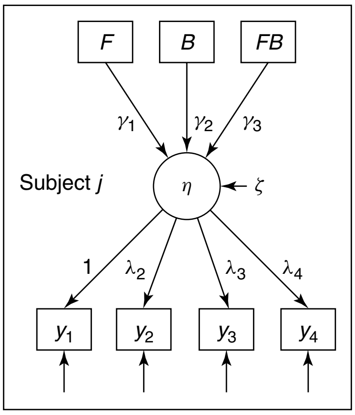


Figure 3 Path diagram of MIMIC model

functioning' (DIF), that is, where the probability of responding correctly to an item differs for instance between black women and others with the same ability (*see* **Differential Item Functioning**). Such item bias would be a problem since it suggests that candidates cannot be fairly assessed by the test. For instance, if black women perform worse on the first item ($i = 1$) we can specify the following model for this item:

$$\text{logit}[\Pr(y_{1j} = 1|\eta_j)] = \beta_1 + \beta_5 F_j B_j + \lambda_1 \eta_j. \quad (9)$$

Results

Table 2 gives maximum likelihood estimates based on 20-point adaptive quadrature estimated using `gllamm` [8]. Estimates for the two-parameter logistic IRT model (without covariates) are given under $\mathcal{M}_1$, for the MIMIC model under $\mathcal{M}_2$ and for the MIMIC model with item bias for black women on the first item under $\mathcal{M}_3$. Deviance and Pearson X^2 statistics are also reported in the table, from which we see that $\mathcal{M}_2$ fits better than $\mathcal{M}_1$. The variance estimate of the disturbance decreases from 2.47 for $\mathcal{M}_1$ to 1.88 for $\mathcal{M}_2$ because some of the variability in ability is 'explained' by the covariates. There is some evidence for a [Female] by [Black] interaction. While being female is associated with lower ability among white people, this is not the case among black people where males and females have similar abilities. Black people have lower mean abilities than both white men and white women. There is little evidence suggesting that item 1 functions differently for black females.

Note that none of the models appear to fit well according to absolute fit criteria (*see* **Model Fit: Assessment of**). For example, for $\mathcal{M}_2$, the deviance is 111.68 with 53 degrees of freedom, although the Table 1 is perhaps too sparse to rely on the χ^2 distribution.

Conclusion

We have discussed generalized structural equation models for noncontinuous responses. Muthén suggested models for continuous, dichotomous, ordinal and censored (tobit) responses based on multivariate normal latent responses and introduced a limited information estimation approach for his model class.

Recently, considerably more general models have been introduced. These models handle (possibly mixes of) responses such as continuous, dichotomous, ordinal, counts, unordered categorical (polytomous), and rankings. The models can be estimated using maximum likelihood or Bayesian analysis.

References

[1] Bartholomew, D.J. & Knott, M. (1999). *Latent Variable Models and Factor Analysis*, Arnold Publishing, London.

[2] Fox, J.P. & Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling, *Psychometrika* **66**, 271–288.

[3] Mislevy, R.J. Estimation of latent group effects, (1985). *Journal of the American Statistical Association* **80**, 993–997.

[4] Moustaki, I. (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables, *British Journal of Mathematical and Statistical Psychology* **56**, 337–357.

[5] Muthén, B.O. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent indicators, *Psychometrika* **49**, 115–132.

[6] Muthén, B.O. & Satorra, A. (1996). Technical aspects of Muthén's LISCOMP approach to estimation of latent variable relations with a comprehensive measurement model, *Psychometrika* **60**, 489–503.

[7] Muthén, L.K. & Muthén, B.O. (2004). *Mplus User's Guide*, Muthén & Muthén, Los Angeles.

[8] Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2004a). GLLAMM Manual. U.C. Berkley Division of Biostatistics Working Paper Series. Working Paper 160. Downloadable from `http://www.bepress.com/ucbbiostat/paper160/`.

[9] Rabe-Hesketh, S., Skrondal, A., Pickles & A. (2004b). Generalized multilevel structural equation modeling, *Psychometrika* **69**, 167–190.

[10] Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects, *Journal of Econometrics* in press.

[11] Sammel, M.D., Ryan, L.M. & Legler, J.M. (1997). Latent variable models for mixed discrete and continuous outcomes, *Journal of the Royal Statistical Society, Series B* **59**, 667–678.

[12] Skrondal, A. & Rabe-Hesketh, S. (2003). Multilevel logistic regression for polytomous data and rankings, *Psychometrika* **68**, 267–287.

[13] Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Chapman & Hall/CRC, Boca Raton.

[14] Spiegelhalter, D.J. Thomas, A., Best, N.G. & Gilks, W.R. (1996). *BUGS 0.5 Bayesian Analysis using Gibbs Sampling. Manual (version ii)*, MRC-Biostatistics Unit, Cambridge, Downloadable from `http://www.mrc-bsu.cam.ac.uk/bugs/documentation/contents.shtml`.

[15] U. S. Department of Defense. (1982). *Profile of American Youth*, Office of the Assistant Secretary of Defense for Manpower, Reserve Affairs, and Logistics, Washington.

ANDERS SKRONDAL AND
SOPHIA RABE-HESKETH

Structural Equation Modeling: Checking Substantive Plausibility

Adi Jaffe and Scott L. Hershberger

Volume 4, pp. 1910–1912

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Structural Equation Modeling: Checking Substantive Plausibility

There is little doubt that without the process of validation, the undertaking of test development means little. Regardless of the reliability of an instrument (*see* **Reliability: Definitions and Estimation**), it is impossible for its developers, let alone its users, to be certain that any conclusions drawn from subjects' scores have meaning. In other words, without an understanding of the substantive plausibility of a set of measurement criteria, its implications cannot be confidently stated. It is widely accepted that for a test to be accepted as 'valid', numerous studies using different approaches are often necessary, and even then, these require an ongoing validation process as societal perceptions of relevant constructs change [2].

The classic work regarding the process of construct validation was that of Cronbach and Meehl [6]. Construct validation involves a three-step process; construct definition, construct operationalization, and empirical confirmation. Some of the methods commonly used to develop empirical support for construct validation are Campbell and Fiske's [5] **multitrait–multimethod** matrix method (MTMM) (*see* **Factor Analysis: Multitrait–Multimethod**; **Multitrait–Multimethod Analyses**) and **factor analysis**. However, contemporary **structural equation modeling** techniques [4] serve as good methods for evaluating what Cronbach and Meehl termed the *nomological network* (i.e., the relationship between constructs of interests and observable variables, as well as among the constructs themselves).

The process of construct validation usually starts with the theoretical analysis of the relationship between relevant variables or constructs. Sometimes known as the substantive component of construct validation, this step requires the definition of the construct of interest and its theoretical relationship to other constructs. For example, while it is a generally accepted fact that the degree of substance addiction is dependent on the length of substance use (i.e., short-term vs. chronic) [10], research has also demonstrated a relationship between drug dependence and parental alcohol norms [8], consequences of use [11], and various personality factors [10]. A theoretical relationship could therefore be established between drug dependence – the construct to be validated – and these previously mentioned variables.

The next step in the process of construct validation requires the operationalization of the constructs of interest in terms of measurable, observed variables that relate to each of the specified constructs. Given the above mentioned example, length of substance use could easily be defined as years, months, or weeks of use, while parental alcohol norms could be assessed using either presently available instruments assessing parental alcohol use (e.g., the Alcohol Dependence Scale [12]) or some other measure of parental norms and expectancies regarding the use of alcohol and drugs. The consequences of past alcohol or drug use can be assessed as either number, or severity, of negative consequences associated with drinking. Various personality factors such as fatalism and loneliness can be assessed using available personality scales (e.g., the Social and Emotional Loneliness Scale for Adults [7]). The final structural equation model – nomological network – is shown in Figure 1.

The last step in the process of construct validation, empirical confirmation, requires the use of structural equation modeling to assess and specify the relationships between the observed variables, their related constructs, and the construct of interest. In its broadest sense, structural equation modeling is concerned with testing complex models for the structure of functional relationships between observed variables and constructs, and between the constructs themselves, one of which is the construct to be validated. Often, constructs in structural equation modeling are referred to as **latent variables**. The functional relationships are described by parameters that indicate the magnitude of the effect (direct or indirect) that independent variables have on dependent variables. Thus, a structural equation model can be considered a series of linear regression equations relating dependent variables to independent variables and other dependent variables. Those equations that describe the relations between observed variables and constructs are the *measurement* part of the model; those equations that describe the relations between constructs are the *structural* part of the model (*see* **All-X Models**; **All-Y Models**). The coefficients determining the

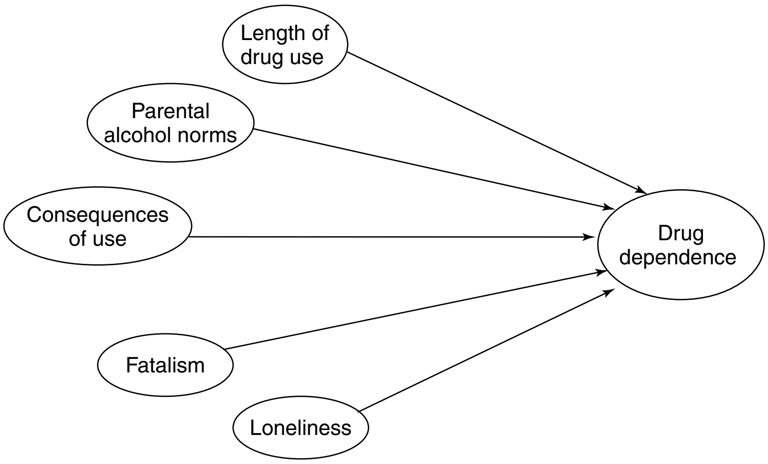


Figure 1 Nomological network for drug dependence

relations are usually the parameters we are interested in solving. By estimating the magnitude and direction of these parameters, one can evaluate the nomological network and hence provide evidence for construct validity. For example, in the model of Figure 1, we are hypothesizing that an observed score on the Alcohol Dependence Scale (not shown) is significantly and positively related to the construct of drug dependence. This relationship is part of the measurement model for drug dependence. Each construct has its own measurement model. To provide another example, we are also hypothesizing that the construct of length of drug use is significantly and positively related to the construct of drug dependence. This relation is part of the structural model linking the construct of interest, drug dependence, to the other constructs.

The most common sequence followed to confirm the nomological network is to first examine the individual measurement models of the constructs and then proceed to examine the structural model relating these constructs [1], although many variations to this sequence have been suggested [3, 9]. However one proceeds to evaluate the different components of the nomological network, ultimately the *global fit* of the nomological network must be evaluated. A great many indices of fit have been developed for this purpose, most of which define global fit in terms of the discrepancy between the observed data and that implied by the model parameters [4]. It is not until each component, both individually and combined, of the nomological network has been confirmed that one has strong evidence of construct validity.

References

[1] Anderson, J.C. & Gerbing, D.W. (1988). Structural equation modeling in practice: a review and recommended two-step approach, *Psychological Bulletin* **103**, 411–423.

[2] Benson, J. (1998). Developing a strong program of construct validation: a test anxiety example, *Educational Measurement: Issues and Practices* **17**, 10–17.

[3] Bollen, K.A. (1989). *Structural Equation modeling: An Introduction*, Wiley, New York.

[4] Bollen, K.A. (2000). Modeling strategies: in search of the Holy Grail, *Structural Equation Modeling* **7**, 74–81.

[5] Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological Bulletin* **56**, 81–105.

[6] Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests, *Psychological Bulletin* **52**, 281–302.

[7] DiTommaso, E., Brannen, C. & Best, L.A. (2004). Measurement and validity characteristics of the short version of the social and emotional loneliness scale for adults, *Educational and Psychological Measurement* **64**, 99–119.

[8] Fals-Stewart, W., Kelley, M.L., Cooke, C.G. & Golden, J.C. (2003). Predictors of the psychosocial adjustment of children living in households of parents in which fathers abuse drugs. The effects of postnatal parental exposure, *Addictive Behaviors* **28**, 1013–1031.

[9] Mulaik, S.A. & Millsap, R.E. (2000). Doing the four-step right, *Structural Equation Modeling* **7**, 36–73.

[10] Olmstead, R.E., Guy, S.M., O'Mally, P.M. & Bentler, P.M. (1991). Longitudinal assessment of the

relationship between self-esteem, fatalism, loneliness, and substance use, *Journal of Social Behavior & Personality* **6**, 749–770.

[11] Robinson, T.E. & Berridge, K.C. (2003). Addiction, *Annual Review of Psychology* **54**, 25–53.

[12] Skinner, H.A. & Allen, B.A. (1982). Alcohol dependence syndrome: measurement and validation, *Journal of Abnormal Psychology* **91**, 199–209.

ADI JAFFE AND SCOTT L. HERSHBERGER

Structural Equation Modeling: Latent Growth Curve Analysis

JOHN B. WILLETT AND KRISTEN L. BUB

Volume 4, pp. 1912–1922

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Structural Equation Modeling: Latent Growth Curve Analysis

Learning and development are ubiquitous. When new skills are acquired, when attitudes and interests develop, people change. Measuring change demands a longitudinal perspective (*see* **Longitudinal Data Analysis**), with multiple waves of data collected on representative people at sensibly spaced intervals (*see* **Panel Study**). Multiwave data is usually analyzed by individual **growth curve modeling** using a multilevel model for change (*see* **Generalized Linear Mixed Models** and [6]). Recently, innovative methodologists [2–4] have shown how the multilevel model for change can be mapped onto the general covariance structure model, such as that implemented in *LISREL* (*see* **Structural Equation Modeling: Software**). This has led to an alternative approach to analyzing change known as *latent growth modeling*. In this chapter, we describe and link these two analogous approaches.

Our presentation uses four waves of data on the reading scores of 1740 Caucasian children from the *Early Childhood Longitudinal Study* (*ECLS-K*; [5]). Children's reading ability was measured in the Fall and Spring of kindergarten and first grade – we assume that test administrations were six months apart, with time measured from entry into kindergarten. Thus, in our analyses, predictor t – representing time – has values 0, 0.5, 1.0 and 1.5 years. Finally, we know the child's gender (*FEMALE*: boy = 0, girl = 1), which we treat as a time-invariant predictor of change[1].

Introducing Individual Growth Modeling

In Figure 1, we display empirical reading records for ten children selected from the larger dataset. In the top *left* panel is the growth record of child #15013, a boy, with observed reading score on the ordinate and time on the abscissa. Reading scores are represented by a '+' symbol and are connected by a smooth freehand curve summarizing the change trajectory. Clearly, this boy's reading ability improves during kindergarten and first grade. In the top *right* panel, we display similar smoothed change trajectories for all ten children (dashed trajectories for boys, solid for girls, plotting symbols omitted to reduce clutter). Notice the dramatic changes in children's observed reading scores over time, and how disparate they are from child to child. The complexity of the collection, and because true reading ability is obscured by measurement error, makes it hard to draw defensible conclusions about gender differences. However, perhaps the girls' trajectories do occupy higher elevations than those of the boys, on average.

Another feature present in the reading trajectories in the top two panels of Figure 1 is the apparent acceleration of the observed trajectories between Fall and Spring of first grade. Most children exhibit moderate growth in reading over the first three waves, but their scores increase dramatically over the last time period. The score of child #15013, for instance, rises modestly between waves 1 and 2 (20 to 28 points), modestly again between waves 2 and 3 (28 to 39 points), and then rapidly (to 66 points) by the fourth wave. Because of this nonlinearity, which was also evident in the entire sample, we transformed children's reading scores before further analysis (Singer & Willett [6, Chapter 6] comment on how to choose an appropriate transformation). We used a natural log transformation in order to 'pull down' the top end of the change trajectory disproportionally, thereby linearizing the accelerating raw trajectory.

In the lower panels of Figure 1, we redisplay the data in the newly transformed logarithmic world. The *log-reading* trajectory of child #15013 is now approximately linear in time, with positive slope. To dramatize this, we have superimposed a *linear* trend line on the transformed plot by simply regressing the log-reading scores on time using ordinary least squares regression (OLS) analysis for that child (*see* **Least Squares Estimation**; **Multiple Linear Regression**). This trend line has a positive slope, indicating that the log-reading score increases during kindergarten and first grade. In the lower right panel, we display OLS-fitted linear trajectories for all ten children in the subsample and reveal the heterogeneity in change that remains across children (albeit change in log-reading score). In subsequent analyses, we model change in the log-reading scores as a linear function of time.

The individual change trajectory can be described by a 'within-person' or 'level-1' individual growth model ([6], Ch. 3). For instance, here we hypothesize

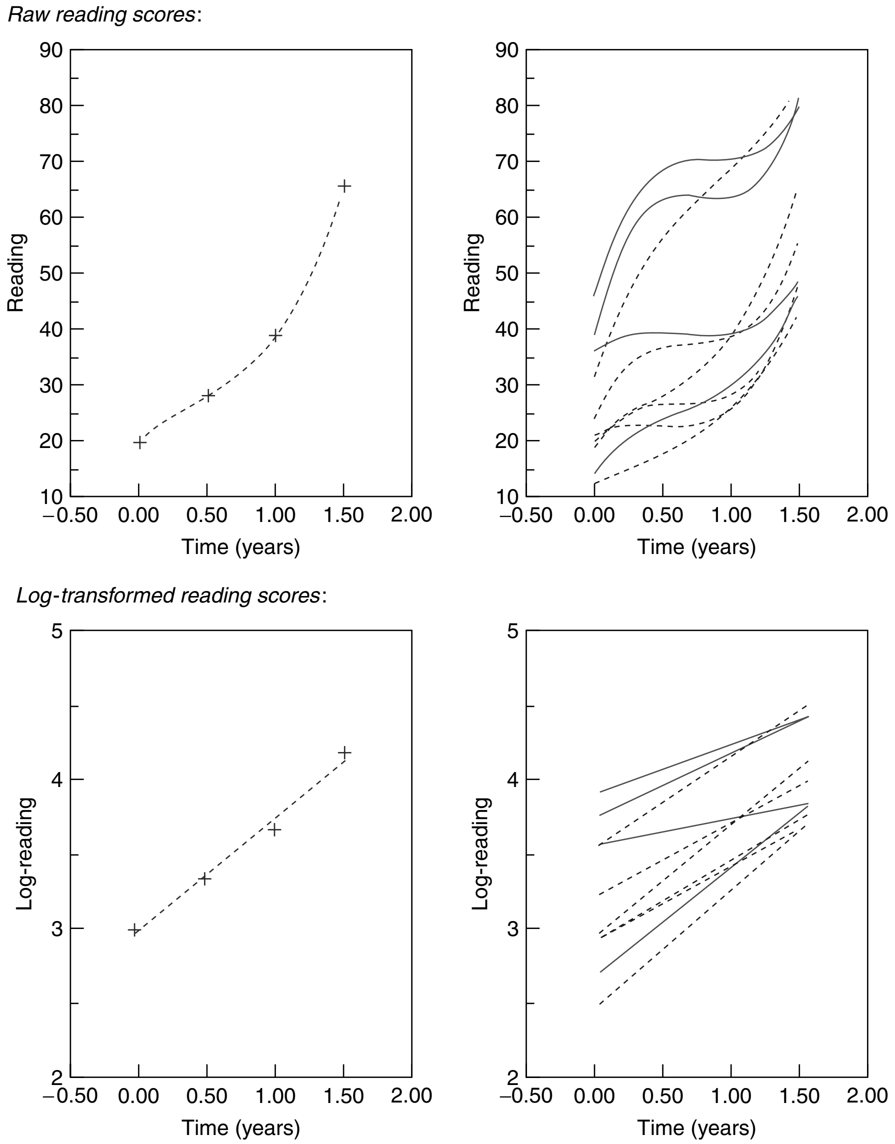


Figure 1 Observed raw and transformed trajectories of reading score over kindergarten and first grade for ten children (boys = dashed; girls = solid). *Top panel*: (a) raw reading score versus time for child #15013, with observed data points (+'s) connected by a smoothed freehand trajectory, (b) smoothed freehand trajectories for all 10 children. *Bottom panel*: (a) log-reading score versus time for child #15013, with an OLS-estimated linear change trajectory, (b) OLS-estimated linear trajectories for all children

that the log-reading score, Y_{ij}, of child i on occasion j is a linear function of time, t:

$$Y_{ij} = \{\pi_{0i} + \pi_{1i}t_j\} + \varepsilon_{ij}, \tag{1}$$

where $i = 1, 2, \ldots, 1740$ and $j = 1, 2, 3, 4$ (with, as noted earlier, $t_1 = 0, t_2 = 0.5, t_3 = 1.0$ and $t_4 = 1.5$ years, respectively). We have bracketed the *systematic* part of the model to separate the orderly temporal dependence from the random errors, ε_{ij}, that accrue on each measurement occasion. Within the brackets, you will find the *individual growth parameters*, π_{0i} and π_{1i}:

- π_{0i} is the *intercept* parameter, describing the child's true 'initial' log-reading score on entry into kindergarten (because entry into kindergarten has been defined as the origin of time).
- π_{1i} is the *slope* ('rate of change') parameter, describing the child's true annual rate of change in log-reading score over time. If π_{1i} is positive, true log-reading score increases with time.

If the model is correctly specified, the individual growth parameters capture the defining features of the log-reading trajectory for child i. Of course, in specifying such models, you need not choose a linear specification – many shapes of trajectory are available, and the particular one that you choose should depend on your theory of change and on your inspection of the data [6, Chapter 6].

One assumption built deeply into individual growth modeling is that, while every child's change trajectory has the same functional form (here, linear in time), different children may have different values of the individual growth parameters. Children may differ in intercept (some have low log-reading ability on entry into kindergarten, others are higher) and in slope (some children change more rapidly with time, others less rapidly). Such heterogeneity is evident in the right-hand lower panel of Figure 1.

We have coded the trajectories in the right-hand panels of Figure 1 by child gender. Displays like these help to reveal systematic differences in change trajectory from child to child, and help you to assess whether interindividual variation in change is related to individual characteristics, like gender. Such 'level-2' questions – about the effect of predictors of change – translate into questions about 'between-person' relationships among the individual growth parameters and predictors like gender. Inspecting the right-hand lower panel of Figure 1, for instance, you can ask whether boys and girls differ in their initial scores (do the intercepts differ by gender?) or in the rates at which their scores change (do the slopes differ by gender?).

Analytically, we can handle this notion in a second 'between-person' or 'level-2' statistical model to represent interindividual differences in change. In the level-2 model, we express how we believe the individual growth parameters (standing in place of the full trajectory) depend on predictors of change. For example, we could investigate the impact of child gender on the log-reading trajectory by positing the following pair of simultaneous level-2 statistical models:

$$\begin{aligned}\pi_{0i} &= \gamma_{00} + \gamma_{01}\,\mathit{FEMALE}_i + \zeta_{0i}\\ \pi_{1i} &= \gamma_{10} + \gamma_{11}\,\mathit{FEMALE}_i + \zeta_{1i},\end{aligned} \tag{2}$$

where the level-2 residuals, ζ_{0i} and ζ_{1i}, represent those portions of the individual growth parameters that are 'unexplained' by the selected predictor of change, *FEMALE*. In this model, the γ coefficients are known as the 'fixed effects' and summarize the population relationship between the individual growth parameters and the predictor. They can be interpreted like regular regression coefficients. For instance, if the initial log-reading ability of girls is higher than boys (i.e., if girls have larger values of π_{0i}, on average) then γ_{01} will be positive (since $\mathit{FEMALE} = 1$, for girls). If girls have higher annual rates of change (i.e., if girls have larger values of π_{1i}, on average), then γ_{11} will be positive. Together, the level-1 and level-2 models in (1) and (2) make up the multilevel model for change ([6], Ch. 3).

Researchers investigating change must fit the multilevel model for change to their longitudinal data. Many methods are available for doing this (see [6], Chs. 2 and 3), the simplest of which is exploratory, as in Figure 1. To conduct data-analyses efficiently, the level-1 and level-2 models are usually fitted *simultaneously* using procedures now widely available in major statistical packages. The models can also be fitted using **covariance structure analysis**, as we now describe.

Latent Growth Modeling

Here, we introduce latent growth modeling by showing how the multilevel model for change can be mapped onto the general covariance structure model. Once the mapping is complete, all parameters of the multilevel model for change can be estimated by fitting the companion covariance structure model using standard covariance structure analysis (CSA) software, such as AMOS, LISREL, EQS, MPLUS, etc. (*see* **Structural Equation Modeling: Software**).

To conduct latent growth analyses, we lay out our data in multivariate format, in which there is a single row in the dataset for each person, with multiple (*multi-*) variables (*-variate*) containing the time-varying information, arrayed horizontally. With four waves of data, multivariate format requires four columns to record each child's growth record, each column associated with a measurement occasion. Any time-invariant predictor of change, like child gender, also has its own column in the dataset. Multivariate formatting is not typical in longitudinal data analysis (which usually requires a 'person-period' or 'univariate' format), but is required here because of the nature of covariance structure analysis. As its name implies, CSA is an analysis of covariance structure in which, as an initial step, a sample covariance matrix (and mean vector) is estimated to summarize the associations among (and levels of) selected variables, including the multiple measures of the outcome across the several measurement occasions. The data-analyst then specifies statistical models appropriate for the research hypotheses, and the mathematical implications of these hypotheses for the structure of the underlying population covariance matrix and mean vector are evaluated against their sample estimates. Because latent growth analysis compares sample and predicted covariance matrices (and mean vectors), the data must be formatted to support the estimation of covariance matrices (and mean vectors) – in other words, in a multivariate format.

Note, finally, that there is no unique column in the multivariate dataset to record time. In our multivariate format dataset, values in the outcome variable's first column were measured at the start of kindergarten, values in the second column were measured at the beginning of spring in kindergarten, etc. The time values – each corresponding to a particular measurement occasion and to a specific column of outcome values in the dataset – are noted by the analyst and programmed directly into the CSA model. It is therefore more convenient to use latent growth modeling to analyze change when panel data are *time-structured* – when everyone has been measured on an identical set of occasions and possesses complete data. Nevertheless, you can use latent growth modeling to analyze panel datasets with limited violations of time-structuring, by regrouping the full sample into subgroups who share identical time-structured profiles and then analyzing these subgroups simultaneously with CSA multigroup analysis (*see* **Factor Analysis: Multiple Groups**).

Mapping the Level-1 Model for Individual Change onto the CSA Y-measurement Model

In (1), we specified that the child's log-reading score, Y_{ij}, depended linearly on time, measured from kindergarten entry. Here, for clarity, we retain symbols t_1 through t_4 to represent the measurement timing but you should remember that each of these time symbols has a known value (0, 0.5, 1.0, and 1.5 years, respectively) that is used when the model is fitted. By substituting into the individual growth model, we can create equations for the value of the outcome on each occasion for child i:

$$\begin{aligned} Y_{i1} &= \pi_{0i} + \pi_{1i}t_1 + \varepsilon_{i1} \\ Y_{i2} &= \pi_{0i} + \pi_{1i}t_2 + \varepsilon_{i2} \\ Y_{i3} &= \pi_{0i} + \pi_{1i}t_3 + \varepsilon_{i3} \\ Y_{i4} &= \pi_{0i} + \pi_{1i}t_4 + \varepsilon_{i4} \end{aligned} \tag{3}$$

that can easily be rewritten in simple matrix form, as follows:

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ 1 & t_4 \end{bmatrix} \begin{bmatrix} \pi_{0i} \\ \pi_{1i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{bmatrix}. \tag{4}$$

While this matrix equation is unlike the representation in (1), it says exactly the same thing – that observed values of the outcome, Y, are related to the times (t_1, t_2, t_3, and t_4), to the individual growth parameters (π_{0i} and π_{1i}), and to the measurement errors ($\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}$, and ε_{i4}). The only difference between (4) and (1) is that all values of the outcome and of time, and all parameters and time-specific residuals, are arrayed neatly as vectors and matrices. (Don't be diverted by the strange vector of zeros introduced immediately to the right of the equals sign – it makes no difference to the *meaning* of the equation, but it will help our subsequent mapping of the multilevel model for change onto the general CSA model).

In fact, the new growth model representation in (4) maps straightforwardly onto the CSA *Y-Measurement Model*, which, in standard *LISREL* notation, is

$$\mathbf{Y} = \boldsymbol{\tau}_y + \boldsymbol{\Lambda}_y\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{5}$$

where $\mathbf{Y}$ is a vector of observed scores, $\boldsymbol{\tau}_y$ is a vector intended to contain the population means of $\mathbf{Y}$, $\boldsymbol{\Lambda}_y$ is a matrix of factor loadings, $\boldsymbol{\eta}$ is a vector of latent (endogenous) constructs, and $\boldsymbol{\varepsilon}$ is a vector of residuals[2]. Notice that the new matrix representation of the individual growth model in (4) matches the

CSA Y-Measurement Model in (5) providing that the observed and latent score vectors are set to:

$$\mathbf{Y} = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{bmatrix}, \quad \boldsymbol{\eta} = \begin{bmatrix} \pi_{0i} \\ \pi_{1i} \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{bmatrix} \qquad (6)$$

and providing that parameter vector $\boldsymbol{\tau}_y$ and loading matrix $\boldsymbol{\Lambda}_y$ are specified as containing the following constants and known times:

$$\boldsymbol{\tau}_y = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{\Lambda}_y = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ 1 & t_4 \end{bmatrix}. \qquad (7)$$

Check this by substituting from (6) and (7) into (5) and multiplying out – you will conclude that, with this specification of score vectors and parameter matrices, the CSA Y-Measurement Model can act as, or contain, the individual growth trajectory from the multilevel model for change.

Notice that (1), (3), (4), (5), and (6) all permit the measurement errors to participate in the individual growth process. They state that level-1 residual ε_{i1} disturbs the true status of the ith child on the first measurement occasion, ε_{i2} on the second occasion, ε_{i3} on the third, and so on. However, so far, we have made no claims about the underlying distribution from which the errors are drawn. Are the errors normally distributed, homoscedastic, and independent over time within-person? Are they heteroscedastic or auto-correlated? Now that the individual change trajectory is embedded in the Y-Measurement Model, we can easily account for level-1 error covariance structure because, under the usual CSA assumption of a multivariate normal distribution for the errors, we can specify the CSA parameter matrix $\boldsymbol{\Theta}_\varepsilon$ to contain hypotheses about the covariance matrix of $\boldsymbol{\varepsilon}$. In an analysis of change, we usually compare nested models with alternative error structures to identify which error structure is optimal. Here, we assume that level-1 errors are distributed normally, independently, and heteroscedastically over time within-person:[3]

$$\boldsymbol{\Theta}_\varepsilon = \begin{bmatrix} {\sigma_{\varepsilon_1}}^2 & 0 & 0 & 0 \\ 0 & {\sigma_{\varepsilon_2}}^2 & 0 & 0 \\ 0 & 0 & {\sigma_{\varepsilon_3}}^2 & 0 \\ 0 & 0 & 0 & {\sigma_{\varepsilon_4}}^2 \end{bmatrix}. \qquad (8)$$

Ultimately, we estimate all level-1 variance components on the diagonal of $\boldsymbol{\Theta}_\varepsilon$ and reveal the action of measurement error on reading score on each occasion.

The key point is that judicious specification of CSA score and parameter matrices forces the level-1 individual change trajectory into the Y-Measurement Model in a companion covariance structure analysis. Notice that, unlike more typical covariance structure analyses, for example, confirmatory factor analysis (*see* **Factor Analysis: Confirmatory**) the $\boldsymbol{\Lambda}_\mathbf{y}$ matrix in (7) is entirely specified as a set of known constants and times rather than as unknown latent factor loadings to be estimated. Using the Y-Measurement Model to represent individual change in this way 'forces' the individual-level growth parameters, π_{0i} and π_{1i}, into the endogenous construct vector $\boldsymbol{\eta}$, creating what is known as the *latent growth vector*, $\boldsymbol{\eta}$. This notion – that the CSA $\boldsymbol{\eta}$-vector can be *forced* to contain the individual growth parameters – is critical in latent growth modeling, because it suggests that level-2 interindividual variation in change can be modeled in the CSA Structural Model, as we now show.

Mapping the Level-2 Model for Interindividual Differences in Change onto the CSA Structural Model

In an analysis of change, at level-2, we ask whether interindividual heterogeneity in change can be predicted by other variables, such as features of the individual's background and treatment. For instance, in our data-example, we can ask whether between-person heterogeneity in the log-reading trajectories depends on the child's gender. Within the growth-modeling framework, this means that we must check whether the individual growth parameters – the true intercept and slope standing in place of the log-reading trajectories – are related to gender. Our analysis therefore asks: Does initial log-reading ability differ for boys and girls? Does the annual rate at which log-reading ability changes depend upon gender? In latent growth modeling, level-2 questions like these, which concern the distribution of the individual growth parameters across individuals and their relationship to predictors, are addressed by specifying a CSA *Structural Model*. Why? Because it is in the CSA structural model that the vector of unknown endogenous constructs $\boldsymbol{\eta}$ – which now contains the all-important individual growth parameters, π_{0i} and π_{1i} – is hypothesized to vary across people.

Recall that the CSA Structural Model stipulates that endogenous construct vector $\boldsymbol{\eta}$ is potentially related to both itself and to exogenous constructs $\boldsymbol{\xi}$ by the following model:

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta}, \tag{9}$$

where $\boldsymbol{\alpha}$ is a vector of intercept parameters, $\boldsymbol{\Gamma}$ is a matrix of regression coefficients that relate exogenous predictors $\boldsymbol{\xi}$ to outcomes $\boldsymbol{\eta}$, $\mathbf{B}$ is a matrix of regression coefficients that permit elements of the endogenous construct vector $\boldsymbol{\eta}$ to predict each other, and $\boldsymbol{\zeta}$ is a vector of residuals. In a covariance structure analysis, we fit this model to data, simultaneously with the earlier measurement model, and estimate parameters $\boldsymbol{\alpha}$, $\boldsymbol{\Gamma}$ and $\mathbf{B}$. The rationale behind latent growth modeling argues that, by structuring (9) appropriately, we can force parameter matrices $\boldsymbol{\alpha}$, $\boldsymbol{\Gamma}$ and $\mathbf{B}$ to contain the fixed effects central to the multilevel modeling of change.

So, what to do? Inspection of the model for systematic interindividual differences in change in (2) suggests that the level-2 component of the multilevel model for change can be reformatted in matrix form, as follows:

$$\begin{bmatrix} \pi_{0i} \\ \pi_{1i} \end{bmatrix} = \begin{bmatrix} \gamma_{00} \\ \gamma_{10} \end{bmatrix} + \begin{bmatrix} \gamma_{01} \\ \gamma_{11} \end{bmatrix} [FEMALE_i] + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \pi_{0i} \\ \pi_{1i} \end{bmatrix} + \begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix}, \tag{10}$$

which is identical to the CSA Structural Model in (9), providing that we force the elements of the CSA $\mathbf{B}$ parameter matrix to be zero throughout:

$$\mathbf{B} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{11}$$

and that we permit the $\boldsymbol{\alpha}$ vector and the $\boldsymbol{\Gamma}$ matrix to be free to contain the fixed-effects parameters from the multilevel model for change:

$$\boldsymbol{\alpha} = \begin{bmatrix} \gamma_{00} \\ \gamma_{10} \end{bmatrix}, \quad \boldsymbol{\Gamma} = \begin{bmatrix} \gamma_{01} \\ \gamma_{11} \end{bmatrix} \tag{12}$$

and providing we can force the potential predictor of change – child gender – into the CSA exogenous construct vector, $\boldsymbol{\xi}$. In this new level-2 specification of the structural model, the latent intercept vector, $\boldsymbol{\alpha}$, contains the level-2 fixed-effects parameters γ_{00} and γ_{10}, defined earlier as the population intercept and slope of the log-reading trajectory for boys (when FEMALE $= 0$). The $\boldsymbol{\Gamma}$ matrix contains the level-2 fixed-effects parameters γ_{01} and γ_{11}, representing increments to the population average intercept and slope for girls, respectively. By fitting this CSA model to data, we can estimate all four fixed effects.

When a time-invariant predictor like *FEMALE* is present in the structural model, the elements of the latent residual vector $\boldsymbol{\zeta}$ in (10) represent those portions of true intercept and true slope that are *unrelated* to the predictor of change – the 'adjusted' values of true intercept and slope, with the linear effect of child gender partialled out. In a covariance structure analysis of the multilevel model for change, we assume that latent residual vector $\boldsymbol{\zeta}$ is distributed normally with zero mean vector and covariance matrix $\boldsymbol{\Psi}$,

$$\boldsymbol{\Psi} = \text{Cov}\{\boldsymbol{\zeta}\} = \begin{bmatrix} \sigma^2_{\zeta_0} & \sigma_{\zeta_{01}} \\ \sigma_{\zeta_{10}} & \sigma^2_{\zeta_1} \end{bmatrix}, \tag{13}$$

which contains the residual (partial) variances and covariance of true intercept and slope, controlling for the predictor of change, *FEMALE*. We estimate these level-2 variance components in any analysis of change.

But there is one missing link that needs resolving before we can proceed. How is the hypothesized predictor of change, *FEMALE*, loaded into the exogenous construct vector, $\boldsymbol{\xi}$? This is easily achieved via the so-far-unused CSA *X-Measurement Model*. And, in the current analysis, the process is disarmingly simple because there is only a single infallible predictor of change, child gender. So, in this case, while it may seem a little weird, the specification of the X-Measurement Model derives from a tautology:

$$FEMALE_i = (0) + (1)(FEMALE_i) + (0). \tag{14}$$

This, while not affecting predictor *FEMALE*, facilitates comparison with the CSA X-Measurement Model:

$$\mathbf{X} = \boldsymbol{\tau}_x + \boldsymbol{\Lambda}_x\boldsymbol{\xi} + \boldsymbol{\delta}. \tag{15}$$

By comparing (14) and (15), you can see that the gender predictor can be incorporated into the analysis by specifying an X-Measurement Model in which:

- Exogenous score vector $\mathbf{X}$ contains one element, the gender predictor, *FEMALE*, itself.
- The $\mathbf{X}$-measurement error vector, $\boldsymbol{\delta}$, contains a single element whose value is fixed at zero,

embodying the assumption that gender is measured infallibly (with ‘zero’ error).

- The $\boldsymbol{\tau}_x$ mean vector contains a single element whose value is fixed at zero. This forces the mean of *FEMALE* (which would reside in $\boldsymbol{\tau}_x$ if the latter were not fixed to zero) into the CSA latent mean vector, κ, which contains the mean of the exogenous construct, $\boldsymbol{\xi}$, in the general CSA model.
- The matrix of exogenous latent factor loadings $\boldsymbol{\Lambda}_x$ contains a single element whose value is fixed at 1. This forces the metrics of the exogenous construct and its indicator to be identical.

Thus, by specifying a CSA X-Measurement Model in which the score vectors are

$$\mathbf{X} = [FEMALE_i], \quad \boldsymbol{\delta} = [\,0\,] \tag{16}$$

and the parameter matrices are fixed at:

$$\boldsymbol{\tau}_{\mathrm{x}} = [\,0\,], \quad \boldsymbol{\Lambda}_{\mathrm{x}} = [\,1\,], \tag{17}$$

we can make the CSA exogenous construct $\boldsymbol{\xi}$ represent child gender. And, since we know that exogenous construct $\boldsymbol{\xi}$ is a predictor in the CSA Structural Model, we have succeeded in inserting the predictor of change, child gender, into the model for interindividual differences in change. As a final consequence of (14) through (17), the population mean of the predictor of change appears as the sole element of the CSA exogenous construct mean vector, $\boldsymbol{\kappa}$:

$$\boldsymbol{\kappa} = \text{Mean}\{\xi\} = [\mu_{FEMALE}] \tag{18}$$

and the population variance of the predictor of change appears as the sole element of CSA exogenous construct covariance matrix $\boldsymbol{\Phi}$:

$$\boldsymbol{\Phi} = \text{Cov}\{\xi\} = \left[\sigma^2_{FEMALE}\right]. \tag{19}$$

Both of these parameter matrices are estimated when the model is fitted to data. And, although we do not demonstrate it here, the X-Measurement Model in (14) through (19) can be reconfigured to accommodate multiple time-invariant predictors of change, and even several indicators of each predictor construct if available. This is achieved by expanding the exogenous indicator and constructing score vectors to include sufficient elements to contain the new indicators and constructs, and the parameter matrix $\boldsymbol{\Lambda}_x$ is expanded to include suitable loadings (Willett and Singer [6; Chapter 8] give an example with multiple predictors).

Table 1 Trajectory of change in the logarithm of children’s reading score over four measurement occasions during kindergarten and first grade, by child gender. Parameter estimates, approximate *P* values, and goodness of fit statistics from the multilevel model for change, obtained with latent growth modeling (n = 1740)

Effect	Parameter	Estimate
Fixed effects:		
	γ_{00}	3.1700***
	γ_{10}	0.5828***
	γ_{01}	0.0732***
	γ_{11}	−0.0096
Variance components:		
	$\sigma^2_{\varepsilon_1}$	0.0219***
	$\sigma^2_{\varepsilon_2}$	0.0228***
	$\sigma^2_{\varepsilon_3}$	0.0208***
	$\sigma^2_{\varepsilon_4}$	0.0077***
	$\sigma^2_{\zeta_0}$	0.0896***
	$\sigma^2_{\zeta_1}$	0.0140***
	$\sigma_{\zeta_0\zeta_1}$	−0.0223***
Goodness of fit:		
	χ^2	1414.25***
	Df	7

$\sim$= p < .10, $* = p < .05$, $** = p < .01$, $*** = p < .00$.

So, the CSA version of the multilevel model for change – now called the *latent growth model* – is complete. It consists of the CSA *X-Measurement, Y-Measurement*, and *Structural Models*, defined in (14) through (19), (4) through (8), and (9) through (13), respectively and is displayed as a path model in Figure 2. In the figure, by fixing the loadings associated with the outcome measurements to their constant and temporal values, we emphasize how the endogenous constructs were forced to become the individual growth parameters, which are then available for prediction by the hypothesized predictor of change. We fitted the latent growth model in (4) through (14) to our reading data on the full sample of 1740 children using LISREL (see Appendix I). Table 1 presents full maximum-likelihood (FML) estimates of all relevant parameters from latent regression-weight matrix $\boldsymbol{\Gamma}$ and parameter matrices $\boldsymbol{\Phi}, \boldsymbol{\alpha}$, and $\boldsymbol{\Psi}$.

The estimated level-2 fixed effects are in the first four rows of Table 1. The first and second rows contain estimates of parameters γ_{00} and γ_{10}, representing true initial log-reading ability ($\hat{\gamma}_{00} =$

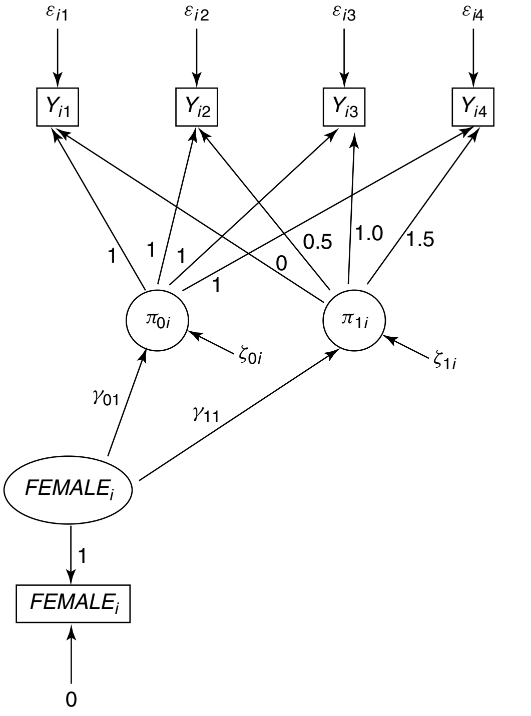


Figure 2 Path diagram for the hypothesized latent growth in reading score. Rectangles represent observed indicators, circles represent latent constructs, and arrows and their associated parameters indicate hypothesized relationships

3.170, p < .001) and true annual rate of change in log-reading ability ($\hat{\gamma}_{10} = 0.583$, p < .001) for boys (for whom *FEMALE* = 0). Anti-logging tells us that, on average, boys: (a) begin kindergarten with an average reading ability of 23.8 (= $e^{3.1700}$), and (b) increase their reading ability by 79% (= $100(e^{0.5828} - 1)$) per year. The third and fourth rows contain the estimated latent regression coefficients γ_{01} (0.073, p < .001) and γ_{11} (−0.010, *p* > .10), which capture differences in change trajectories between boys and girls. Girls have a higher initial level of 3.243 (= 3.170 + 0.073) of log-reading ability, whose anti-log is 25.6 and a statistically significant couple of points higher than the boys. However, we cannot reject the null hypothesis associated with γ_{11} (−0.010, p > .10) so, although the estimated annual rate of increase in log-reading ability for girls is 0.572 (= 0.5828 − 0.0096), a little smaller than boys, this difference is not statistically significant. Nonetheless, anti-logging, we find that girls' reading ability increases by about 78% (= $100(e^{0.5828} - 1)$) per year, on average. We display fitted log-reading and reading trajectories for prototypical boys and girls in Figure 3 – once de-transformed, the trajectories are curvilinear and display the acceleration we noted earlier in the raw data.

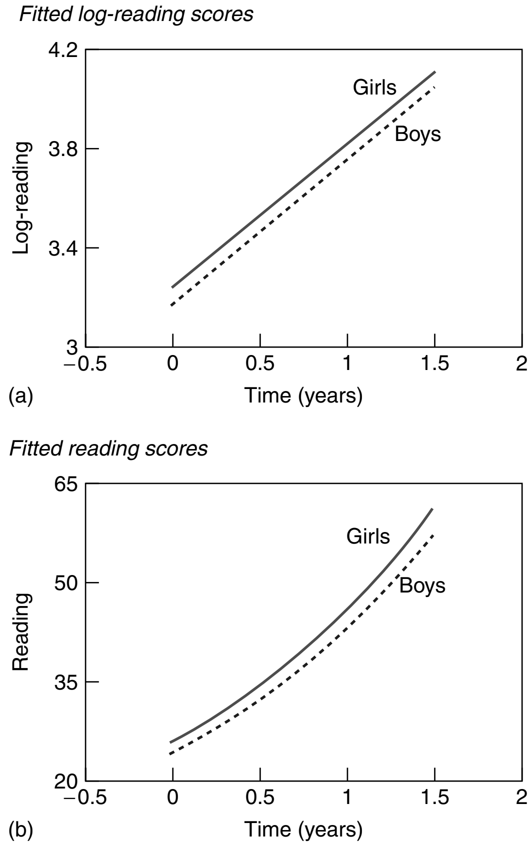


Figure 3 Fitted log-reading and reading trajectories over kindergarten and first grade for prototypical Caucasian children, by gender

Next, examine the random effects. The fifth through eighth rows of Table 1 contain estimated level-1 error variances, one per occasion, describing the measurement fallibility in log-reading score over time. Their estimated values are 0.022, 0.023, 0.021, and 0.008, respectively, showing considerable homoscedasticity over the first three occasions but measurement error variance decreases markedly in the spring of first grade. The tenth through twelfth rows of Table 1 contain the estimated level-2 variance components, representing estimated partial (residual) variances and partial covariance of true initial status and rate of change, after controlling for child gender (*see* **Partial Correlation Coefficients**). We reject the null hypothesis associated with each variance component, and conclude that there is predictable true variation remaining in both initial status and rate of change.

Conclusion: Extending the Latent Growth-Modeling Framework

In this chapter, we have shown how a latent growth-modeling approach to analyzing change is created by mapping the multilevel model for change onto the general CSA model. The basic latent growth modeling approach that we have described can be extended in many important ways:

- **You can include any number of waves of longitudinal data**, by simply increasing the number of rows in the relevant score vectors. Including more waves generally leads to greater precision for the estimation of the individual growth trajectories and greater reliability for measuring change.
- **You need not space the occasions of measurement equally**, although it is most convenient if everyone in the sample is measured on the *same set* of irregularly spaced occasions. However, if they are not, then latent growth modeling can still be conducted by first dividing the sample into subgroups of individuals with identical temporal profiles and using multigroup analysis to fit the multilevel model for change simultaneously in all subgroups.
- **You can specify curvilinear individual change**. Latent growth modeling can accommodate polynomial individual change of any order (provided sufficient waves of data are available), or any other curvilinear change trajectory in which individual status is linear in the growth parameters.
- **You can model the covariance structure of the level-1 measurement errors explicitly**. You need not accept the independence and homoscedasticity assumptions of classical analysis unchecked. Here, we permitted level-1 measurement errors to be heteroscedastic, but other, more general, error covariance structures can be hypothesized and tested.
- **You can model change in several domains simultaneously**, including both exogenous and endogenous domains. You simply extend the empirical growth record and the measurement models to include rows for each wave of data available, in each domain.
- **You can model *intervening effects***, whereby an exogenous predictor may act directly on endogenous change and also indirectly via the influence of intervening factors, each of which may be time-invariant or time varying.

In the end, you must choose your analytic strategy to suit the problems you face. Studies of change can be designed in enormous variety and the multilevel model for change can be specified to account for all manner of trajectories and error structures. But, it is always wise to have more than one way to deal with data – latent growth modeling often offers a flexible alternative to more traditional approaches.

Notes

[1] The dataset is available at `http://gseacademic.harvard.edu/~willetjo/`.

[2] Readers unfamiliar with the general *CSA* model should consult Bollen [1].

[3] Supplementary analyses suggested that this was reasonable.

Appendix I Specimen LISREL Program

```
/*Specify the number of variables
   (indicators) to be read from
   the external data-file of
   raw data*/
   data ni=6
/*Identify the location of the
   external data-file*/
   raw fi = C:\Data\ECLS.dat
/*Label the input variables and
   select those to be analyzed*/
   label
   id Y1 Y2 Y3 Y4 FEMALE
   select
   2 3 4 5 6 /
/*Specify the hypothesized covariance
   structure model*/
   model   ny=4 ne=2 ty=ze ly=fu,
             fi te=di,fi          c
           nx=1 nk=1 lx=fu,fi tx=fr
```

```
            td=ze ph=sy,fr       c
          al=fr ga=fu,fr be=ze
            ps=sy,fr
/*Label the individual growth
   parameters as endogenous
   constructs (eta's)*/
   le
   pi0 pi1
/*Label the predictor of change as
   an exogenous construct (ksi) */
   lk
   FEMALE
/*Enter the required ``1's'' and
   measurement times into the
   Lambda-Y matrix*/
   va 1 ly(1,1) ly(2,1) ly(3,1)
     ly(4,1)
   va 0.0 ly(1,2)
   va 0.5 ly(2,2)
   va 1.0 ly(3,2)
   va 1.5 ly(4,2)
/*Enter the required scaling factor
   ``1'' into the Lambda-X matrix*/
   va 1.0 lx(1,1)
/*Free up the level-1 residual
   variances to be estimated*/
   fr te(1,1) te(2,2) te(3,3) te(4,4)
/*Request data-analytic output to 5
   decimal places*/
   ou nd=5
```

References

[1] Bollen, K. (1989). *Structural Equations with Latent Variables*, Wiley, New York.

[2] McArdle, J.J. (1986). Dynamic but structural equation modeling of repeated measures data, in *Handbook of Multivariate Experimental Psychology*, Vol. 2. J.R. Nesselrode & R.B. Cattell, eds, Plenum Press, New York, pp. 561–614.

[3] Meredith, W. & Tisak, J. (1990). Latent curve analysis, *Psychometrika* **55**, 107–122.

[4] Muthen, B.O. (1991). Analysis of longitudinal data using latent variable models with varying parameters, in *Best Methods for the Analysis of Change: Recent Advances, Unanswered Questions, Future Directions*, L.M. Collins & J.L. Horn, eds, American Psychological Association, Washington, pp. 1–17.

[5] NCES (2002). *Early Childhood Longitudinal Study*. National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education, Washington.

[6] Singer, J.D. & Willett, J.B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*, Oxford University Press, New York.

John B. Willett and Kristen L. Bub

Structural Equation Modeling: Mixture Models

Jeroen K. Vermunt and Jay Magidson

Volume 4, pp. 1922–1927

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Structural Equation Modeling: Mixture Models

Introduction

This article discusses a modeling framework that links two well-known statistical methods: **structural equation modeling** (SEM) and **latent class** or **finite mixture modeling**. This hybrid approach was proposed independently by Arminger and Stein [1], Dolan and Van der Maas [4], and Jedidi, Jagpal and DeSarbo [5]. Here, we refer to this approach as mixture SEM or latent class SEM.

There are two different ways to view mixture SEM. One way is as a refinement of multivariate normal (MVN) mixtures (*see* **Finite Mixture Distributions**), where the within-class covariance matrices are smoothed according to a postulated SEM structure. MVN mixtures have become a popular tool for **cluster analysis** [6, 10], where each cluster corresponds to a latent (unobservable) class. Names that are used when referring to such a use of mixture models are latent profile analysis, mixture-model clustering, model-based clustering, probabilistic clustering, Bayesian classification, and latent class clustering. Mixture SEM restricts the form of such latent class clustering, by subjecting the class-specific mean vectors and covariance matrices to a postulated SEM structure such as a one-factor, a latent-growth, or an autoregressive model. This results in MVN mixtures that are more parsimonious and stable than models with unrestricted covariance structures.

The other way to look at mixture SEM is as an extension to standard SEM, similar to multiple group analysis. However, an important difference between this and standard multiple group analysis, is that in mixture SEM group membership is not observed. By incorporating latent classes into an SEM model, various forms of unobserved heterogeneity can be detected. For example, groups that have identical (unstandardized) factor loadings but different error variances on the items in a factor analysis, or groups that show different patterns of change over time. Dolan and Van der Maas [4] describe a nice application from developmental psychology, in which (as a result of the existence of qualitative development stages) children who do not master certain types of tasks have a mean and covariance structure that differs from the one for children who master the tasks.

Below, we first introduce standard MVN mixtures. Then, we show how the SEM framework can be used to restrict the means and covariances. Subsequently, we discuss parameter estimation, model testing, and software. We end with an empirical example.

Multivariate Normal Mixtures

Let $\mathbf{y}_i$ denote a P-dimensional vector containing the scores for individual i on a set of P observed continuous random variables. Moreover, let K be the number of mixture components, latent classes, or clusters, and π_k the prior probability of belonging to latent class or cluster k or, equivalently, the size of cluster k, where $1 \leq k \leq K$. In a mixture model, it is assumed that the density of $\mathbf{y}_i$, $f(\mathbf{y}_i|\boldsymbol{\theta})$, is a mixture or a weighted sum of K class-specific densities $f_k(\mathbf{y}_i|\boldsymbol{\theta}_k)$ [4, 10]. That is,

$$f(\mathbf{y}_i|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i|\boldsymbol{\theta}_k). \tag{1}$$

Here, $\boldsymbol{\theta}$ denotes the vector containing all unknown parameters, and $\boldsymbol{\theta}_k$ the vector of the unknown parameters of cluster k.

The most common specification for the class-specific densities $f_k(\mathbf{y}_i|\boldsymbol{\theta}_k)$ is multivariate normal (*see* **Catalogue of Probability Density Functions**), which means that the observed variables are assumed to be normally distributed within latent classes, possibly after applying an appropriate nonlinear transformation. Denoting the class-specific mean vector by $\boldsymbol{\mu}_k$, and the class-specific covariance matrix by $\boldsymbol{\Sigma}_k$, we obtain the following class-specific densities:

$$f_k(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-P/2}|\boldsymbol{\Sigma}_k|^{-1/2} \times \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_k)\right\}. \tag{2}$$

In the most general specification, no restrictions are imposed on $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ parameters; that is, the model-based clustering problem involves estimating a separate set of means, variances, and covariances for each latent class. Although in most clustering applications, the main objective is finding classes that differ with respect to their means or locations, in the

MVN mixture model, clusters may also have different shapes.

An unrestricted MVN mixture model with K latent classes contains $(K-1)$ unknown class sizes, $K \cdot P$ class-specific means, $K \cdot P$ class-specific variances and $K \cdot P \cdot (P-1)/2$ class-specific covariances. As the number of indicators and/or the number of latent classes increases, the number of parameters to be estimated may become quite large, especially the number of free parameters in $\mathbf{\Sigma}_k$. Thus, to obtain more parsimony and stability, it is not surprising that restrictions are typically imposed on the class-specific covariance matrices.

Prior to using SEM models to restrict the covariances, a standard approach to reduce the number of parameters is to assume local independence. Local independence means that all within-cluster covariances are equal to zero, or, equivalently, that the covariance matrices, $\mathbf{\Sigma}_k$, are diagonal matrices. Models that are less restrictive than the local independence model can be obtained by fixing some but not all covariances to zero, or, equivalently, by assuming certain pairs of y's to be mutually dependent within latent classes.

Another approach to reduce the number of parameters, is to assume the equality or homogeneity of variance–covariance matrices across latent classes; that is, $\mathbf{\Sigma}_k = \mathbf{\Sigma}$. Such a homogeneous or class-independent error structure yields clusters having the same forms but different locations. This type of constraint is equivalent to the restrictions applied to the covariances in linear discriminant analysis. Note that this between-class equality constraint can be applied in combination with any structure for $\mathbf{\Sigma}$.

Banfield and Raftery [2] proposed reparameterizing the class-specific covariance matrices by an eigenvalue decomposition:

$$\mathbf{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k. \tag{3}$$

The parameter λ_k is a scalar, $\mathbf{D}_k$ is a matrix with eigenvectors, and $\mathbf{A}_k$ is a diagonal matrix whose elements are proportional to the eigenvalues of $\mathbf{\Sigma}_k$. More precisely, $\lambda_k = |\mathbf{\Sigma}_k|^{1/d}$, where d is the number of observed variables, and $\mathbf{A}_k$ is scaled such that $|\mathbf{A}_k| = 1$.

A nice feature of the above decomposition is that each of the three sets of parameters has a geometrical interpretation: λ_k indicates what can be called the volume of cluster k, $\mathbf{D}_k$ its orientation, and $\mathbf{A}_k$ its shape. If we think of a cluster as a clutter of points in a multidimensional space, the volume is the size of the clutter, while the orientation and shape parameters indicate whether the clutter is spherical or ellipsoidal. Thus, restrictions imposed on these matrices can directly be interpreted in terms of the geometrical form of the clusters. Typical restrictions are to assume matrices to be equal across classes, or to have the forms of diagonal or identity matrices [3].

Mixture SEM

As an alternative to simplifying the $\mathbf{\Sigma}_k$ matrices using the eigenvalue decomposition, the mixture SEM approach assumes a **covariance-structure model**. Several authors [1, 4, 5] have proposed using such a mixture specification for dealing with unobserved heterogeneity in SEM. As explained in the introduction, this is equivalent to restricting the within-class mean vectors and covariance matrices by an SEM. One interesting SEM structure for $\mathbf{\Sigma}_k$ that is closely related to the eigenvalue decomposition described above is a factor-analytic model (*see* **Factor Analysis: Exploratory**) [6, 11]. Under the factor-analytic structure, the within-class covariances are given by:

$$\mathbf{\Sigma}_k = \mathbf{\Lambda}_k \mathbf{\Phi}_k \mathbf{\Lambda}'_k + \mathbf{\Theta}_k. \tag{4}$$

Assuming that there are Q factors, $\mathbf{\Lambda}_k$ is a $P \times Q$ matrix with factor loadings, $\mathbf{\Phi}_k$ is a $Q \times Q$ matrix containing the variances of, and the covariances between, the factors, and $\mathbf{\Theta}_k$ is a $P \times P$ diagonal matrix containing the unique variances. Restricted covariance structures are obtained by setting $Q < P$ (for instance, $Q = 1$), equating factor loadings across indicators, or fixing some factor loading to zero. Such specifications make it possible to describe the covariances between the y variables within clusters by means of a small number of parameters.

Alternative formulations can be used to define more general types of SEM models. Here, we use the Lisrel submodel that was also used by Dolan and Van der Maas [4]. Other alternatives are the full Lisrel [5], the RAM [8], or the conditional mean and covariance structure [1] formulations.

In our Lisrel submodel formulation, the SEM for class k consists of the following two (sets of) equations:

$$\mathbf{y}_i = \boldsymbol{\nu}_k + \mathbf{\Lambda}_k \boldsymbol{\eta}_{ik} + \boldsymbol{\varepsilon}_{ik} \tag{5}$$

$$\boldsymbol{\eta}_{ik} = \boldsymbol{\alpha}_k + \mathbf{B}_k\boldsymbol{\eta}_{ik} + \boldsymbol{\varsigma}_{ik}. \tag{6}$$

The first equation concerns the measurement part of the model, in which the observed variables are regressed on the latent factors $\boldsymbol{\eta}_{ik}$. Here, $\boldsymbol{\nu}_k$ is a vector of intercepts, $\boldsymbol{\Lambda}_k$ a matrix with factor loadings and $\boldsymbol{\varepsilon}_{ik}$ a vector with residuals. The second equation is the structural part of the model, the path model for the factors. Vector $\boldsymbol{\alpha}_k$ contains the intercepts, matrix $\mathbf{B}_k$ the path coefficients and vector $\boldsymbol{\varsigma}_{ik}$ the residuals. The implied mean and covariance structures for latent class k are

$$\boldsymbol{\mu}_k = \boldsymbol{\nu}_k + \boldsymbol{\Lambda}_k(\mathbf{I} - \mathbf{B}_k)^{-1}\boldsymbol{\alpha}_k \tag{7}$$

$$\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_k(\mathbf{I} - \mathbf{B}_k)^{-1}\boldsymbol{\Phi}_k(\mathbf{I} - \mathbf{B}'_k)^{-1}\boldsymbol{\Lambda}'_k + \boldsymbol{\Theta}_k, \tag{8}$$

where $\boldsymbol{\Theta}_k$ and $\boldsymbol{\Phi}_k$ denote the covariance matrices of the residuals $\boldsymbol{\varepsilon}_{ik}$ and $\boldsymbol{\varsigma}_{ik}$. These equations show the connection between the SEM parameters, and the parameters of the MVN mixture model.

Covariates

An important extension of the mixture SEM described above is obtained by including covariates to predict class membership, with possible direct effects on the item means. Conceptually, it makes sense to distinguish (endogenous) variables that are used to identify the latent classes, from (exogenous) variables that are used to predict to which cluster an individual belongs.

Using the same basic structure as in 1, this yields the following mixture model:

$$f(\mathbf{y}_i|\mathbf{z}_i, \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(\mathbf{z}_i)\, f_k(\mathbf{y}_i|\boldsymbol{\theta}_k). \tag{9}$$

Here, $\mathbf{z}_i$ denotes person i's covariate values. Alternative terms for the z's are concomitant variables, grouping variables, external variables, exogenous variables, and inputs. To reduce the number of parameters, the probability of belonging to class k given covariate values $\mathbf{z}_i$, $\pi_k(\mathbf{z}_i)$, will generally be restricted by a multinomial logit model; that is, a logit model with 'linear effects' and no higher order interactions.

An even more general specification is obtained by allowing covariates to have direct effects on the indicators, which yields

$$f(\mathbf{y}_i|\mathbf{z}_i, \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k(\mathbf{z}_i)\, f_k(\mathbf{y}_i|\mathbf{z}_i, \boldsymbol{\theta}_k). \tag{10}$$

The conditional means of the y variables are now directly related to the covariates, as proposed by Arminger and Stein [1]. This makes it possible to relax the implicit assumption in the previous specification, that the influence of the z's on the y's goes completely via the latent classes (see, for example, [9]).

Estimation, Testing, and Software

Estimation

The two main estimation methods in mixture SEM and other types of MVN mixture modeling are **maximum likelihood** (ML) and maximum posterior (MAP). The log-likelihood function required in ML and MAP approaches can be derived from the probability density function defining the model. Bayesian MAP estimation involves maximizing the log-posterior distribution, which is the sum of the log-likelihood function and the logs of the priors for the parameters (*see* **Bayesian Statistics**).

Although generally, there is not much difference between ML and MAP estimates, an important advantage of the latter method is that it prevents the occurrence of boundary or terminal solutions: probabilities and variances cannot become zero. With a very small amount of prior information, the parameter estimates are forced to stay within the interior of the parameter space. Typical priors are Dirichlet priors for the latent class probabilities, and inverted-Wishart priors for the covariance matrices. For more details on these priors, see [9].

Most mixture modeling software packages use the EM algorithm, or some modification of it, to find the ML or MAP estimates. In our opinion, the ideal algorithm starts with a number of EM iterations, and when close enough to the final solution, switches to Newton–Raphson. This is a way to combine the advantages of both algorithms – the stability of EM even when far away from the optimum, and the speed of Newton–Raphson when close to the optimum (*see* **Optimization Methods**).

A well-known problem in mixture modeling analysis is the occurrence of local solutions. The best way to prevent ending with a local solution is to use multiple sets of starting values. Some computer programs for mixture modeling have automated the search for good starting values using several sets of random starting values.

When using mixture SEM for clustering, we are not only interested in the estimation of the model parameters, but also in the classification of individual into clusters. This can be based on the posterior class membership probabilities

$$\pi_k(\mathbf{y}_i, \mathbf{z}_i, \boldsymbol{\pi}, \boldsymbol{\theta}) = \frac{\pi_k(\mathbf{z}_i)\ f_k(\mathbf{y}_i|\mathbf{z}_i, \boldsymbol{\theta}_k)}{\sum_{k=1}^{K} \pi_k(\mathbf{z}_i)\ f_k(\mathbf{y}_i|\mathbf{z}_i, \boldsymbol{\theta}_k)}. \qquad (11)$$

The standard classification method is modal allocation, which amounts to assigning each object to the class with the highest posterior probability.

Model Selection

The model selection issue is one of the main research topics in mixture-model clustering. Actually, there are two issues involved: the first concerns the decision about the number of clusters, the second concerns the form of the model, given the number of clusters. For an extended overview on these topics, see [6].

Assumptions with respect to the forms of the clusters, given their number, can be tested using standard likelihood-ratio tests between nested models, for instance, between a model with an unrestricted covariance matrix and a model with a restricted covariance matrix. Wald tests and Lagrange multiplier tests can be used to assess the significance of certain included or excluded terms, respectively. However, these kinds of chi-squared tests cannot be used to determine the number of clusters.

The approach most often used for model selection in mixture modeling is to use information criteria, such as AIC, BIC, and CAIC (**Akaike**, Bayesian, and Consistent Akaike Information Criterion). The most recent development is the use of computationally intensive techniques like parametric bootstrapping [6] and Markov Chain Monte Carlo methods [3] to determine the number of clusters, as well as their forms.

Another approach for evaluating mixture models is based on the uncertainty of classification, or, equivalently, the separation of the clusters. Besides the estimated total number of misclassifications, Goodman–Kruskal lambda, Goodman–Kruskal tau, or entropy-based measures can be used to indicate how well the indicators predict class membership.

Software

Several computer programs are available for estimating the various types of mixture models discussed in this paper. Mplus [7] and Mx [8] are syntax-based programs that can deal with a very general class of mixture SEMs. Mx is somewhat more general in terms of model possible constraints. Latent GOLD [9] is a fully Windows-based program for estimating MVN mixtures with covariates. It can be used to specify restricted covariance structures, including a number of SEM structures such as a one-factor model within blocks of variables, and a compound-symmetry (or random-effects) structure.

An Empirical Example

To illustrate mixture SEM, we use a longitudinal data set made available by Patrick J. Curran at '`http://www.duke.edu/~curran/`'. The variable of interest is a child's reading recognition skill measured at four two-year intervals using the Peabody Individual Achievement Test (PIAT) Reading Recognition subtest. The research question of interest is whether a one-class model with its implicit assumption that a single pattern of reading development holds universally is correct, or whether there are different types of reading recognition trajectories among different latent groups. Besides information on reading recognition, we have information on the child's gender, the mother's age, the child's age, the child's cognitive stimulation at home, and the child's emotional support at home. These variables will be used as covariates. The total sample size is 405, but only 233 children were measured at all assessments. We use all 405 cases in our analysis, assuming that the missing data is missing at random (MAR) (*see* **Dropouts in Longitudinal Data**). For parameter estimation, we used the Latent GOLD and Mx programs.

One-class to three-class models (without covariates) were estimated under five types of SEM structures fitted to the within-class covariance matrices. These SEM structures are local independence (LI), saturated (SA), random effects (RE), autoregressive (AR), and one factor (FA). The BIC values reported in Table 1 indicate that two classes are needed when using a SA, AR, or FA structure.[1] As is typically the case, working with a misspecified covariance structure (here, LI or RE), yields an overestimation of

Table 1 Test results for the child's reading recognition example

Model		log-likelihood	# parameters	*BIC*
A1.	1-class LI	−1977	8	4003
A2.	2-class LI	−1694	17	3490
A3.	3-class LI	−1587	26	3330
B1.	1-class SA	−1595	14	3274
B2.	2-class SA	−1489	29	3151
B3.	3-class SA	−1459	44	3182
C1.	1-class RE	−1667	9	3375
C2.	2-class RE	−1561	19	3237
C3.	3-class RE	−1518	29	3211
D1.	1-class AR	−1611	9	3277
D2.	2-class AR	−1502	19	3118
D3.	3-class AR	−1477	29	3130
E1.	1-class FA	−1611	12	3294
E2.	2-class FA	−1497	25	3144
E3.	3-class FA	−1464	38	3157
F.	D2 + covariates	−1401	27	2964

the number of classes. Based on the BIC criterion, the two-class AR model (Model D2) is the model that is preferred. Note that this model captures the dependence between the time-specific measures with a single path coefficient, since the coefficients associated with the autoregressive component of the model are assumed to be equal for each pair of adjacent time points.

Subsequently, we included the covariates in the model. Child's age was assumed to directly affect the indicators, in order to assure that the encountered trajectories are independent of the child's age at the first occasion. Child's gender, mother's age, child's cognitive stimulation, and child's emotional support were assumed to affect class membership. According to the BIC criterion, this model (Model F) is much better than the model without covariates (Model D2).

According to Model F, Class 1 contains 61% and class 2 39% of the children. The estimated means for class 1 are 2.21, 3.59, 4.51, and 5.22, and for class 2, 3.00, 4.80, 5.81, and 6.67. These results show that class 2 starts at a higher level and grows somewhat faster than class 1. The estimates of the class-specific variances are 0.15, 0.62, 0.90 and 1.31 for class 1, and 0.87, 0.79, 0.94, and 0.76 for class 2. This indicates that the within-class heterogeneity increases dramatically within class 1, while it is quite stable within class 2. The estimated values of the class-specific path coefficients are 1.05 and 0.43, respectively, indicating that even with the incrementing variance, the autocorrelation is larger in latent class 1 than in latent class 2.[2]

The age effects on the indicators are highly significant. As far as the covariate effects on the log-odds of belonging to class 2 instead of class 1 are concerned, only the mother's age is significant. The older the mother, the higher the probability of belonging to latent class 2.

Notes

1. BIC is defined as minus twice the log-likelihood plus $\ln(N)$ times the number of parameters, where N is the sample size (here 450).
2. The autocorrelation is a standardized path coefficient that can be obtained as the product of the unstandardized coefficient and the ratio of the standard deviations of the independent and the dependent variable in the equation concerned. For example, the class 1 autocorrelation between time points 1 and 2 equals $1.05(\sqrt{0}.15/\sqrt{0}.62)$.

References

[1] Arminger, G. & Stein, P. (1997). Finite mixture of covariance structure models with regressors: loglikelihood function, distance estimation, fit indices, and a complex example, *Sociological Methods and Research* **26**, 148–182.

[2] Banfield, J.D. & Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering, *Biometrics* **49**, 803–821.

[3] Bensmail, H., Celeux, G., Raftery, A.E. & Robert, C.P. (1997). Inference in model based clustering, *Statistics and Computing* **7**, 1–10.

[4] Dolan, C.V. & Van der Maas, H.L.J. (1997). Fitting multivariate normal finite mixtures subject to structural equation modeling, *Psychometrika* **63**, 227–253.

[5] Jedidi, K., Jagpal, H.S. & DeSarbo, W.S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity, *Marketing Science* **16**, 39–59.

[6] McLachlan, G.J. & Peel, D. (2000). *Finite Mixture Models*, John Wiley & Sons, New York.

[7] Muthén, B. & Muthén, L. (1998). *Mplus: User's Manual*, Muthén & Muthén, Los Angeles.

[8] Neale, M.C., Boker, S.M., Xie, G. & Maes, H.H. (2002). *Mx: Statistical Modeling*, VCU, Department of Psychiatry, Richmond.

[9] Vermunt, J.K. & Magidson, J. (2000). *Latent GOLD's User's Guide*, Statistical Innovations, Boston.

[10] Vermunt, J.K. & Magidson, J. (2002). Latent class cluster analysis, in *Applied Latent Class Analysis*, J.A. Hagenaars & A.L. McCutcheon, eds, Cambridge University Press, Cambridge, pp. 89–106.

[11] Yung, Y.F. (1997). Finite mixtures in confirmatory factor-analysis models, *Psychometrika* **62**, 297–330.

JEROEN K. VERMUNT AND JAY MAGIDSON

Structural Equation Modeling: Multilevel

Fiona Steele

Volume 4, pp. 1927–1931

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Structural Equation Modeling: Multilevel

Multilevel Factor Models

Suppose that for each individual i we observe a set of R continuous responses $\{y_{ri} : r = 1, \dots, R\}$. In standard **factor analysis**, we assume that the pairwise correlations between the responses are wholly explained by their mutual dependency on one or more underlying factors, also called **latent variables**. If there is only one such factor, the factor model may be written:

$$y_{ri} = \alpha_r + \lambda_r \eta_i + e_{ri}, \qquad (1)$$

where α_r is the grand mean for response r, η_i is a factor with loading λ_r for response r, and e_{ri} is a residual. We assume that both the factor and the residuals are normally distributed. In addition, we assume that the residuals are uncorrelated, which follows from the assumption that the correlation between the responses is due to their dependency on the factor; conditional on this factor, the responses are independent.

A further assumption of model (1) is that the y's are independent across individuals. Often, however, individuals will be clustered in some way, for example, in areas or institutions, and responses for individuals in the same cluster are potentially correlated. In the context of regression analysis, multilevel or **hierarchical models** have been developed to account for within-cluster correlation and to explore between-cluster variation. Multilevel models include cluster-level residuals or random effects that represent unobserved variables operating at the cluster level; conditional on the random effects, individuals' responses are assumed to be independent (*see* **Generalized Linear Mixed Models**). The factor model has also been extended to handle clustering, leading to a multilevel factor model (see [3, 5]). In the multilevel factor model, in addition to having residuals at both the individual and cluster level as in a multilevel regression model, there may be factors at both levels. Suppose, for example, that academic ability is assessed using a series of tests. An individual's score on each of these tests is likely to depend on his overall ability (represented by an individual-level factor) and the ability of children in the same school (a school-level factor). A multilevel extension of (1) with a single factor at the cluster level may be written:

$$y_{rij} = \alpha_r + \lambda_r^{(1)} \eta_{ij}^{(1)} + \lambda_r^{(2)} \eta_j^{(2)} + u_{rj} + e_{rij}, \qquad (2)$$

where y_{rij} is response r for individual $i (i = 1, \dots, n_j)$ in cluster $j (j = 1, \dots, J)$, $\eta_{ij}^{(1)}$ and $\eta_j^{(2)}$ are factors at the individual and cluster levels (levels 1 and 2 respectively) with loadings $\lambda_r^{(1)}$ and $\lambda_r^{(2)}$, and u_{rj} and e_{rij} are residuals at the individual and cluster levels. We assume $\eta_{ij}^{(1)} \sim N(0, \sigma^2_{\eta(1)})$, $\eta_j^{(2)} \sim N(0, \sigma^2_{\eta(2)})$, $u_{rj} \sim N(0, \sigma^2_{ur})$ and $e_{rij} \sim N(0, \sigma^2_{er})$.

As is usual in factor analysis, constraints on either the factor loadings or the factor variances are required in order to fix the scale of the factors. In the example that follows, we will constrain the first loading of each factor to one, which constrains each factor to have the same scale as the first response. An alternative is to constrain the factor variances to one. If both the factors and responses are standardized to have unit variance, factor loadings can be interpreted as correlations between a factor and the responses. If responses are standardized and constraints are placed on factor loadings rather than factor variances, we can compute standardized loadings for a level k factor (omitting subscripts) as $\lambda_r^{(k)*} = \lambda_r^{(k)} \sigma_{\eta(k)}$.

The model in (2) may be extended in a number of ways. Goldstein and Browne [2] propose a general factor model with multiple factors at each level, correlations between factors at the same level, and covariate effects on the responses.

An Example of Multilevel Factor Analysis

We will illustrate the application of multilevel factor analysis using a dataset of science scores for 2439 students in 99 Hungarian schools. The data consist of scores on four test booklets: a core booklet with components in earth science, physics and biology, two biology booklets and one in physics. Therefore, there are six possible test scores (one earth science, three biology, and two physics). Each student responds to a maximum of five tests, the three tests in the core booklet, plus a randomly selected pair of tests from the other booklets. Each test is marked out of ten. A detailed description of the data is given in [1]. All analysis presented here is based on standardized test scores.

Table 1 Pairwise correlations (variances on diagonal) at school and student levels

	E. Sc. core	Biol. core	Biol. R3	Biol. R4	Phys. core	Phys. R2
School level						
E. Sc. core	0.16					
Biol. core	0.68	0.22				
Biol. R3	0.51	0.68	0.07			
Biol. R4	0.46	0.67	0.46	0.23		
Phys. core	0.57	0.89	0.76	0.62	0.24	
Phys. R2	0.56	0.77	0.58	0.64	0.77	0.16
Student level						
E. Sc. core	0.84					
Biol. core	0.27	0.78				
Biol. R3	0.12	0.13	0.93			
Biol. R4	0.14	0.27	0.19	0.77		
Phys. core	0.26	0.42	0.10	0.29	0.77	
Phys. R2	0.22	0.34	0.15	0.39	0.43	0.83

Table 2 Estimates from two-level factor model with one factor at each level

	Student level			School level		
	$\lambda_r^{(1)}$ (SE)	$\lambda_r^{(1)*}$	σ_{er}^2 (SE)	$\lambda_r^{(2)}$ (SE)	$\lambda_r^{(2)*}$	σ_{ur}^2 (SE)
E. Sc. core	1[a]	0.24	0.712 (0.023)	1[a]	0.36	0.098 (0.021)
Biol. core	1.546 (0.113)	0.37	0.484 (0.022)	2.093 (0.593)	0.75	0.015 (0.011)
Biol. R3	0.583 (0.103)	0.14	0.892 (0.039)	0.886 (0.304)	0.32	0.033 (0.017)
Biol. R4	1.110 (0.115)	0.27	0.615 (0.030)	1.498 (0.466)	0.53	0.129 (0.029)
Phys. core	1.665 (0.128)	0.40	0.422 (0.022)	2.054 (0.600)	0.73	0.036 (0.013)
Phys. R2	1.558 (0.133)	0.37	0.526 (0.030)	1.508 (0.453)	0.54	0.057 (0.018)

[a] Constrained parameter.

Table 1 shows the correlations between each pair of standardized test scores, estimated from a multivariate multilevel model with random effects at the school and student levels. Also shown are the variances at each level; as is usual with attainment data, the within-school (between-student) variance is substantially larger than the between-school variance. The correlations at the student level are fairly low. Goldstein [1] suggests that this is due to the fact that there are few items in each test so the student-level reliability is low. Correlations at the school level (i.e., between the school means) are moderate to high, suggesting that the correlations at this level might be well explained by a single factor.

The results from a two-level factor model, with a single factor at each level, are presented in Table 2. The factor variances at the student and school levels are estimated as $0.127(\text{SE} = 0.016)$ and $0.057(\text{SE} = 0.024)$ respectively. As at each level the standardized loadings have the same sign, we interpret the factors as student-level and school-level measures of overall attainment in science. We note, however, that the student-level loadings are low, which is a result of the weak correlations between the test scores at this level (Table 1). Biology R3 has a particularly low loading; the poor fit for this test is reflected in a residual variance estimate (0.89), which is close to the estimate obtained from the multivariate model (0.93). Thus, only a small amount of the variance in the scores for this biology test is explained by the student-level factor, that is, the test has a low communality. At the school level the factor loadings are higher, and the school-level residual variances from the factor model are substantially lower than those from the multivariate model. The weak correlations between the test scores and the student-level factor suggest that another factor at this level may improve the model. When a second student-level factor is added to the model, however, the loadings on one factor have very large standard errors (results not shown). Goldstein and Browne [2] consider an alternative specification with two correlated student-level factors, but we do not pursue this here. In their model, the loadings for physics on one factor are constrained to zero, and on the other factor, zero constraints are placed on the loadings for all tests except physics.

Multilevel Structural Equation Models

While it is possible to include covariates in a multilevel factor model (see [2]), we are often interested in the effects of covariates on the underlying factors rather than on each response. We may also wish to allow a factor to depend on other factors. A **structural equation model** (SEM) consists of two components: (a) a measurement model in which the multivariate responses are assumed to depend on factors (and possibly covariates), and (b) a structural model in which factors depend on covariates and possibly other factors.

A simple multilevel SEM is:

$$y_{rij} = \alpha_r + \lambda_r^{(1)}\eta_{ij}^{(1)} + u_{rj} + e_{rij} \tag{3}$$

$$\eta_{ij}^{(1)} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}. \tag{4}$$

In the structural model (4), the individual-level factor is assumed to be a linear function of a covariate x_{ij} and cluster-level random effects u_j, which are assumed to be normally distributed with variance σ_u^2. Individual-level residuals e_{ij} are assumed to be normally distributed with variance σ_e^2. To fix the scale of $\eta_{ij}^{(1)}$, the intercept in the structural model, β_0, is constrained to zero and one of the factor loadings is constrained to one.

The model defined by (3) and (4) is a multilevel version of what is commonly referred to as a multiple indicators multiple causes (MIMIC) model. If we substitute (4) in (3), we obtain a special case of the multilevel factor model with covariates, in which u_j is a cluster-level factor with loadings equal to the individual-level factor loadings. If we believe that the factor structure differs across levels, a cluster-level factor can be added to the measurement model (3) and u_j removed from (4). A further equation could then be added to the structural model to allow the cluster-level factor to depend on cluster-level covariates. Another possible extension is to allow for dependency between factors, either at the same or different levels.

In the MIMIC model x is assumed to have an indirect effect on the y's through the factor. The effect of x on y_r is $\lambda_r^{(1)}\beta_1$. If instead we believe that x has a direct effect on the y's, we can include x as an explanatory variable in the measurement model. A model in which the same covariate affects both the y's and a factor is not identified (see [4] for a demonstration of this for a single-level SEM).

An Example of Multilevel Structural Equation Modeling

We illustrate the use of multilevel SEM by applying the MIMIC model of (3) and (4) to the Hungarian

Table 3 Estimates from a two-level MIMIC model

Measurement model	$\lambda_r^{(1)}$ (SE)	σ_{er}^2 (SE)	σ_{ur}^2 (SE)
E. Sc. core	1[a]	0.717 (0.023)	0.095 (0.019)
Biol. core	1.584 (0.095)	0.490 (0.021)	0.023 (0.011)
Biol. R3	0.628 (0.083)	0.892 (0.039)	0.033 (0.017)
Biol. R4	1.150 (0.100)	0.613 (0.030)	0.128 (0.028)
Phys. core	1.727 (0.109)	0.406 (0.021)	0.033 (0.012)
Phys. R2	1.491 (0.105)	0.537 (0.029)	0.053 (0.018)
Structural model	**Estimate (SE)**		
β_1	−0.133 (0.018)		
σ_e^2	0.117 (0.013)		
σ_u^2	0.077 (0.015)		

[a] Constrained parameter.

science data. We consider one covariate, gender, which is coded 1 for girls and 0 for boys. The results are shown in Table 3. The parameter estimates for the measurement model are similar to those obtained for the factor model (Table 2). The results from the structural model show that girls have significantly lower overall science attainment (i.e., lower values on $\eta_{ij}^{(1)}$) than boys. A standardized coefficient may be calculated as $\beta_1^* = \beta_1/\sigma_e$. In our example $\hat{\beta}_1^* = -0.39$, which is interpreted as the difference in standard deviation units between girls' and boys' attainment, after adjusting for school effects (u_j). We may also compute the proportion of the residual variance in overall attainment that is due to differences between schools, which in this case is $0.077/(0.077 + 0.117) = 0.40$.

Estimation and Software

A multilevel factor model may be estimated in several ways using various software packages. A simple estimation procedure, described in [1], involves fitting a multivariate multilevel model to the responses to obtain estimates of the within-cluster and between-cluster covariance matrices, possibly adjusting for covariate effects. These matrices are then analyzed using any SEM software (*see* **Structural Equation Modeling: Software**). Muthén [6] describes another two-stage method, implemented in MPlus (`www.statmodel.com`), which involves analyzing the within-cluster and between-cluster covariances simultaneously using procedures for multigroup analysis. Alternatively, estimation may be carried out in a single step using Markov chain Monte Carlo (MCMC) methods (see [2]), in MLwiN (`www.mlwin.com`) or WinBUGS (`www.mrc-bsu.cam.ac.uk/bugs/`).

Multilevel structural equation models may also be estimated using Muthén's two-stage method. Simultaneous analysis of the within-covariance and between-covariance matrices allows cross-level constraints to be introduced. For example, as described above, the MIMIC model is a special case of a general factor model with factor loadings constrained to be equal across levels. For very general models, however, a one-stage approach may be required; examples include random coefficient models and models for mixed response types where multivariate normality cannot be assumed (see [7] for further discussion of the limitations of two-stage procedures). For these and other general SEMs, Rabe–Hesketh et al. [7] propose **maximum likelihood estimation**, which has been implemented in gllamm (`www.gllamm.org`), a set of Stata programs. An alternative is to use Monte Carlo Markov Chain methods, which are available in WinBUGS (*see* **Markov Chain Monte Carlo and Bayesian Statistics**).

A common problem in the analysis of multivariate responses is **missing data**. For example, in the Hungarian study, responses are missing by design as each student was tested on the core booklet plus two randomly selected booklets from the remaining three. The estimation procedures and software for multilevel SEMs described above can handle missing data, under a 'missing at random' assumption.

In the analysis of the Hungarian science data, we used MLwiN to estimate the multilevel factor model and WinBUGS to estimate the multilevel MIMIC model. The parameter estimates and standard errors presented in Tables 2 and 3 are the means and standard errors from 20 000 samples, with a burn-in of 2000 samples.

Further Topics

We have illustrated the application and interpretation of simple multilevel factor models and SEMs in analyses of the Hungarian science data. In [2], extensions to more than one factor at the same level and correlated factors are considered. Other generalizations include allowing for additional levels, which may be hierarchical or cross-classified with another level, and random coefficients. For instance, in the science attainment example, schools may be nested in areas and the effect of gender on attainment may vary across schools and/or areas.

We have restricted our focus to models for multilevel continuous responses. In many applications, however, responses will be categorical or a mixture of different types. For a discussion of multilevel SEMs for binary, polytomous, or mixed responses, see [7] and **Structural Equation Modeling: Categorical Variables**.

We have not considered multilevel structures that arise from longitudinal studies, where the level one units are repeated measures nested within individuals at level two. (*See* **Multilevel and SEM Approaches to Growth Curve Modeling** for a discussion of multilevel SEMs for longitudinal data.)

References

[1] Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd Edition, Arnold, London.

[2] Goldstein, H. & Browne, W.J. (2002). Multilevel factor analysis modelling using Markov Chain Monte Carlo (MCMC) estimation, in *Latent Variable and Latent Structure Models*, G. Marcoulides & I. Moustaki, eds, Lawrence Erlbaum, pp. 225–243.

[3] McDonald, R.P. & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data, *British Journal of Mathematical and Statistical Psychology* **42**, 215–232.

[4] Moustaki, I. (2003). A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables, *British Journal of Mathematical and Statistical Psychology* **69**, 337–357.

[5] Muthén, B.O. (1989). Latent variable modelling in heterogeneous populations, *Psychometrika* **54**, 557–585.

[6] Muthén, B.O. (1994). Multilevel covariance structure analysis, *Sociological Methods and Research* **22**, 376–398.

[7] Rabe–Hesketh, S., Skrondal, A. & Pickles, A. (2004). Generalized multilevel structural equation modelling, *Psychometrika* **69**, 183–206.

FIONA STEELE

Structural Equation Modeling: Nonstandard Cases

George A. Marcoulides

Volume 4, pp. 1931–1934

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Structural Equation Modeling: Nonstandard Cases

A statistical model is only valid when certain assumptions are met. The assumptions of **structural equation models (SEM)** can be roughly divided into two types: structural and distributional [22]. Structural assumptions demand that no intended (observed or theoretical) variables are omitted from the model under consideration, and that no misspecifications are made in the equations underlying the proposed model. Distributional assumptions include linearity of relationships, completeness of data, multivariate normality (*see* **Multivariate Normality Tests**), and adequate sample size. With real data obtained under typical data gathering situations, violations of these distributional assumptions are often inevitable. So, two general questions can be asked about these distributional assumptions: (a) What are the consequences of violating them? (b) What strategies should be used to cope with them? In this article, each of these questions is addressed.

Linearity of Relationships

Not all relationships examined in the social and behavioral sciences are linear. Fortunately, various procedures are available to test such nonlinearities. Kenny and Judd [11] formulated the first nonlinear SEM model. Their approach used the product of observed variables to define **interaction effects** in **latent variables**. The equation $y = \alpha + \gamma_1\xi_1 + \gamma_2\xi_2 + \gamma_3\xi_1\xi_2 + \zeta$ was used to describe both the direct effects of ξ_1 and ξ_2 on y and the interactive effect $\xi_1\xi_2$. To model the interaction of ξ_1 and ξ_2, multiplicative values of the interactive effect were created. Jöreskog and Yang [10] expanded the approach and illustrated how, even with more elaborate models, a one-product term variable is sufficient to identify all the parameters in the model. But even this approach is difficult to apply in practice due to the complicated nonlinear constraints that must be specified, and the need for large samples. If the interacting variable is discrete (e.g., gender), or can be made so by forming some data groupings, a multisample approach can be easily applied (*see* **Factor Analysis: Multiple Groups**). Based on the multisample approach, the interaction effects become apparent as differences in the parameter estimates when the same model is applied to the grouped sets of data created. Jöreskog [8] recently introduced the latent variable score approach to test interaction effects. The factor score approach does not require the creation of product variables or the sorting of the data based on a categorization of the potential interacting variables. The approach can also be easily implemented using the PRELIS2 and SIMPLIS programs [9, 25] (*see* **Structural Equation Modeling: Software**). Various chapters in Schumacker and Marcoulides [26] discuss both the technical issues and the different methods of estimation available for dealing with nonlinear relationships.

Multivariate Normality

Four estimation methods are available in most SEM programs: Unweighted Least Squares (ULS), Generalized Least Squares (GLS), Asymptotically Distribution Free (ADF) – also called *Weighted Least Squares* (WLS) (*see* **Least Squares Estimation**), and **Maximum Likelihood (ML)**. ML and GLS are used when data are normally distributed. The simplest way to examine normality is to consider univariate **skewness** and **kurtosis**. A measure of multivariate skewness and kurtosis is called *Mardia's coefficient* and its normalized estimate [4]. For normally distributed data, Mardia's coefficient will be close to zero, and its normalized estimate nonsignificant. Another method for judging bivariate normality is based on a plot of the χ^2 percentiles and the mean distance measure of individual observations. If the distribution is normal, the plot of the χ^2 percentiles and the mean distance measure should resemble a straight line [13, 19].

Research has shown that the ML and GLS methods can be used even with minor deviations from normality [20] – the parameter estimates generally remain valid, although the standard errors may not. With more serious deviations the ADF method can be used as long as the sample size is large. With smaller sample sizes, the Satorra-Bentler robust method of parameter estimation (a special type of ADF method) should be used.

Another alternative to handling nonnormal data is to make the data more 'normal-looking' by applying a transformation on the raw data. Numerous transformation methods have been proposed in the literature. The most popular are square root transformations, power transformations, reciprocal transformations, and logarithmic transformations.

The presence of categorical variables may also cause nonnormality. Muthén [14] developed a categorical and continuous variable methodology (implemented in M*plus* (*see* **Structural Equation Modeling: Software**) [16]), which basically permits the analysis of any combination of dichotomous, ordered polytomous, and measured variables. With data stemming from designs with only a few possible response categories (e.g., 'Very Satisfied', 'Somewhat Satisfied', and 'Not Satisfied'), the ADF method can also be used with polychoric (for assessing the degree of association between ordinal variables) or polyserial (for assessing the degree of association between an ordinal variable and a continuous variable) correlations. Ignoring the categorical attributes of data obtained from such items can lead to biased results. Fortunately, research has shown that when there are five or more response categories (and the distribution of data is normal) the problems from disregarding the categorical nature of responses are likely to be minimized.

Missing Data

Data sets with missing values are commonly encountered. **Missing data** are often dealt with by ad hoc procedures (e.g., listwise deletion, pairwise deletion, mean substitution) that have no theoretical justification, and can lead to biased estimates. But there are a number of alternative theory-based procedures that offer a wide range of good data analytic options. In particular, three ML estimation algorithms are available: (a) the multigroup approach [1, 15], which can be implemented in most existing SEM software; (b) full information maximum likelihood (FIML) estimation, which is available in LISREL [9], EQS6 [5], AMOS [2], M*plus* [16], and M*x* [18]; and (c) the Expectation Maximization (EM) algorithm, which is available in SPSS Missing Values Analysis, EMCOV [6], and NORM [24]. All the available procedures assume that the missing values are either missing completely at random (MCAR) or missing at random (MAR) (*see* **Dropouts in Longitudinal Data**; **Dropouts in Longitudinal Studies: Methods of Analysis**). Under MCAR, the probability of a missing response is independent of both the observed and unobserved data. Under MAR, the probability of a missing response depends only on the observed responses. The MCAR and MAR conditions are also often described as ignorable nonresponses. But sometimes respondents and nonrespondents with the same observed data values differ systematically with respect to the missing values for nonrespondents. Such cases are often called *nonignorable nonresponses* or data missing not at random (MNAR), and, although there is usually no correction method available, sometimes, a case-specific general modeling approach might work.

To date, the FIML method has been shown to yield unbiased estimates under both MCAR and MAR scenarios, including data with mild deviations from multivariate normality. This suggests that FIML can be used in a variety of empirical settings with missing data. **Multiple imputation** is also becoming a popular method to deal with missing data. Multiple data sets are created with plausible values replacing the missing data, and a complete data set is subsequently analyzed [7]. For dealing with nonnormal missing data, the likelihood-based approaches developed by Arminger and Sobel [3], and Yuan and Bentler [27] may work better.

Outliers

Real data sets almost always include **outliers**. Sometimes, outliers are harmless and do not change the results, regardless of whether they are included or deleted from the analyses. But, sometimes, they have much influence on the results, particularly on parameter estimates. Assuming outliers can be correctly identified, deleting them is often the preferred approach. Another approach is to fit the model to the data without the outliers and inspect the results to examine the impact of including them in the analysis.

Power and Sample Size Requirements

A common modeling concern involves using the appropriate sample size. Although various sample size rules-of-thumb have been proposed in the literature (e.g., 5–10 observations per parameter, 50 observations per variable, 10 times the number of free model parameters), no one rule can be applied to all situations encountered. This is because the sample size needed for a study depends on many factors including the complexity of the model, distribution of the variables, amount of missing data, reliability of the variables, and strength of the relationships among the variables. Standard errors are also particularly sensitive to sample size issues. For example, if the standard errors in a proposed model are overestimated, significant effects can easily be missed. In contrast, if standard errors are underestimated, significant effects can be overemphasized. As a consequence, proactive Monte Carlo analyses should be used to help determine the sample size needed to achieve accurate Type I error control and parameter estimation precision. The methods introduced by Satorra and Saris [21, 23], and MacCallum, Brown, and Sugawara [12] can be used to assess the sample size in terms of power of the goodness of fit of the model. Recently, Muthén and Muthén [17] illustrated how M*plus* can be used to conduct a Monte Carlo study to help decide on sample size and determine power.

References

[1] Allison, P.D. (1987). Estimation of linear models with incomplete data, in *Sociological Methodology*, C.C. Clogg, ed., Jossey-Bass, San Francisco pp. 71–103.

[2] Arbuckle, J.L. & Wothke, W. (1999). *AMOS 4.0 User's Guide*, Smallwaters, Chicago.

[3] Arminger, G. & Sober, M.E. (1990). Pseudo-maximum likelihood estimation of mean and covariance structures with missing data, *Journal of the American Statistical Association* **85**, 195–2003.

[4] Bentler, P.M. (1995). *EQS Structural Equations Program Manual*, Multivariate Software, Encino.

[5] Bentler, P.M. (2002). *EQS6 Structural Equations Program Manual*, Multivariate Software, Encino.

[6] Graham, J.W. & Hofer, S.M. (1993). *EMCOV.EXE Users Guide*, Unpublished manuscript, University of Southern California.

[7] Graham, J.W. & Hofer, S.M. (2000). Multiple imputation in multivariate research, in *Modeling Longitudinal and Multilevel Data*, T. Little, K.U. Schnabel & J. Baumert, eds., Lawrence Erlbaum Associates, Mahwah.

[8] Jöreskog, K.G. (2000). *Latent Variable Scores and their Uses*, Scientific Software International, Inc, Lincolnwood, (Acrobat PDF file: `http://www.sscentral.com/`).

[9] Jöreskog, K.G., Sörbom, D., Du Toit, S. & Du Toit, M. (1999). *LISREL 8: New Statistical Features*, Scientific Software International, Chicago.

[10] Jöreskog, K.G. & Yang, F. (1996). Nonlinear structural equation models: the Kenny-Judd model with interaction effects, in *Advanced Structural Equation Modeling: Issues and Techniques*, G.A. Marcoulides & R.E. Schumacker, eds., Lawrence Erlbaum Associates, Mahwah, pp. 57–89.

[11] Kenny, D.A. & Judd, C.M. (1984). Estimating the non-linear and interactive effects of latent variables, *Psychological Bulletin* **96**, 201–210.

[12] MacCallum, R.C., Browne, M.W. & Sugawara, H.M. (1996). Power analysis and determination of sample size for covariance structure models, *Psychological Methods* **1**, 130–149.

[13] Marcoulides, G.A. & Hershberger, S.L. (1997). *Multivariate Statistical Methods: A First Course*, Lawrence Erlbaum Associates, Mahwah.

[14] Muthén, B.O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators, *Psychometrika* **49**, 115–132.

[15] Muthén, B.O., Kaplan, D. & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random, *Psychometrika* **52**, 431–462.

[16] Muthén, L.K. & Muthén, B.O. (2002a). *Mplus User's Guide*, Muthén & Muthén, Los Angels.

[17] Muthén, L.K. & Muthén, B.O. (2002b). How to use a Monte Carlo study to decide on sample size and determine power, *Structural Equation Modeling* **9**, 599–620.

[18] Neale, M.C. (1994). *Mx: Statistical Modeling*, Department of Psychiatry, Medical College of Virginia, Richmond.

[19] Raykov, T. & Marcoulides, G.A. (2000). *A First Course in Structural Equation Modeling*, Lawrence Erlbaum Associates, Mahwah.

[20] Raykov, T. & Widaman, K.F. (1995). Issues in applied structural equation modeling research, *Structural Equation Modeling* **2**, 289–318.

[21] Saris, W.E. & Satorra, A. (1993). Power evaluations in structural equation models, in *Testing Structural Equation Models*, K.A. Bollen & J.S. Long, eds, Sage Publications, Newbury Park pp. 181–204.

[22] Satorra, A. (1990). Robustness issues in structural equation modeling: a review of recent developments, *Quality & Quantity* **24**, 367–386.

[23] Satorra, A. & Saris, W.E. (1985). The power of the likelihood ratio test in covariance structure analysis, *Psychometrika* **50**, 83–90.

[24] Schafer, J.L. (1998). *Analysis of Incomplete Multivariate Data*, Chapman & Hall, New York.

[25] Schumacker, R.E. (2002). Latent variable interaction modeling, *Structural Equation Modeling* **9**, 40–54.

[26] Schumacker. R.E. & Marcoulides, G.A., eds (1998). *Interaction and Nonlinear Effects in Structural Equation Modeling*, Lawrence Erlbaum Associates, Mahwah.

[27] Yuan, K.H. & Bentler, P.M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data, *Sociological Methodology* **30**, 165–200.

George A. Marcoulides

Structural Equation Modeling: Nontraditional Alternatives

EDWARD E. RIGDON

Volume 4, pp. 1934–1941

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Structural Equation Modeling: Nontraditional Alternatives

Introduction

Within the broad family of multivariate analytical techniques (*see* **Multivariate Analysis: Overview**), the boundaries that associate similar techniques or separate nominally different techniques are not always clear. **Structural equation modeling** (SEM) can be considered a generalization of a wide class of multivariate modeling techniques. Nevertheless, there are a variety of techniques which, in comparison to conventional SEM, are similar enough on some dimensions but different enough in procedure to be labeled variant SEM methods. (Not all of the creators of these methods would appreciate this characterization.) Each of these methods addresses one or more of the problems which can stymie or complicate conventional SEM analysis, so SEM users ought to be familiar with these alternative techniques. The aim here is to briefly introduce some of these techniques and give SEM users some sense of when these methods might be employed instead of, or in addition to, conventional SEM techniques.

Partial Least Squares

Partial Least Squares (PLS) is now the name for a family of related methods. For the most part, the essential distinction between these methods and conventional SEM is the same as the distinction between principal component analysis and factor analysis. Indeed, PLS can be thought of as 'a constrained form of component modeling' [5], as opposed to constrained modeling of common factors. The strengths and weaknesses of PLS versus SEM largely follow from this distinction.

Origin

PLS was invented by Herman Wold, in the 1960s, under the inspiration of Karl Jöreskog's pioneering work in SEM. Wold, a renowned econometrician (he invented the term, 'recursive model', for example) was Jöreskog's mentor. Wold's intent was to develop the same kind of structural analysis technique but starting from the basis of principal component analysis. In particular, reacting to SEM's requirements for large samples, multivariate normality, and substantial prior theory development, Wold aimed to develop a structural modeling technique that was compatible with small sample size, arbitrary distributions, and what he termed *weak theory*. Like many terms associated with PLS, the meaning of 'weak theory' is not precisely clear. McDonald [12] strongly criticized the methodology for the *ad hoc* nature of its methods.

Goals

Unlike conventional SEM, PLS does not aim to test a model in the sense of evaluating discrepancies between empirical and model-implied covariance matrices. Eschewing assumptions about data distributions or even about sample size, PLS does not produce an overall test statistic like conventional SEM's χ^2. Instead, the stated aim of PLS is to maximize prediction of (that is, to minimize the unexplained portion of) dependent variables, especially dependent observed variables. Indeed, PLS analysis will show smaller residual variances, but larger residual covariances, than conventional SEM analysis of the same data. PLS analysis 'develops by a dialog between the investigator and the computer' [20] – users estimate models and make revisions until they are satisfied with the result.

Estimation of model parameters is purely a byproduct of this optimization effort. Unlike conventional SEM, PLS parameter estimates are not generally consistent, in the sense of converging on population values as sample size approaches infinity. Instead, PLS claims the property of 'consistency at large' [20]. PLS parameter estimates converge on population values as both sample size and the number of observed variables per construct both approach infinity. In application, PLS tends to overestimate structural parameters (known here as 'inner relations') and underestimate measurement parameters (known here as 'outer relations') [20]. The literature argues that this is a reasonable trade, giving up the ability to accurately estimate essentially hypothetical model parameters [7] for an improved ability to predict the data; in other words, it is a focus on the ends rather than the means.

Inputs

At a minimum, Partial Least Squares can be conducted on the basis of a correlation matrix and a structure for relations between the variables. Ideally, the user will have the raw data. Since PLS does not rely on distributional assumptions and asymptotic properties, standard errors for model parameters are estimated via **jackknifing** (randomly omitting data points and reestimating model parameters, and observing the empirical variation in the parameter estimates), while model quality is evaluated, in part, through a blindfolding technique (repeatedly estimating the model parameters with random data points omitted, each time using those parameter estimates to predict the values of the missing data points, and observing the accuracy of these estimates) known as the Stone–Geisser test [7]. Wold [20] specified that the data could be 'scalar, ordinal, or interval'.

Because PLS proceeds in a stepwise fashion through a series of regressions, rather than attempting to estimate all parameters simultaneously, and because it does not rely on distributional assumptions, sample size requirements for PLS are said to be substantially lower than those for conventional SEM [3], but there is little specific guidance. On the one hand, Wold [20] asserted that PLS comes into its own in situations that are 'data-rich but theory-primitive', and the 'consistency at large' property suggests that larger sample size will improve results. Chin [3], borrowing a regression heuristic, suggests finding the larger of either (a) the largest number of arrows from observed variables pointing to any one block variable (see below), or (b) the largest number of arrows from other block variables pointing to any one block variable, and multiplying by 10.

While PLS does not focus on model testing in the same way as conventional SEM, PLS users are still required to specify an initial model or structure for the variables. PLS is explicitly not a model-finding tool like exploratory factor analysis [5] (*see* **Factor Analysis: Exploratory**). Each observed variable is uniquely associated with one block variable. These block variables serve the same role in a PLS model as common factor **latent variables** do in SEM. However, block variables in PLS are not latent variables [12] – they are weighted composites of the associated observed variables, and hence observable themselves. This means that the block variables share in all aspects of the observed variables, including random error, hence the bias in parameter estimation as compared to conventional SEM. But it also means that the PLS user can always assign an unambiguous score for each block variable for each case. In SEM, because the latent variables are not composites of the observed variables, empirical 'factor scores' (expressions of a latent variable in terms of the observed variables) can never be more than approximate, nor is there one clearly best method for deriving these scores.

Besides specifying which observed variables are associated with which block, the user must also specify how each set of observed variables is associated with the block variable. The user can choose to regress the observed variables on the block variable or to regress the block variable on the observed variables. If the first choice is made for all blocks, this is known as 'Mode A'. If the second choice is made for all blocks, this is known as 'Mode B', while making different choices for different blocks is known as 'Mode C' (see Figure 1). Users cannot make different choices for different variables within a block, but it is expected that users will try estimating the model with different specifications in search of a satisfactory outcome.

The best-known implementations of PLS are limited to recursive structural models. Relations between block variables may not include reciprocal relations, feedback loops, or correlated errors between block

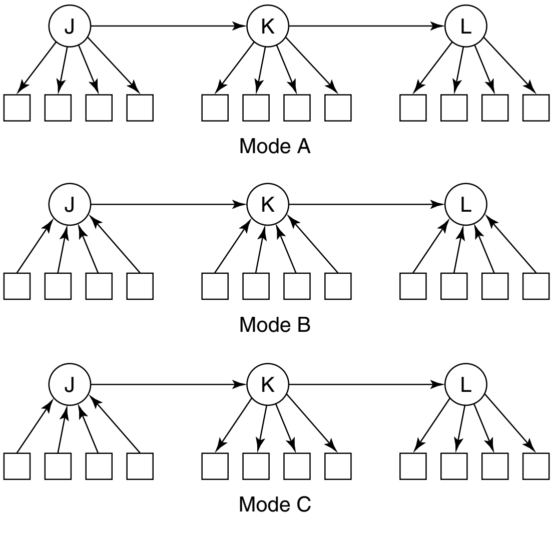


Figure 1 Three modes for estimating relations between observed variables and block variables in partial least squares

variables. Hui [8] developed a procedure and program for PLS modeling of nonrecursive models (*see* **Recursive Models**), but it seems to have been little used.

Execution

PLS parameter estimation proceeds iteratively. Each iteration involves three main steps [20]. The first step establishes or updates the value of each block variable as a weighted sum of the measures in the block. Weights are standardized to yield a block variable with a variance of 1. The second step updates estimates of the 'inner relations' and 'outer relations' parameters. The 'inner relations' path weights are updated through least squares regressions, as indicated by the user's model. For the 'outer relations' part of this step, each block variable is replaced by a weighted sum of all other block variables to which it is directly connected. Wold [20] specified that these sums were weighted simply by the sign of the correlation between the two blocks – hence each block variable is replaced by a 'sign-weighted sum'. For example, the J block variable in Figure 1 would be replaced by the K block variable, weighted by the sign of the correlation between J and K. The K block variable would be replaced by a sum of the J and L block variables, each weighted by the signs of the (J, K) and (K, L) correlations. From here, the procedure for estimating the 'outer relations' weights depends on the choice of mode. For Mode A blocks, each observed variable in a given block is regressed on the sign-weighted composite for its block, in a set of independent bivariate regressions (*see* **Multivariate Multiple Regression**). For Mode B blocks, the sign-weighted composite is regressed on all the observed variables in the block, in one multiple regression. Then the next iteration begins by once again computing the values of the block variables as weighted sums of their observed variables. Estimation iterates until the change in values becomes smaller than some convergence criterion. Chin [3] notes that while different PLS incarnations have used slightly different weighting schemes, the impact of those differences has never been substantial.

Output and Evaluation

PLS produces estimates of path weights, plus R^2 calculations for dependent variables and a block correlation matrix. For Mode A blocks, PLS also produces loadings, which are approximations to the loadings one would derive from a true factor model. Given raw data, PLS will also produce jackknifed standard errors and a value for the Stone–Geisser test of 'predictive relevance'.

Falk and Miller [5] cautiously offer a number of rules of thumb for evaluating the quality of the model resulting from a PLS analysis. At the end, after possible deletions, there should still be three measures per block. Loadings, where present, should be greater than 0.55, so that the communalities (loadings squared) should be greater than 0.30. Every predictor should explain at least 1.5% of the variance of every variable that it predicts. R^2 values should be above 0.10, and the average R^2 values across all dependent block variables should be much higher.

PLS Variants and PLS Software

Again, PLS describes a family of techniques rather than just one. Each step of the PLS framework is simple in concept and execution, so the approach invites refinement and experimentation. For example, Svante Wold, Herman Wold's son, applied a form of PLS to chemometrics (loosely, the application of statistics to chemistry), as a tool for understanding the components of physical substances. As a result, the Proc PLS procedure currently included in the SAS package is designed for this chemometrics form of PLS, rather than for Herman Wold's PLS. There are undoubtedly many proprietary variations of the algorithm and many proprietary software packages being used commercially. The most widely used PLS software must surely be Lohmöller's [11] LVPLS. The program is not especially user-friendly, as it has not benefited from regular updates. Lohmöller's program and a user manual are distributed, free, by the Jefferson Psychometrics Laboratory at the University of Virginia (`http://kiptron.psyc.virginia.edu/disclaimer.html`). Also in circulation is a beta version of Chin's PLSGraph [4], a PLS program with a graphical user interface.

Tetrad

In many ways, Tetrad represents the polar opposite of PLS. There is a single Tetrad methodology, although that methodology incorporates many tools and algorithms. The rationale behind the methodology is fully specified and logically impeccable, and the methodology aims for optimality in well-understood terms. Going beyond PLS, which was designed to enable analysis in cases where theory is weak, the creators of Tetrad regard theory as largely irrelevant: 'In the social sciences, there is a great deal of talk about the importance of "theory" in constructing causal explanations of bodies of data.... In many of these cases the necessity of theory is badly exaggerated.' [17]. Tetrad is a tool for searching out plausible causal inferences from correlational data, within constraints. Tetrad and related tools have found increasing application by organizations facing a surplus of data and a limited supply of analysts. These tools and their application fall into a field known as 'knowledge discovery in databases' or 'data mining' [6]. Academic researchers may turn to such tools when they gain access to secondary commercial data and want to learn about structures underlying that data. Extensive online resources are available from the Tetrad Project at Carnegie Mellon University (`http://www.phil.cmu.edu/projects/tetrad/`) and from Pearl (`http://bayes.cs.ucla.edu/jp_home.html` and `http://bayes.cs.ucla.edu/BOOK-2K/index.html`).

Origin

Tetrad was developed primarily by Peter Spirtes, Clark Glymour, and Richard Scheines [17], a group of philosophers at Carnegie-Mellon University who wanted to understand how human beings develop causal reasoning about events. Prevailing philosophy lays out rigorous conditions which must be met before one can defensibly infer that A causes B. Human beings make causal inferences every day without bothering to check these conditions – sometimes erroneously, but often correctly. From this basis, the philosophers moved on to study the conditions under which it was possible to make sound causal inferences, and the kinds of procedures that would tend to produce the most plausible causal inferences from a given body of data. Their research, in parallel with contributions by Judea Pearl [14] and others, produced algorithms which codified procedures for quickly and automatically uncovering possible causal structures that are consistent with a given data set. The Tetrad program gives researchers access to these algorithms.

Goals

As noted above, the explicit aim of Tetrad is to uncover plausible inferences about the causal structure that defines relationships among a set of variables. Tetrad's creators deal with causality in a straightforward way. By contrast, conventional SEM deals gingerly with causal claims, having hardly recovered from Ling's [10] caustic review of Kenny's [9] early SEM text, *Correlation and Causality*. Today, any mention of the phrase, 'causal modeling' (an early, alternate name for structural equation modeling), is almost sure to be followed by a disclaimer about the near impossibility of making causal inferences from correlational data alone. SEM users, as well as the researchers associated with the Tetrad program, are well aware of the typically nonexperimental nature of their data, and of the various potential threats to the validity of causal inference from such data. Nevertheless, the aim of Tetrad is to determine which causal inferences are consistent with a given data set. As Pearl [14] notes, however, the defining attribute of the resulting inferences is not 'truth' but 'plausibility': the approach 'identifies the mechanisms we can plausibly infer from nonexperimental data; moreover, it guarantees that any alternative mechanism will be less trustworthy than the one inferred because the alternative would require more contrived, hindsighted adjustment of parameters (i.e., functions) to fit the data'.

Inputs

Tetrad proceeds by analysis of an empirical correlation matrix. Sampling is a key issue for researchers in this area. Both [14] and [18] devote considerable attention to the problem of heterogeneity – of a given data set including representative of different populations. Even if all of the represented populations conform to the same structural equation model, but with different population values for some parameters, the combined data set may fit poorly to that same model, with the parameters freely estimated [13]. This phenomenon, discussed in this literature under the title, 'Simpson's Paradox' (*see* **Paradoxes**), explains why researchers must take care to ensure that data are sampled only from homogeneous populations. As noted below, Tetrad does not actually estimate model parameters, so sample size need only be large enough to stably estimate the correlation matrix. Still, larger samples do improve precision.

Unlike conventional SEM, however, users of Tetrad are not expected to have a full theoretical model, or any prior knowledge about the causal structure. Still, Tetrad does exploit prior knowledge. Researchers can impose constraints on the relations between variables – requiring either that a certain relation must exist, or that it must not exist – whether those constraints arise from broad theory or from logical considerations related to such factors as time order. Thus, one might allow 'parent's occupation' to have a causal effect on 'child's occupation', but might disallow a relationship in the opposite direction. Tetrad does not accommodate nonrecursive relationships – the authors view such structures as incompatible with common sense notions of cause and effect.

Far more important than broad 'theory', for the Tetrad user, are logical analysis and the set of assumptions which they are willing to make. One key assumption involves 'causal sufficiency' [17]. A set of observed variables is causally sufficient if the causal structure of those variables can be explained purely by relations among those variables themselves. If a set of observed variables are causally sufficient, then there literally are no latent or unobserved variables involved in a Tetrad analysis. If the researcher does not believe that the set of variables is causally sufficient, they can employ Tetrad algorithms that search for latent variables – for variables outside the data set that explain relations between the observed variables.

Thus, besides providing the correlation matrix, the researcher faces two key choices. First, will they assume causal sufficiency, or not? The assumption greatly simplifies and speeds execution, but a false assumption here invalidates the analysis. In addition, the researcher must select an alpha or Type I error probability. Two key tools for model selection are the tetrad and the partial correlation. A tetrad is a function of the covariances or correlations of four variables:

$$\tau_{ABCD} = \sigma_{AB}\sigma_{CD} - \sigma_{AC}\sigma_{BD} \tag{1}$$

A tetrad that is equal to zero is called a *vanishing tetrad*. Different model structures may or may not imply different sets of vanishing tetrads. For example, if four observed variables are all reflective measures of the same common factor (see Figure 2), then all tetrads involving the four factors vanish. But even if one of the four variables actually measures a different factor, all tetrads still vanish [17].

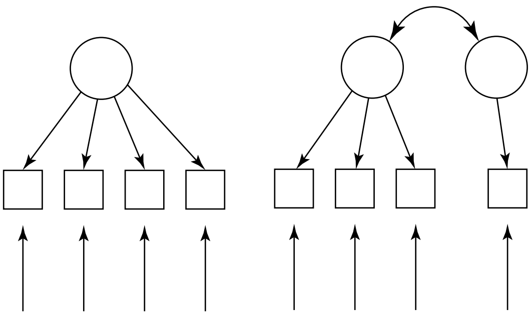


Figure 2 Two measurement models with different substantive implications, yet both implying that all tetrads vanish

The partial correlation of variables B and C, given A, is:

$$\rho_{BC.A} = \frac{\rho_{BC} - \rho_{AB} \times \rho_{AC}}{\sqrt{1-\rho_{AB}^2}\sqrt{1-\rho_{AC}^2}} \tag{2}$$

A zero partial correlation suggests no direct relationship between two variables. For example, for three mutually correlated variables A, B, and C, $\rho_{BC.A} = 0$ suggests either that A is direct or indirect predictor of both B and C or that A mediates the relationship between B and C. By contrast, if A and B predict C, then $\rho_{AB.C} \neq 0$, even if A and B are uncorrelated [17].

When the program is determining whether tetrads and partial correlations are equal to 0, it must take account of random sampling error. A smaller value for alpha, such as .05, will mean wider confidence intervals when Tetrad is determining whether a certain quantity is equal to 0. As a result, the program will identify more potential restrictions which can be imposed, and will produce a more parsimonious and determinate model. A larger value, such as .10 or greater, will lead to fewer 'nonzero' results, producing a less determinate and more saturated model, reflecting greater uncertainty for any given sample size.

Execution

Tetrad's method of analysis is based on a rigorous logic which is detailed at length in [14] and [17], and elsewhere. In essence, these authors argue that if a certain causal structure actually underlies a data set, then the vanishing (and nonvanishing) tetrads and zero (and nonzero) partial correlations and other correlational patterns which are logically implied by that structure will be present in the data, within random sampling error. Only those patterns that are implied by the true underlying structure should be present. Certainly, there may be multiple causal structures that are consistent with a certain data set, but some of these may be ruled out by the constraints imposed by the researcher. (The Tetrad program is designed to identify all causal structures that are compatible with both the data and the prior constraints.) Therefore, if a researcher determines which correlational patterns are present, and which causal structures are consistent with the empirical evidence and with prior constraints, then, under assumptions, the most plausible causal inference is that those compatible causal structures are, in fact, the structures that underlie the data.

Tetrad uses algorithms to comb through all of the available information – correlations, partial correlations, tetrads, and prior constraints – in its search for plausible causal structures. These algorithms have emerged from the authors' extensive research. Different algorithms are employed if the researcher does not embrace the causal sufficiency assumption. Some of these algorithms begin with an unordered set of variables, where the initial assumption is that the set of variables are merely correlated, while others require a prior ordering of variables. The ideal end-state is a model of relations between the variables that is fully directed, where the observed data are explained entirely by the causal effects of some variables upon other variables. In any given case, however, it may not be possible to achieve such a result given the limitations of the data and the prior constraints. The algorithms proceed by changing mere correlations into causal paths, or by deleting direct relationships between pairs of variables, until the system is as fully ordered as possible.

Outputs

The chief output of Tetrad analysis is information about what sorts of constraints on relations between variables can be imposed based on the data. Put another way, Tetrad identifies those permitted causal structures that are most plausible. If causal sufficiency is not assumed, Tetrad also indicates what sorts of latent variables are implied by its analysis, and which observed variables are affected. Tetrad also identifies situations where it cannot resolve the direction of causal influence between two variables.

Unlike conventional SEM or even PLS, Tetrad produces no parameter estimates at all. Computation of tetrads and partial correlations does not require parameter estimates, so Tetrad does not require them. Users who want parameter estimates, standard errors, and fit indices, and so forth might estimate the model(s) recommended by Tetrad analysis using some conventional SEM package. By contrast, researchers who encounter poor fit when testing a model using a conventional SEM package might consider using Tetrad to determine which structures are actually consistent with their data. Simulation studies [18] suggest researchers are more likely to recover a true causal structure by using Tetrad's search algorithms, which may return multiple plausible models, than they will by using the model modification tools in conventional SEM software.

Confirmatory Tetrad Analysis

While Partial Least Squares and Tetrad are probably the two best-known methodologies in this class, others are certainly worth mentioning. Bollen and Ting ([1, 2, 19]) describe a technique called *confirmatory tetrad analysis* (CTA). Unlike Tetrad, CTA aims to test whether a prespecified model is consistent with a data set. The test proceeds by determining, from the structure of the model, which tetrads ought to vanish, and then empirically determining whether or not those tetrads are actually zero. As compared to SEM, CTA has two special virtues. First, as noted above, tetrads are computed directly from the correlation matrix, without estimating model parameters. This is a virtue in situations where the model in question is not statistically identified [1]. Being not identified means that there is no one unique best set of parameter estimates. SEM users have the freedom to specify models that are not statistically identified, but the optimization procedures in conventional SEM packages generally fail when they encounter such a model. Identification is often a problem for models that include causal or formative indicators [2], for example. Such indicators are observed variables that are predictors of, rather than being predicted by, latent variables, as in Mode B estimation in PLS (see Figure 1). In conventional SEM, when a latent variable has only formative indicators, identification problems are quite likely. For such a model, CTA can formally test the model, by way of a χ^2 test statistic [1].

A second virtue of CTA is that two competing structural equation models which are not nested in the conventional sense may be nested in terms of the vanishing tetrads that they each imply [1]. When two models are nested in conventional SEM terms, it means, in essence, that the free parameters of one model are a strict subset of those in the other model. Competing nested models can be evaluated using a χ^2 difference test, but if the models are not nested, then the χ^2 difference test cannot be interpreted. Other model comparison procedures exist within conventional SEM, but they do not readily support hypothesis testing. If the vanishing tetrads implied by one model are a strict subset of the vanishing tetrads implied by the other, then the two models are nested in tetrad terms, and a comparison of such models yields a χ^2 difference test. Models may be nested in tetrad terms even if they are not nested in conventional SEM terms, so conventional SEM users may find value in CTA, in such cases. In addition, avoiding parameter estimation may enable model testing at lower sample sizes than are required by conventional SEM.

The largest hurdle for CTA appears to be the problem of redundant tetrads ([1], [19]). For example, if $\tau_{ABCD} = \rho_{AB}\rho_{CD} - \rho_{AC}\rho_{BD} = 0$, then it necessarily follows that $\tau_{ACBD} = \rho_{AC}\rho_{BD} - \rho_{AB}\rho_{CD}$ is also 0. Thus, CTA involves determining not only which vanishing tetrads are implied by a model, but also which vanishing tetrads are redundant. Testing the model, or comparing the competing models, proceeds from that point.

HyBlock

William Rozeboom's HyBall ([15, 16]) is a comprehensive and flexible package for conducting exploratory factor analysis (EFA). One module in this package, known as HyBlock, performs a kind of structured exploratory factor analysis that may be a useful alternative to the sparse measurement models of conventional SEM. In using HyBlock, a researcher must first allocate the variables in a data set into blocks of related variables. The user must also specify how many common factors are to be extracted from each block and the structural linkages between blocks. Like PLS and Tetrad, HyBlock is limited to recursive structural models. HyBlock conducts the blockwise factor analysis and estimates the between-block structural model, generating both parameter estimates and familiar EFA diagnostics. Thus, one might view HyBlock as an alternative to PLS that is firmly grounded in factor analysis. Alternatively, one might view HyBlock as an alternative to conventional SEM for researchers who believe that their measures are factorially complex.

Conclusion

Conventional SEM is distinguished by large-sample empirical testing of a prespecified, theory-based model involving observed variables which are linked to latent variables through a sparse measurement model. Methodologies which vary from these design criteria will continue to have a place in the multivariate toolkit.

References

[1] Bollen, K.A. & Ting, K.-F. (1993). Confirmatory tetrad analysis, in *Sociological Methodology*, P.V. Marsden, ed., American Sociological Association, Washington, 147–175.

[2] Bollen, K.A. & Ting, K.-F. (2000). A tetrad test for causal indicators, *Psychological Methods* **5**, 3–22.

[3] Chin, W. (1998). The partial least squares approach for structural equation modeling, in G.A. Marcoulides, ed., *Modern Business Research Methods*, *Lawrence Erlbaum Associates*, Mahwah, 295–336.

[4] Chin, W. (2001). *PLS-Graph User's Guide, Version 3.0*, Houston: Soft Modeling, Inc.

[5] Falk, R.F. & Miller, N.B. (1992). *A Primer for Soft Modeling*, University of Akron, Akron.

[6] Frawley, W.J., Piatetsky-Shapiro, G. & Matheus, C.J. (1991). Knowledge discovery in databases: an overview, in *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro & W.J. Frawley, eds, AAAI Press, Menlo Park, pp. 1–27.

[7] Geisser, S. (1975). The predictive sample reuse method with applications, *Journal of the American Statistical Association* **70**, 320–328.

[8] Hui, B.S. (1982). On building partial least squares with interdependent inner relations, in K.G. Joreskog & H. Wold, eds, *Systems Under Indirect Observation, Part II*, North-Holland, 249–272.

[9] Kenny, D.A. (1979). *Correlation and Causality*, Wiley, New York.

[10] Ling, R.F. (1982). Review of 'Correlation and Causation [sic]', *Journal of the American Statistical Association* **77**, 489–491.

[11] Lohmöller, J.-B. (1984). *LVPS Program Manual: Latent Variables Path Analysis with Partial Least Squares Estimation*, Zentralarchiv für empirische Sozialforschung, Universität zu Köln, Köln.

[12] McDonald, R.P. (1996). Path analysis with composite variables, *Multivariate Behavioral Research* **31**, 239–270.

[13] Muthén, B. (1989). Latent variable modeling in heterogeneous populations, *Psychometrika* **54**, 557–585.

[14] Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge.

[15] Rozeboom, W.W. (1991). HYBALL: a method for subspace-constrained oblique factor rotation, *Multivariate Behavioral Research* **26**, 163–177.

[16] Rozeboom, W.W. (1998). *HYBLOCK: A routine for exploratory factoring of block-structured data. Working paper*, University of Alberta, Edmonton.

[17] Spirtes, P., Glymour, C. & Scheines, R. (2001). *Causation, Prediction and Search*, 2nd Edition, MIT Press, Cambridge.

[18] Spirtes, P., Scheines, R. & Glymour, C. (1990). Simulation studies of the reliability of computer-aided model specification using the TETRAD II, EQS, and LISREL programs, *Sociological Methods & Research* **19**, 3–66.

[19] Ting, K.-F. (1995). Confirmatory tetrad analysis in SAS, *Structural Equation Modeling* **2**, 163–171.

[20] Wold, H. (1985). Partial least squares, in *Encyclopedia of Statistical Sciences*, Vol. 6, S. Kotz & N.L. Johnson, eds, Wiley, New York, pp. 581–591.

Edward E. Rigdon

Structural Equation Modeling: Overview

Randall E. Schumacker

Volume 4, pp. 1941–1947

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Structural Equation Modeling: Overview

Structural equation modeling (SEM) has been historically referred to as linear structural relationships, **covariance structure analysis**, or latent variable modeling. SEM has traditionally tested hypothesized theoretical models that incorporate a correlation methodology with correction for unreliability of measurement in the observed variables. SEM models have currently included most statistical applications using either observed variables and/or **latent variables**, for example, **multiple linear regression**, **path analysis**, **factor analysis**, latent growth curves (*see* **Structural Equation Modeling: Latent Growth Curve Analysis**), multilevel, and interaction models (*see* **Generalized Linear Mixed Models**). Six basic steps are involved in structural equation modeling: model specification, model identification, model estimation, model testing, model modification, and model validation.

Model Specification

A researcher uses all relevant theory and related research to develop a theoretical model, that is, specifies a theoretical model. A theoretical model establishes how latent variables are related. The researcher wants the specific theoretical model to be confirmed by the observed sample variance–covariance data. A researcher must decide which observed variables to include in the theoretical model and how these observed variables measure latent variables. *Model specification* implies that the researcher specifies observed and latent variable relationships in a theoretical model and designates which parameters in the model are important. A model is properly specified when the true population model is consistent with the theoretical model being tested, that is, the sample variance–covariance matrix is sufficiently reproduced by the theoretical model. The goal of SEM is to determine whether the theoretical model generated the sample variance–covariance matrix. The sample variance–covariance matrix therefore implies some underlying theoretical model (covariance structure).

The researcher can determine the extent to which the theoretical model reproduces the sample variance–covariance matrix. The theoretical model produces an implied variance–covariance matrix whose elements are subtracted from the original sample variance–covariance matrix to produce a residual variance–covariance matrix. If the residual values are larger than expected, then the theoretical model is *misspecified.* The theoretical model misspecification can be due to errors in either not including an important variable or in including an unimportant variable. A misspecified theoretical model results in bias parameter estimates for variables that may be different from the true population model. This bias is known as *specification error* and indicates that the theoretical model may not fit the sample variance–covariance data and be statistically acceptable.

Model Identification

A researcher must resolve the model identification problem prior to the estimation of parameters for observed and latent variables in the theoretical model. *Model identification* is whether a unique set of parameter estimates can be computed, given the theoretical model and sample variance–covariance data. If many different parameter estimates are possible for the theoretical model, then the model is not identified, that is, indeterminacy exists in the model. The sample variance–covariance data may also fit more than one theoretical model equally well. Model indeterminacy occurs when there are not enough constraints in the theoretical model to obtain unique parameter estimates for the variables. Model identification problems are solved by imposing additional constraints in the theoretical model, for example, specifying latent variable variance.

Observed and latent variable parameters in a theoretical model must be specified as a *free* parameter, a *fixed* parameter, or a *constrained* parameter. A *free* parameter is a parameter that is unknown and a researcher wants to estimate it. A *fixed* parameter is a parameter that is not free, rather fixed to a specific value, for example, 0 or 1. A *constrained* parameter is a parameter that is unknown, but set to equal one or more other parameters. *Model identification* therefore involves setting variable parameters as fixed, free, or constrained (*see* **Identification**).

There have been traditionally three types of model identification distinctions. The three types are distinguished by whether the sample variance–covariance matrix can uniquely estimate the variable parameters in the theoretical model. A theoretical model is *underidentified* when one or more variable parameters cannot be uniquely determined, *just-identified* when all of the variable parameters are uniquely determined, and *overidentified* when there is more than one way of estimating a variable parameter(s). The just- or overidentified model distinctions are considered model identified, while the underidentified distinction yields unstable parameter estimates, and the degrees of freedom for the theoretical model are zero or negative. An underidentified model can become identified when additional constraints are imposed, that is, the model degrees of freedom equal one or greater.

There are several conditions for model identification. A necessary, but not sufficient, condition for model identification, is the *order condition*, under which the number of free variable parameters to be estimated must be less than or equal to the number of distinct values in the sample variance–covariance matrix, that is, only the diagonal variances and one set of off-diagonal covariances are counted. The number of distinct values in the sample variance–covariance matrix is equal to $p(p+1)/2$, where p is the number of observed variables. A saturated model (all variables are related in the model) with p variables has $p(p+3)/2$ free variable parameters. For a sample variance–covariance matrix S with 3 observed variables, there are 6 distinct values $[3(3+1)/2 = 6]$ and 9 free (independent) parameters $[3(3+3)/2]$ that can be estimated. Consequently, the number of free parameters estimated in any theoretical model must be less than or equal to the number of distinct values in the variance–covariance matrix. While the *order condition* is necessary, other sufficient conditions are required, for example, the *rank condition*. The *rank condition* requires an algebraic determination of whether each parameter in the model can be estimated from the sample variance–covariance matrix and is related to the determinant of the matrix.

Several different solutions for avoiding model identification problems are available to the researcher. The first solution involves a decision about which observed variables measure each latent variable. A fixed parameter of one for an observed variable or a latent variable will set the measurement scale for that latent variable, preventing scale *indeterminacy* in the theoretical model for the latent variable. The second solution involves a decision about specifying the theoretical model as *recursive* or *nonrecursive*. A *recursive* model is when all of the variable relationships are unidirectional, that is, no bidirectional paths exist between two latent variables (*see* **Recursive Models**). A *nonrecursive* model is when a bidirectional relationship (reciprocal path) between two latent variables is indicated in the theoretical model. In *nonrecursive* models, the correlation of the latent variable errors should be included to correctly estimate the parameter estimates for the two latent variables. The third solution is to begin with a less complex theoretical model that has fewer parameters to estimate. The less complex model would only include variables that are absolutely necessary and only if this model is identified would you develop a more complex theoretical model. A fourth solution is to save the model-implied variance–covariance matrix and use it as the original sample variance–covariance matrix. If the theoretical model is identified, then the parameter estimates from both analyses should be identical. A final solution is to use different starting values in separate analyses. If the model is identified, then the estimates should be identical. A researcher can check model identification by examining the degrees of freedom for the theoretical model, the rank test, or the inverse of the information matrix.

Model Estimation

A researcher can designate initial parameter estimates in a theoretical model, but more commonly, the SEM software automatically provides initial default estimates or start values. The *initial default estimates* are computed using a noniterative two-stage least squares estimation method. These initial estimates are consistent and rather efficient relative to other iterative methods. After initial start values are selected, SEM software uses one of several different estimation methods available to calculate the final observed and latent variable parameter estimates in the theoretical model, that is, estimates of the population parameters in the theoretical model (*see* **Structural Equation Modeling: Software**).

Our goal is to obtain parameter estimates in the theoretical model that compute an implied variance–covariance matrix Σ, which is as close as possible to the sample variance–covariance matrix S. If all residual variance–covariance matrix elements are zero, then $S - \Sigma = 0$ and $\chi^2 = 0$, that is, a perfect fit of the theoretical model to the sample variance–covariance data. Model estimation therefore involves the selection of a *fitting function* to minimize the difference between Σ and S. Several fitting functions or estimation methods are currently available: unweighted or ordinary least squares (ULS or OLS) (*see* **Least Squares Estimation**), generalized least squares (GLS), **maximum likelihood (ML)**, weighted least squares (WLS) (*see* **Least Squares Estimation**), and asymptotic distribution free (ADF). Another goal in model estimation is to use the correct fit function to obtain parameter estimates that are *unbiased, consistent, sufficient*, and *efficient*, that is, *robust* (*see* **Estimation**).

The ULS or OLS parameter estimates are consistent, but have no distributional assumptions or associated statistical tests, and are scale-dependent, that is, changes in the observed variable measurement scale yield different sets of parameter estimates. The GLS and ML parameter estimates are not scale-dependent so any transformed observed variables will yield parameter estimates that are related. The GLS and ML parameter estimates have desirable asymptotic properties (large sample properties) that yield minimum error variance and unbiased estimates. The GLS and ML estimation methods assume multivariate normality (*see* **Catalogue of Probability Density Functions**) of the observed variables (the sufficient conditions are that the observations are independent, identically distributed, and kurtosis is zero). The WLS and ADF estimation methods generally require a large sample size and do not require observed variables to be normally distributed.

Model estimation with binary and ordinal scaled observed variables introduces a parameter estimation problem in structural equation modeling (*see* **Structural Equation Modeling: Categorical Variables**). If observed variables are ordinal scaled or nonnormally distributed, then GLS and ML estimation methods yield parameter estimates, standard errors, and test statistics that are not robust. SEM software uses different techniques to resolve this problem: a categorical variable matrix (CVM) that does not use Pearson product-moment correlations or an asymptotic variance–covariance matrix based on polychoric correlations of two ordinal variables, polyserial correlations of an ordinal and an interval variable, and Pearson product-moment correlations of two interval variables. All three types of correlations (Pearson, polychoric, and polyserial) are then used to create an asymptotic covariance matrix for analysis in the SEM software. A researcher *should not* use mixed types of correlation matrices or variance–covariance matrices in SEM software, rather create a CVM or asymptotic variance–covariance matrix.

The type of estimation method to use with different theoretical models is still under investigation. The following recommendations, however, define current practice. A theoretical model with interval scaled multivariate normal observed variables should use the ULS or OLS estimation method. A theoretical model with interval scaled multivariate nonnormal observed variables should use GLS, WLS, or ADF estimation methods. A theoretical model with ordinal scaled observed variables should use the CVM approach or an asymptotic variance–covariance matrix with WLS or ADF estimation methods.

Model Testing

Model testing involves determining whether the sample data fits the theoretical model once the final parameter estimates have been computed, that is, to what extent is the theoretical model supported by the sample variance–covariance matrix. Model testing can be determined using an omnibus global test of the entire theoretical model or by examining the individual parameter estimates in the theoretical model.

An omnibus global test of the theoretical model can be determined by interpreting several different model fit criteria. The various model fit criteria are based on a comparison of the theoretical model variance–covariance matrix Σ to the sample variance–covariance matrix S. If Σ and S are similar, then the sample data fit the theoretical model. If Σ and S are quite different, then the sample data do not fit the theoretical model. The model fit criteria are computed on the basis of knowledge of the saturated model (all variable relationships defined), independence model (no variable relationships defined), sample size, degrees of freedom and/or the chi-square value, and range in value from 0 (no fit) to 1 (perfect fit) for several subjective model fit criteria.

The model fit criteria are categorized according to model fit, model parsimony, and model comparison. Model fit is interpreted using the chi-square

(χ^2),goodness-of-fit index (GFI), adjusted goodness-of-fit index (AGFI), or root-mean-square residual (RMR). The model fit criteria are based on a difference between the sample variance–covariance matrix (S) and the theoretical model–reproduced variance–covariance matrix (Σ). Model comparison is interpreted using the Tucker–Lewis index (TLI), Bentler–Bonett Non-Normed fit index (NNFI), Bentler–Bonett Normed fit index (NFI), or the Bentler comparative fit index (CFI) (*see* **Goodness of Fit**). The model comparison criteria compare a theoretical model to an independence model (no variable relationships defined). Model parsimony is interpreted using the normed chi-square (NC), parsimonious fit index (PNFI or PCFI), or **Akaike information criterion (AIC)**. The model parsimony criteria are determined by the number of estimated parameters required to achieve a given value for chi-square, that is, an overidentified model is compared with a restricted model.

Model testing can also involve interpreting the individual parameter estimates in a theoretical model for statistical significance, magnitude, and direction. Statistical significance is determined by testing whether a free parameter is different from zero, that is, parameter estimates are divided by their respective standard errors to yield a test statistic. Another interpretation is whether the sign of the parameter estimate agrees with expectations in the theoretical model. For example, if the expectation is that more education will yield a higher income level, then an estimate with a positive sign would support that expectation. A third interpretation is whether the parameter estimates are within an expected range of values. For example, variances should not have negative values and correlations should not exceed one. The interpretation of parameter estimates therefore considers whether parameter estimates are statistically significant, are in the expected direction, and fall within an expected range of acceptable values; thus, parameter estimates should have a practical and meaningful interpretation.

Model Modification

Model modification involves adding or dropping variable relationships in a theoretical model. This is typically done when sample data do not fit the theoretical model, that is, parameter estimates and/or model test criteria are not reasonable. Basically, the initial theoretical model is modified and the new modified model is subsequently evaluated.

There are a number of procedures available for model modification or what has been termed a *specification search*. An intuitive way to consider modifying a theoretical model is to examine the statistical significance of each parameter estimated in the model. If a parameter is not statistically significant, then drop the variable from the model, which essentially sets the parameter estimate to zero in the modified model. Another intuitive method is to examine the residual matrix, that is, the differences between the sample variance–covariance matrix S and the theoretical model reproduced variance–covariance matrix Σ elements. The residual values in the matrix should be small in magnitude and similar across variables. Large residual values overall indicate that the theoretical model was not correctly specified, while a large residual value for a single variable indicates a problem with that variable only. A large standardized residual value for a single variable indicates that the variable's covariance is not well defined in the theoretical model. The theoretical model would be examined to determine how this particular covariance could be explained, for example, by estimating parameters of other variables.

SEM software currently provides model modification indices for variables whose parameters were not estimated in the theoretical model. The modification index for a particular variable indicates how much the omnibus global chi-square value is expected to decrease in the modified model if a parameter is estimated for that variable. A modification index of 50 for a particular variable suggests that the omnibus global chi-square value for the modified model would be decreased by 50. Large modification indices for variables offer suggestions on how to modify the theoretical model by adding variable relationships in the modified model to yield a better fitting model.

Other model modification indices in SEM software are the *expected parameter change, Lagrange multiplier*, and *Wald statistics*. The expected parameter change (EPC) statistic indicates the estimated change in the magnitude and direction of variable parameters if they were to be estimated in a modified model, in contrast to the expected decrease in the omnibus global chi-square value. The EPC is especially informative when the sign of a variable parameter is not in the expected direction, that is, positive instead of negative. The EPC in this situation would suggest that the parameter for the variable be fixed. The lagrange multiplier (LM) statistic is used to evaluate the effect of freeing a set of fixed parameters in a theoretical model. The Lagrange multiplier statistic can consider a set of variable parameters and is therefore considered the multivariate analogue of the modification index. The Wald (W) statistic is used to evaluate whether variables in a theoretical model should be dropped. The Wald (W) statistic can consider a set of variable parameters and is therefore considered the multivariate analogue of the *individual variable critical values*.

Empirical research suggests that model modification is most successful when the modified model is similar to the underlying population model that reflects the sample variance–covariance data. Theoretically, many different models might fit the sample variance–covariance data. Consequently, new specification search procedures generate all possible models and list the best fitting models based on certain model fit criteria, for example, chi-square, AIC, BIC. For example, a multiple regression equation with 17 independent variables predicting a dependent variable would yield 2^{17}th or 131,072 regression models, not all of which would be theoretically meaningful. SEM software permits the formulation of all possible models; however, the outcome of any specification search should still be guided by theory and practical considerations, for example, the time and cost of acquiring the data.

Model Validation

Model validation involves checking the stability and accuracy of a theoretical model. Different methods are available using SEM software: replication, cross-validation, simulation, **bootstrap**, **jackknife**, and specification search. A researcher should ideally seek model validation using additional random samples of data (replication), that is, multiple sample analysis with the same theoretical model. The other validation methods are used in the absence of replication to provide evidence of model validity, that is, the stability and accuracy of the theoretical model.

Cross-validation involves randomly splitting a large sample data set into two smaller data sets. The theoretical model is analyzed using each data set to compare parameter estimates and model fit statistics. In the simulation approach, a theoretical model is compared to a known population model; hence, a population data set is simulated using a random number generator with a known population model. The bootstrap technique is used to determine the stability of parameter estimates by using a random sample data set as a pseudo-population data set to repeatedly create randomly sampled data sets with replacement. The theoretical model is analyzed using all of the bootstrap data sets and results are compared. The jackknife technique is used to determine the impact of outliers on parameter estimates and fit statistics by creating sample data sets where a different data value is excluded each time. The exclusion of a single data value from each sample data set identifies whether an outlier data value is influencing the results. Specification search examines all possible models and selects the best model on the basis of a set of model fit criteria. This approach permits a comparison of the initial theoretical model to other plausible models that are supported by the sample data.

SEM Software

Several structural equation modeling software packages are available to analyze theoretical models. A theoretical model is typically drawn using squares or rectangles to identify observed variables, small circles or ellipses to identify observed variable measurement error, larger circles or ellipses to identify latent variables, curved arrows to depict variable correlation, and straight arrows to depict prediction of dependent variables by independent variables. A graphical display of the theoretical model therefore indicates direct and indirect effects amongst variables in the model.

Four popular SEM packages are Amos, EQS, Mplus, and LISREL. Amos employs a graphical user interface with drawing icons to create a theoretical model and link the model to a data set, for example, SPSS save file. EQS incorporates a statistics package, model diagrammer, and program syntax to analyze theoretical models. EQS can use either a graphical display or program syntax to analyze models. Mplus integrates random effect, factor, and latent class analysis in both cross-sectional and longitudinal settings for single-level and multilevel designs. LISREL is a matrix command language that specifies the type of matrices to use for a specific theoretical model including which parameters are free and

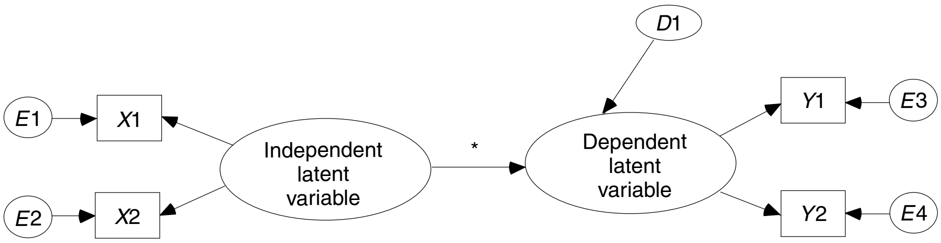


Figure 1 Basic Structural Equation Model

fixed. LISREL includes a data set preprocessor program, PRELIS, to edit and analyze sample data. LISREL also includes a program with simple language commands, SIMPLIS, to input data and specify a theoretical model (*see* **Structural Equation Modeling: Software**).

All four SEM software packages include excellent documentation, tutorials, and data set examples to illustrate how to analyze different types of theoretical models. The SEM software packages are available at their respective Internet web sites and include student versions:

Amos: `http://www.spss.com/amos`
EQS: `http://www.mvsoft.com/`
Mplus `http://www.statmodel.com/`
LISREL: `http://www.ssicentral.com/`

SEM Example

The theoretical model (Figure 1) depicts a single independent latent variable predicting a single dependent latent variable. The independent latent variable is defined by two observed variables, $X1$ and $X2$, with corresponding measurement error designated as $E1$ and $E2$. The dependent latent variable is defined by two observed variables, $Y1$ and $Y2$, with corresponding measurement error designated as $E3$ and $E4$. The parameter estimate or structure coefficient of interest to be estimated is indicated by the asterisk (*) on the straight arrow from the independent latent variable to the dependent latent variable with $D1$ representing the error in prediction. The computer output will also yield an R-squared value that indicates how well the independent latent variable predicts the dependent latent variable.

Further Reading

Geoffrey, M. (1997). *Basics of Structural Equation Modeling*, Sage Publications, Thousand Oaks, ISBN: 0-8039-7408-6; 0-8039-7409-4 (pbk).

Marcoulides, G. & Schumacker, R.E., eds (1996). *Advanced Structural Equation Modeling: Issues and Techniques*, Lawrence Erlbaum Associates, Mahwah.

Marcoulides, G.A. & Schumacker, R.E. (2001). *Advanced Structural Equation Modeling: New Developments and Techniques*, Lawrence Erlbaum Associates, Mahwah.

Schumacker, R.E. & Marcoulides, G.A., eds (1998). *Interaction and Nonlinear Effects in Structural Equation Modeling*, Lawrence Erlbaum Associates, Mahwah.

Schumacker, R.E. & Lomax, R.G. (2004). *A Beginner's Guide to Structural Equation Modeling*, 2nd Edition, Lawrence Erlbaum Associates, Mahwah.

Structural Equation Modeling: A Multidisciplinary Journal. Lawrence Erlbaum Associates, Inc, Mahwah.

Tenko, R. & Marcoulides, G.A. (2000). *A First Course in Structural Equation Modeling*, Lawrence Erlbaum Associates, Mahwah, ISBN: 0-8058-3568-7; 0-8058-3569-5 (pbk).

(*See also* **Linear Statistical Models for Causation: A Critical Review**; **Residuals in Structural Equation, Factor Analysis, and Path Analysis Models**; **Structural Equation Modeling: Checking Substantive Plausibility**; **Structural Equation Modeling: Nontraditional Alternatives**)

RANDALL E. SCHUMACKER

Structural Equation Modeling: Software

Edward E. Rigdon

Volume 4, pp. 1947–1951

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Structural Equation Modeling: Software

Introduction

The first widely distributed special-purpose software for estimating **structural equation models** (SEM) appeared more than twenty years ago. The earliest packages were written with the Fortran programming language and designed to run on mainframe computers, in an age when computing time was a scarce commodity and users punched instructions on stacks of cards. Features and terminology reminiscent of that earlier time still survive in the latest versions of some older SEM packages. While SEM software has become much more sophisticated, professional and user-friendly over the years, SEM software still is not what it could be. There is still substantial room for improvement.

Today, researchers can choose from a variety of packages that differ markedly in terms of their intellectual heritage, interface, statistical sophistication, flexibility, integration with other statistical packages, and price. Researchers may find that a few key criteria will substantially narrow their list of possible choices. Then again, packages regularly add capabilities, often mimicking their competitors, so any purchase or use decision should be based on the latest information.

This overview includes only packages that are primarily designed to estimate conventional structural equation models, with extensions. It excludes packages which are designed for SEM variants such as Tetrad and **Partial Least Squares** (*see* **Structural Equation Modeling: Nontraditional Alternatives**). It also excludes sophisticated modeling packages such as aML (`http://www.applied-ml.com/`) or GLLAMM (`http://www.gllamm.org/`), which are powerful and flexible tools but which lack many of the features, such as a broad array of fit indices and other fit diagnostics from the SEM literature, that users would expect in a SEM package.

Choices

At least a dozen packages are available whose primary or major purpose is the estimation of structural equation models. Typically, each package began as a tool created by leading SEM researchers to facilitate their own analyses, only gradually being adapted to the needs of a larger body of customers. This origin as a tool for a SEM expert may partly explain why these packages seem to offer so little help to the average or novice user. This article starts with a brief overview of most of the known packages, in alphabetical order.

The Amos package, written by James Arbuckle and distributed by SPSS (`http://www.spss.com/amos/`) was unusual when it first appeared. It was perhaps the first SEM package to be designed for a graphical computing environment, like Microsoft Windows®. Taking advantage of the environment's capabilities, Amos allowed users to specify models by drawing them, and offered users a set of drawing tools for the purpose. (Amos also includes a command language called Amos Basic.) Amos was also an early leader in implementing advanced missing data techniques (*see* **Missing Data**). While these capabilities have been copied, to a certain degree, by other leading programs, Amos retains a reputation for being easy to use, and the package has continued to add innovations in this area.

Proc Calis, written by Wolfgang Hartmann, is a procedure within the SAS package (`http://www.sas.com/`). In the early 1980s, Proc Calis was arguably the most sophisticated SEM package available. Its ability to specify constraints on model parameters as nonlinear functions of other parameters was instrumental in allowing researchers to model quadratic effects and multiplicative interactions between latent variables [2] (*see* **Structural Equation Modeling: Nonstandard Cases**). Over the intervening years, however, Proc Calis has not added features and extended capabilities to keep pace with developments.

EQS, written by Peter Bentler, has long been one of the leading SEM packages, thanks to the extensive contributions of its author, both to the program and to the field of SEM (`http://www.mvsoft.com/`). EQS was long distinguished by special features for dealing with nonnormal data, such as a **kurtosis**-adjusted χ^2 statistic [4], and superior procedures for modeling ordinal data (*see* **Ordinal Regression Models**).

LISREL, written by Karl Jöreskog and Dag Sörbom, pioneered the field of commercial SEM software, and it still may be the single most widely used and well-known package (`http://www.

ssicentral.com/). Writers have regularly blurred the distinction between LISREL as software and SEM as statistical method. Along with EQS, LISREL was an early leader in offering procedures for modeling ordinal data. Despite their program's advantage in terms of name recognition, however, the pressure of commercial and academic competition has forced the authors to continue updating their package. LISREL's long history is reflected both in its clear Fortran legacy and its enormous worldwide knowledge base.

Mplus (http://www.statmodel.com/), written by Linda and Bengt Muthén, is one of the newest entrants, but it inherits a legacy from Bengt Muthén's earlier program, LISCOMP. Besides maintaining LISCOMP's focus on nonnormal variables, Mplus has quickly built a reputation as one of the most statistically sophisticated SEM packages. Mplus includes tools for finite mixture modeling and latent class analysis that go well beyond conventional SEM, but which may point to the future of the discipline.

Mx (http://griffin.vcu.edu/mx/), written by Michael Neale, may be as technically sophisticated as any product on the market, and it has one distinguishing advantage: it's free. The software, the manual, and a graphical interface are all available free via the Internet. At its core, perhaps, Mx is really a matrix algebra program, but it includes everything that most users would expect in a SEM program, as well as leading edge capabilities in modeling incomplete data and in finite mixture modeling (*see* **Finite Mixture Distributions**).

SEPATH, written by James Steiger, is part of the Statistica statistical package (http://www.statsoftinc.com/products/advanced.html # structural). SEPATH incorporates many of Steiger's innovations relating to the analysis of correlation matrices. It is one of the few packages that automatically provides correct estimated standard errors for parameter estimates from SEM analysis of a Pearson correlation matrix (*see* **Correlation and Covariance Matrices**). Most packages still produce biased estimated standard errors in this case, unless the user takes some additional steps.

Other SEM packages may be less widely distributed but they are still worth serious consideration. LINCS, written by Ronald Schoenberg, and MECOSA, written by Gerhard Arminger, use the GAUSS (http://www.aptech.com/) matrix algebra programming language and require a GAUSS installation. RAMONA, written by Michael Browne and Gerhard Mels, is distributed as part of the Systat statistical package (http://www.systat.com/products/Systat/productinfo/?sec=1006). SEM, by John Fox, is written in the open-source R statistical programming language (http://socserv.socsci.mcmaster.ca/jfox/Misc/sem/index.html). Like Mx, Fox's SEM is free, under an open-source license.

Finally, STREAMS (http://www.mwstreams.com/) is not a SEM package itself, but it may make SEM packages more user-friendly and easier to use. STREAMS generates language for, and formats output from, a variety of SEM packages. Users who need to take advantage of specific SEM package capabilities might use STREAMS to minimize the training burden.

Criteria

Choosing a SEM software package involves potential tradeoffs among a variety of criteria. The capabilities of different SEM packages change regularly – Amos, EQS, LISREL, Mplus, and Mx have all announced or released major upgrades in the last year or two–so it is important to obtain updated information before making a choice.

Users may be inclined to choose a SEM package that is associated with a more general statistical package which they already license. Thus, licensees of SPSS, SAS, Statistica, Systat, or GAUSS might favor Amos, Proc Calis, SEPath, RAMONA or MECOSA, respectively. Keep in mind that leading SEM packages will generally have the ability to import data in a variety of file formats, so users do not need to use the SEM software associated with their general statistical package in order to be able to share data across applications.

Many SEM packages are associated with specific contributors to SEM, as noted previously. Researchers who are familiar with a particular contributor's approach to SEM may prefer to use the associated software. Beyond general orientation, a given contributor's package may be the only one that includes some of the contributor's innovations. Over time, successful innovations do tend to be duplicated across programs, but users cannot assume that any given package includes tools for dealing with all modeling situations. Currently, for example, Amos does not include any procedures for modeling ordinal data, while Proc Calis does not support multiple group analysis, except when all groups have exactly the same sample size. Some features are largely exclusive, at least for the moment. Mplus and Mx are the only packages with tools for mixture modeling and latent class analysis. This allows these programs to model behavior that is intrinsically categorical, such as voting behavior, and also provides additional options for dealing with heterogeneity in a data set [3]. Similarly, only LISREL and Mplus have procedures for obtaining correct statistical results from modeling ordinal data without requiring exceptionally large sample sizes.

'Ease of use' is a major consideration for most users, but there are many different ways in which a package can be easy to use. Amos pioneered the graphical specification of models, using drawing tools to specify networks of observed and latent variables. Other packages followed suit, to one extent or another. Currently, LISREL allows users to modify a model with drawing tools but not to specify an initial model graphically. On the other hand, specifying a model with drawing tools can become tedious when the model involves many variables, and there is no universal agreement on how to graphically represent certain statistical features of a model. Menus, 'wizards' or other helps may actually better facilitate model specification in such cases. Several packages allow users to specify models through equation-like statements, which can provide a convenient basis for specifying relations among groups of variables.

'Ease of use' can mean ease of interpreting results from an analysis. Researchers working with multiple, competing models can easily become overwhelmed by the volume of output. In this regard, Amos and Mx have special features to facilitate comparisons among a set of competing models.

'Ease of use' can also mean that it is easy to find help when something goes wrong. Unfortunately, no SEM package does an excellent job of helping users in such a situation. To a great extent, users of any package find themselves relying on fellow users for support and advice. Thus, the popularity and history of a particular package can be important considerations. On that score, veteran packages like LISREL offer a broad population of users and a deep knowledge base, while less widely used packages like Proc Calis present special problems.

'Ease of use' could also relate to the ability to try out a product before committing to a major purchase. Many SEM packages offer 'student' or 'demo' versions of the software, usually offering full functionality but only for a limited number of variables. Some packages do not offer a demo version, which makes the purchase process risky and inconvenient. Obviously, free packages like Mx do not require demo versions.

Finally, users may be concerned about price. Developing a full-featured SEM package is no small task. Add to that the costs of supporting and promoting the package, factor in the small user base, relative to more general statistical packages, and it is not surprising that SEM packages tend to be expensive. That said, users should give special consideration to the Mx package, which is available free via the Internet. Mx offers a graphical interface and a high degree of flexibility, although specifying some model forms will involve some programming work. Still, templates are available for conducting many types of analysis with Mx.

Room for Improvement

SEM software has come a long way from the opaque, balky, idiosyncratic, mainframe-oriented packages of the early 1980s, but today's SEM packages still frustrate and inconvenience users and fail to facilitate SEM analysis as well as they could, given only the tools available today. SEM packages have made it easier to specify basic models, but specifying advanced models may require substantial programming, often giving rise to tedious errors, even though there is very little user discretion involved, once the general form of the advanced model is chosen. Users sometimes turn to bootstrap and Monte Carlo methods, as when sample size is too small for stable estimation, and several packages offer this capability. Yet, users find themselves 'jumping through hoops' to incorporate these results into their analyses, even though, once a few choices are made, there really is nothing left but tedium. There is much more to be done in designing SEM packages to maximize the efficiency of the researcher.

Most SEM packages do a poor job of helping users when something goes wrong. For example, when a user's structural equation model is not identified – meaning that the model's parameters cannot be uniquely estimated – SEM packages either simply fail or point the user to one particular parameter that is involved in the problem. Pointing to one parameter, however, may direct the user more to the symptoms and away from the fundamental problem in the model. Bekker, Merckens and Wansbeek [1] demonstrated a procedure, implemented via a Pascal program, that indicates all model parameters that are not identified, but this procedure has not been adopted by any major SEM package.

Several packages have taken steps in the area of visualization, but much more could be done. Confirmatory factor analysis measurement models imply networks of constraints on the patterns of covariances among a set of observed variables. When such a model performs poorly, there are sets of covariances that do not conform to the implied constraints. Currently, SEM packages will point to particular elements of the empirical covariance matrix where the model does a poor job of reproducing the data, and they will also point to particular parameter constraints that contribute to lack of fit. But again, as with identification, this is not the same as showing the user just how the data contradict the network of constraints implied by the model.

Packages could probably improve the advice that they provide about how a researcher might improve a poorly fitting model. Spirtes, Scheines, and Glymour [5] have demonstrated algorithms for quickly finding structures that are consistent with data, subject to constraints. Alongside the diagnostics currently provided, SEM packages could offer more insight to researchers by incorporating algorithms from the stream of research associated with the Tetrad program.

SEM users often find themselves struggling in isolation to interpret results that are somewhat vague, even though there is a large body of researcher experience with SEM generally and with any given program in particular. One day, perhaps, programs will not only generate results but will also help researchers evaluate those results, drawing on this body of experience to give the individual research greater perspective. This type of innovation – giving meaning to numbers – is emerging from the field of artificial intelligence, and it will surely come to structural equation modeling, one day.

References

[1] Bekker, P.A., Merckens, A. & Wansbeek, T.J. (1994). *Identification, Equivalent Models, and Computer Algebra*, Academic Press, Boston.

[2] Kenny, D.A. & Judd, C.M. (1984). Estimating the nonlinear and interactive effects of latent variables, *Psychological Bulletin* **96**, 201–210.

[3] Muthén, B.O. (1989). Latent variable modeling in heterogeneous populations: presidential address to the psychometric society, *Psychometrika* **54**, 557–585.

[4] Satorra, A. & Bentler, P.M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis, in 1988 Proceedings of the Business and Economic Statistics Section of the American Statistical *Association*, Washington, DC, 308–313.

[5] Spirtes, P., Scheines, R. & Glymour, C. (1990). Simulation studies of the reliability of computer-aided model specification using the TETRAD II, EQS, and LISREL Programs, *Sociological Methods & Research* **19**, 3–66.

(*See also* **Software for Statistical Analyses**)

EDWARD E. RIGDON

Structural Equation Modeling and Test Validation

Bruno D. Zumbo

Volume 4, pp. 1951–1958

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Structural Equation Modeling and Test Validation

Ideally, test and measurement validation entails theoretical as well as empirical studies (*see* **Validity Theory and Applications**). Moreover, the term validation implies a process that takes place over time, often in a sequentially articulated fashion. The choice of statistical methods and research methodology for empirical data analyses is of course central to the viability of validation studies. The purpose of this entry is to describe developments in test and measurement validation as well as an important advancement in the statistical methods used in test validation research, structural equation modeling. In particular, a generalized linear **structural equation model** (GLISEM) that is a **latent variable** extension of a **generalized linear model** (GLIM) is introduced and shown to be particularly useful as a statistical methodology for test and measurement validation research.

A Brief Overview of Current Thinking in Test Validation

Measurement or test score validation is an ongoing process wherein one provides evidence to support the appropriateness, meaningfulness, and usefulness of the specific inferences made from scores about individuals from a given sample and in a given context. The concept, method, and process of validation are central to constructing and evaluating measures used in the social, behavioral, health, and human sciences, for without validation, any inferences made from a measure are potentially meaningless.

The above definition highlights two central features in current thinking about validation. First, it is not the measure per se that is being validated but rather the inferences one makes from a measure. This distinction between the validation of a scale and the validation of the inferences from scores obtained from a scale may appear subtle at first blush but, in fact, it has significant implications for measurement and testing because it highlights that the validity of the inferences one makes from test scores is somewhat bounded by place, time, and use of the scores resulting from a measurement operation.

The second central feature in the above definition is the clear statement that inferences made from all empirical measures, irrespective of their apparent objectivity, have a need for validation. That is, it matters not whether one is using an observational checklist, an 'objective' educational, economic, or health indicator such as number of students finishing grade 12, or a more psychological measure such as a self-report depression measure, one must be concerned with the validity of the inferences.

It is instructive to contrast contemporary thinking in **validity theory** with what is commonly seen in many introductory texts in research methodology in the social, behavioral, health, and human sciences.

The Traditional View of Validity

The traditional view of validity focuses on (a) validity as a property of the measurement tool, (b) a measure is either valid or invalid, various types of validity – usually four – with the test user, evaluator, or researcher typically assuming only one of the four types is needed to have demonstrated validity, (c) validity as defined by a set of statistical methodologies, such as correlation with a gold-standard, and (d) reliability is a necessary, but not sufficient, condition for validity.

The traditional view of validity can be summarized in Table 1.

The process of validation then simply portrayed as picking the most suitable strategy from Table 1 and conducting the statistical analyses. The basis for much validation research is often described as a correlation with the 'gold standard'; this correlation is commonly referred to as a validity coefficient.

The Contemporary View of Validity

Several papers are available that describe important current developments in validity theory [4, 5, 9, 12, 13, 20]. The purpose of the contemporary view of validity, as it has evolved over the last two decades, is to expand upon the conceptual framework and power of the traditional view of validity seen in most introductory methodology texts. In brief, the recent history of validity theory is perhaps best captured by the following observations.

Table 1 The traditional categories of validity

Type of validity	What does one do to show this type of validity?
Content	Ask experts if the items (or behaviors) tap the construct of interest.
Criterion-related:	
A. Concurrent	Select a criterion and correlate the measure with the criterion measure obtained in the present
B. Predictive	Select a criterion and correlate the measure with the criterion measure obtained in the future
Construct (A. Convergent and B. Discriminant):	Can be done several different ways. Some common ones are (a) correlate to a 'gold standard', (b) factor analysis, (c) multitrait multimethod approaches

1. Validity is no longer a property of the measurement tool but rather of the inferences made from the scores.
2. Validity statements are not dichotomous (valid/invalid) but rather are described on a continuum.
3. Construct validity is the central most important feature of validity.
4. There are no longer various types of validity but rather different sources of evidence that can be gathered to aid in demonstrating the validity of inferences.
5. Validity is no longer defined by a set of statistical methodologies, such as correlation with a gold-standard but rather by an elaborated theory and supporting methods.
6. As one can see in Zumbo's [20] volume, there is a move to consider the *consequences* of inferences from test scores. That is, along with the elevation of construct validity to an overall validity framework for evaluating test interpretation and use came the consideration of the role of ethical and social consequences as validity evidence contributing to score meaning. This movement has been met with some resistance. In the end, Messick [14] made the point most succinctly when he stated that one should not be simply concerned with the obvious and gross negative consequences of score interpretation, but rather one should consider the more subtle and systemic consequences of 'normal' test use. The matter and role of consequences still remains controversial today and will regain momentum in the current climate of large-scale test results affecting educational financing and staffing, as well as health care outcomes and financing in the United States and Canada.
7. Although it was initially set aside in the move to elevate construct validity, content-based evidence is gaining momentum again in part due to the work of Sireci [19].
8. Of all the threats to valid inferences from test scores, test translation is growing in awareness due to the number of international efforts in testing and measurement (see, for example, [3]).
9. And finally, there is debate as to whether reliability is a necessary but not sufficient condition for validity; it seems that this issue is better cast as one of measurement precision so that one strives to have as little measurement error as possible in their inferences. Specifically, reliability is a question of *data quality,* whereas validity is a question of *inferential quality.* Of course, reliability and validity theory are interconnected research arenas, and quantities derived in the former bound or limit the inferences in the latter.

In a broad sense, then, validity is about evaluating the inferences made from a measure. All of the methods discussed in this encyclopedia (e.g., factor analysis, **reliability**, **item analysis**, **item response modeling**, **regression**, etc.) are directed at building the evidential basis for establishing valid inferences. There is, however, one class of methods that are particularly central to the validation process, **structural equation models**. These models are particularly important to test validation research because they are a marriage of regression, path analysis, and latent variable modeling (often called **factor analysis**). Given that the use of latent variable structural equation models presents one of the most exciting new developments with implications for validity theory, the next section discusses these models in detail.

Generalized Linear Structural Equation Modeling

In the framework of modern statistical theory, test validation research involves the analysis of covariance matrices among the observed empirical data that arise from a validation study using **covariance structure models**. There are two classes of models that are key to validation research: **confirmatory factor analysis** (CFA) (*see* **Factor Analysis: Confirmatory**) and multiple indicators multiple causes (MIMIC) models. The former have a long and rich history in validation research, whereas the latter are more novel and are representative of the merger of the structural equation modeling and item response theory traditions to what will be referred to as generalized linear structural equation models. Many very good examples and excellent texts describing CFA are widely available (e.g., [1, 2, 10]). MIMIC models are a relatively novel methodology with only heavily statistical descriptions available.

An Example to Motivate the Statistical Problem

Test validation with SEM will be described using the Center for Epidemiologic Studies Depression scale (CES-D) as an example. The CES-D is useful as a demonstration because it is commonly used in the life and social sciences. The CES-D is a 20-item scale introduced originally by Lenore S. Radloff to measure depressive symptoms in the general population. The CES-D prompts the respondent to reflect upon his/her last week and respond to questions such as 'My sleep was restless' using an ordered or Likert response format of 'not even one day', '1 to 2 days', '3 to 4 days', '5 to 7 days' during the last week. The items typically are scored from zero (not even one day) to three (5–7 days). Composite scores, therefore, range from 0 to 60, with higher scores indicating higher levels of depressive symptoms. The data presented herein is a subsample of a larger data set collected in northern British Columbia, Canada. As part of a larger survey, responses were obtained from 600 adults in the general population -290 females with an average age of 42 years with a range of 18 to 87 years, and 310 males with an average age of 46 years and a range of 17 to 82 years.

Of course, the composite scale score is not the phenomenon of depression, per se, but rather is related to depression such that a higher composite scale score reflects higher levels of the latent variable depression. Cast in this way, two central questions of test validation are of interest: (a) Given that the items are combined to created one scale score, do they measure just one latent variable? and (b) Are the age and gender of the respondents predictive of the latent variable score on the CES-D? The former question is motivated by psychometric necessities whereas the latter question is motivated by theoretical predictions.

CFA Models in Test Validation

The first validation question described above is addressed by using CFA. In the typical CFA model, the score obtained on each item is considered to be a linear function of a latent variable and a stochastic error term. Assuming p items and one latent variable, the linear relationship may be represented in matrix notation as

$$y = \Lambda\eta + \varepsilon, \tag{1}$$

where y is a $(p \times 1)$ column vector of continuous scores for person i on the p items, Λ is a $(p \times 1)$ column vector of loadings (i.e., regression coefficients) of the p items on the latent variable, η is the latent variable score for person i, and ε is $(p \times 1)$ column vector of measurement residuals. It is then straightforward to show that for items that measure one latent variable, (1) implies the following equation:

$$\Sigma = \Lambda\Lambda' + \Psi, \tag{2}$$

where Σ is the $(p \times p)$ population covariance matrix among the items and Ψ is a $(p \times p)$ matrix of covariances among the measurement residuals or unique factors, Λ' is the transpose of Λ, and Λ is as defined above. In words, (2) tells us that the goal of CFA, like all factor analyses, is to account for the covariation among the items by some latent variables. In fact, it is this accounting for the observed covariation that is fundamental definition of a latent variable – that is, a latent variable is defined by local or conditional independence.

More generally, CFA models are members of a larger class of general linear structural models for a p-variate vector of variables in which the empirical data to be modeled consist of the $p \times p$ unstructured estimator, the sample covariance matrix, S, of the population covariance matrix, $\sum$. A confirmatory factor model is specified by a vector of q unknown parameters, θ, which in turn may generate a covariance matrix, $\sum(\theta)$, for the model. Accordingly, there are various estimation methods such as generalized **least-squares** or **maximum likelihood** with their own criterion to yield an estimator $\hat{\theta}$ for the parameters, and a legion of test statistics that indicate the similarity between the estimated model and the population covariance matrix from which a sample has been drawn (i.e., $\sum = \sum(\theta)$). That is, formally, one is trying to ascertain whether the covariance matrix implied by the measurement model is the same as the observed covariance matrix,

$$S \cong \hat{\Lambda}\hat{\Lambda}' + \hat{\Psi} = \Sigma(\hat{\theta}) = \hat{\Sigma}, \tag{3}$$

where the symbols above the Greek letters are meant to imply sample estimates of these population quantities.

As in regression, the goal of CFA is to minimize the error (in this case, the off-diagonal elements of the residual covariance matrix) and maximize the fit between the model and the data. Most current indices of model fit assess how well the model reproduces the observed covariance matrix.

In the example with the CES-D, a CFA model with one latent variable was specified and tested using a recent version of the software LISREL (*see* **Structural Equation Modeling: Software**). Because the CES-D items are ordinal (and hence not continuous) in nature (in our case a four-point response scale) a polychoric covariance matrix was used as input for the analyses. Using a polychoric matrix is an underlying variable approach to modeling ordinal data (as opposed to an item response theory approach). For a polychoric correlation matrix (*see* **Polychoric Correlation**), an underlying continuum for the polytomous scores is assumed and the observed responses are considered manifestations of respondents exceeding a certain number of latent thresholds on that underlying continuum. Conceptually, the idea is to estimate the latent thresholds and model the observed cross-classification of response categories via the underlying latent continuous variables. Formally, for item j with response categories $c = 0, 1, 2, \ldots, C - 1$, define the latent variable y* such that

$$y_j = c \text{ if } \tau_c < y_j^* < \tau_{c+1}, \tag{4}$$

where τ_c, τ_{c+1} are the latent thresholds on the underlying latent continuum, which are typically spaced at nonequal intervals and satisfy the constraint $-\infty = \tau_0 < \tau_1 < \cdots < \tau_{C-1} < \tau_C = \infty$. It is worth mentioning at this point that the latent distribution does not necessarily have to be normally distributed, although it commonly is due to its well understood nature and beneficial mathematical properties, and that one should be willing to believe that this model with an underlying latent dimension is actually realistic for the data at hand.

Suffice it to say that an examination of the fit indices for our example data with the CES-D, such as the root mean-squared error of approximation (RMSEA), a measure of model fit, showed that the one latent variable model was considered adequate, RMSEA = 0.069, with a 90% confidence interval for RMSEA of 0.063 to 0.074.

The single population CFA model, as described above, has been generalized to allow one to test the same model simultaneously across several populations. This is a particularly useful statistical strategy if one wants to ascertain whether their measurement instrument is functioning the same away in subpopulations of participants (e.g., if a measure functioning the same for males and females). This multigroup CFA operates with the same statistical engine described above with the exception of taking advantage of the statistical capacity of partitioning a likelihood ratio Chi-square and hence testing a series of nested models for a variety of tests of scale level measurement invariance (see [1], for details).

MIMIC Models in Test Validation

The second validation question described above (i.e., are age and gender predictive of CES-D scale scores?) is often addressed by using ordinary least-squares **regression** by regressing the observed composite score of the CES-D onto age and the dummy coded gender variables. The problem with this approach is that the regression results are biased by the measurement error in the observed composite score. Although widely known among psychometricians and statisticians, this bias is ignored in a lot of day-to-day validation research.

The more optimal statistical analysis than using OLS regression is to use SEM and MIMIC models. MIMIC models were first described by Jöreskog and Goldberger [7]. MIMIC models, in their essence, posit a model stating that a set of possible observed explanatory variables (sometimes called *predictors or covariates*) affects latent variables, which are themselves indicated by other observed variables. In our example of the CES-D, the age and gender variables are predictors of the CES-D latent variable, which itself is indicated by the 20 CES-D items. Our example highlights an important distinction between the original MIMIC models discussed over the last three decades and the most recent developments in MIMIC methodology – in the original MIMIC work the indicators of the latent variable(s) were all continuous variables. In our case, the indicators for the CES-D latent variables (i.e., the CES-D items) are ordinal or Likert variables. This complicates the MIMIC modeling substantially and, until relatively recently, was a major impediment to using MIMIC models in validation research.

The recent MIMIC model for ordinal indicator variables is, in short, an example of the merging of statistical ideas in generalized linear models (e.g., logit and probit models) and structural equation modeling into a generalized linear structural modeling framework [6, 8, 16, 17, 18]. This new framework builds on the correspondence between factor analytic models and **item response theory** (IRT) models (see, e.g., [11]) and is a very general class of models that allow one to estimate group differences, investigate predictors, easily compute IRT with multiple latent variables (i.e., multidimensional IRT), investigate differential item functioning, and easily model complex data structures involving complex item and test formats such as testlets, item bundles, test method effects, or correlated errors all with relatively short scales, such as the CES-D.

A recent paper by Moustaki, Jöreskog, and Mavridis [15] provides much of the technical detail for the generalized linear structural equation modeling framework discussed in this entry; therefore, I will provide only a sketch of the statistical approach to motivate the example with the CES-D. In this light, it should be noted that these models can be fit with either Mplus or PRELIS-LISREL. I chose to use the PRELIS-LISREL software, and hence my description of the generalized linear structural equation model will use Jöreskog's notation.

To write a general model allowing for predictors of the observed (manifest) and latent variables, one extends (1) with a new matrix that contains the predictors x

$$\begin{aligned} y^* &= \Lambda z + Bx + u, \text{ where} \\ z &= Dw + \delta, \end{aligned} \tag{5}$$

and u is an error term representing a specific factor and measurement error and y^* is an unobserved continuous variable underlying the observed ordinal variable denoted y, z is a vector of latent variables, w is a vector of fixed predictors (also called *covariates*), D is a matrix of regression coefficients and δ is a vector of error terms which follows a $N(0, I)$. Recall that in (1) the variable being modeled is directly observed (and assumed to be continuous), but in (5) it is not.

Note that because the PRELIS-LISREL approach does not specify a model for the complete p-dimensional response pattern observed in the data, one needs to estimate the model in (5) with PRELIS-LISREL one follows two steps. In the first step (the PRELIS step), one models the univariate and bivariate marginal distributions to estimate the thresholds and the joint covariance matrix of y^*, x, and w and their asymptotic covariance matrix. In the PRELIS step there is no latent variable imposed on the estimated joint covariance matrix hence making that matrix an unconstrained covariance matrix that is just like a sample covariance matrix, S, in (3) above for continuous variables. It can therefore be used in LISREL for modeling just as if y^* was directly observed using (robust) maximum likelihood or weighted least-squares estimation methods.

Turning to the CES-D example, the validity researcher is interested in the question of whether age and gender are predictive of CES-D scale scores. Figure 1 is the resulting generalized MIMIC model. One can see in Figure 1 that the correlation of age and gender is, as expected from descriptive statistics of age for each gender, negative. Likewise, if one were to examine the t values in the LISREL output, both the age and gender predictors are statistically significant. Given the female respondents are coded 1 in the binary gender variable, as a group the female respondents scored higher on the latent variable of depression. Likewise, the older respondents tended to have a lower level of depression compared to the younger respondents in this sample, as reflected in the negative regression coefficient in Figure 1. When the predictive relationship of age was investigated separately for males and females via

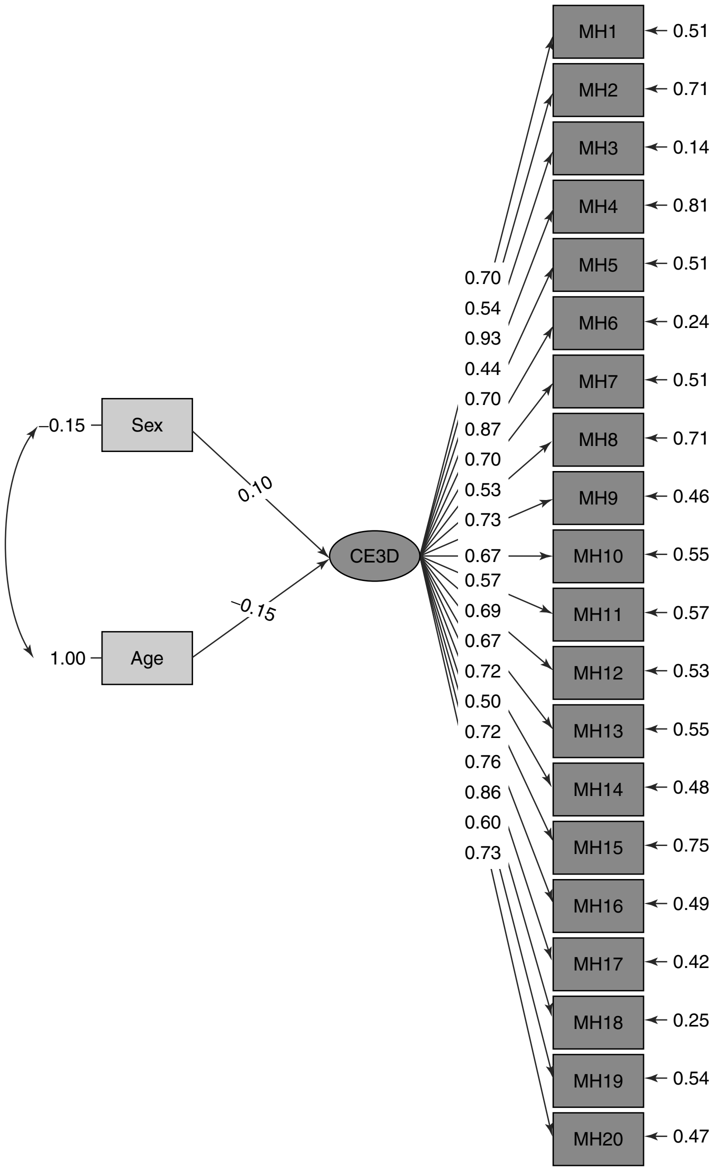


Figure 1 MIMIC model of age and gender for the CES-D (Standardized solution)

this generalized MIMIC model, age was a statistically significant (negative) predictor for the female respondents and age was not a statistically significant for male respondents. Age is unrelated to depression level for men, whereas older women in this sample are less depressed than younger women. This sort of predictive validity information is useful to researchers using the CES-D and hence supports, as described at the beginning of this entry, the inferences made from CES-D test scores.

References

[1] Byrne, B.M. (1998). *Structural Equation Modeling with LISREL, PRELIS and SIMPLIS: Basic Concepts, Applications and Programming*, Lawrence Erlbaum, Hillsdale, N.J.

[2] Byrne, B.M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*, Lawrence Erlbaum, Mahwah.

[3] Hambleton, R.K. & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures, in *Validity Theory and the Methods Used in Validation: Perspectives from the Social and Behavioral Sciences*, B.D. Zumbo, ed., Kluwer Academic Press, Netherlands, pp. 153–171.

[4] Hubley, A.M. & Zumbo, B.D. (1996). A dialectic on validity: Where we have been and where we are going, *The Journal of General Psychology* **123**, 207–215.

[5] Johnson, J.L. & Plake, B.S. (1998). A historical comparison of validity standards and validity practices, *Educational and Psychological Measurement* **58**, 736–753.

[6] Jöreskog, K.G. (2002). Structural equation modeling with ordinal variables using LISREL. Retrieved December 2002 from `http://www.ssicentral.com/lisrel/ordinal.htm`

[7] Jöreskog, K.G. & Goldberger, A.S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable, *Journal of the American Statistical Association* **10**, 631–639.

[8] Jöreskog, K.G. & Moustaki, I. (2001). Factor analysis of ordinal variables: a comparison of three approaches, *Multivariate Behavioral Research* **36**, 341–387.

[9] Kane, M.T. (2001). Current concerns in validity theory, *Journal of Educational Measurement* **38**, 319–342.

[10] Kaplan, D. (2000). *Structural Equation Modeling: Foundations and Extensions*, Sage Publications, Newbury Park.

[11] Lu, I.R.R., Thomas, D.R. & Zumbo, B.D. (in press). Embedding IRT in structural equation models: A comparison with regression based on IRT scores, *Structural Equation Modeling*.

[12] Messick, S. (1989). Validity, in *Educational Measurement*, R.L. Linn, ed., 3rd Edition, American Council on Education/Macmillan, New York, pp. 13–103.

[13] Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning, *American Psychologist* **50**, 741–749.

[14] Messick, S. (1998). Test validity: A matter of consequence, in *Validity Theory and the Methods Used in Validation: Perspectives from the Social and Behavioral Sciences*, B.D. Zumbo, ed., Kluwer Academic Press, pp. 35–44.

[15] Moustaki, I., Jöreskog, K.G. & Mavridis, D. (2004). Factor models for ordinal variables with covariate effects on the manifest and latent variables: a comparison of LISREL and IRT approaches, *Structural Equation Modeling* **11**, 487–513.

[16] Muthen, B.O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables, *Journal of Educational Statistics* **10**, 121–132.

[17] Muthen, B.O. (1988). Some uses of structural equation modeling in validity studies: extending IRT to external variables, in *Test validity*, H. Wainer & H. Braun, eds, Lawrence Erlbaum, Hillsdale, pp. 213–238.

[18] Muthen, B.O. (1989). Latent variable modeling in heterogeneous populations, *Psychometrika* **54**, 551–585.

[19] Sireci, S.G. (1998). The construct of content validity, in *Validity Theory and the Methods used in Validation: Perspectives from the Social and Behavioral Sciences*, B.D. Zumbo, ed., Kluwer Academic Press, pp. 83–117.

[20] Zumbo, B.D., ed. (1998). Validity theory and the methods used in validation: perspectives from the social and behavioral sciences, Special issue of the journal *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement* **45**(1–3), 1–359.

BRUNO D. ZUMBO

Structural Zeros

Vance W. Berger and Jialu Zhang

Volume 4, pp. 1958–1959

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Structural Zeros

Empty or zero cells in a **contingency table** can be classified as either structural zeros or random (sampling) zeros. A sampling zero occurs when the observed cell count in the table is zero, while its expected value is not. This is especially likely when both the sample size and the cell probability are small. For any positive cell probability, however, increasing the sample size sufficiently will ensure that with high probability, the cell count will not be zero; that is, there will not be a random zero. In contrast, increasing the sample size does not have this effect on structural zeros. This is because a cell with a structural zero has an expected value of zero. Clearly, as a nonnegative random variable, this means that its variance is also zero, and that not only did no observations in the data set at hand fall into that cell, but in fact that no observation *could* fall into that cell. The cell count is zero with probability one.

Sampling zeros are part of the data and contribute to the likelihood function (*see* **Maximum Likelihood Estimation**) and model fitting, while structural zeros are not part of the data [1]. Therefore, they do not contribute to the likelihood function or model fitting. A contingency table containing structural zeros is, in some sense, an incomplete table and special analysis methods are needed to deal with structural zeros. Agresti [1] gave an example of a contingency table (Table 1) with a structural zero [1]. The study investigated the effect of primary pneumonia infections in calves on secondary pneumonia infections. The 2 × 2 table has primary infection and secondary infection as the row variable and the column variable, respectively. Since a secondary infection is not possible without a primary infection, the lower left cell has a structural zero. If any of the other three cells had turned out to have a zero count, then they would have been sampling zeros.

Table 1 An Example of a Structural Zero in a 2 × 2 Contingency Table

	Secondary infection		
Primary	Yes	No	
Yes	a (π_{11})	b (π_{12})	π_{1+}
No	–	c (π_{22})	π_{2+}
	π_{+1}	π_{+2}	

Other examples in which structural zeros may arise include cross-classifications by number of children in a household and number of smoking children in a household, number of felonies committed on a given day in a given area and number of these felonies for which at least one suspect was arrested and charged with the felony in question, and number of infections experienced and number of serious infections experienced for patients in a study. Yu and Zelterman [4] presented a triangular table with families classified by both number of siblings (1–6) and number of siblings with interstitial pulmonary fibrosis (0–6). The common theme is that the initial classification bears some resemblance to a Poisson variable, or a count variable, and the secondary variable bears some resemblance to a binary classification of each Poisson observation.

The structural zero occurs because of the restriction on the number of binary 'successes' imposed by the total Poisson count; that is, there cannot be a family with two children and three children smokers, or a day with six felonies and eight felonies with suspects arrested, or a patient with no infection but one serious infection. Table 1 of [4] is triangular for this reason too; that is, every cell above the main diagonal, in which the number of affected siblings would exceed the number of siblings, is a structural zero.

For two-way frequency tables, the typical analyses are based on Pearson's χ^2 test or Fisher's exact test (*see* **Exact Methods for Categorical Data**). These are tests of association (*see* **Measures of Association**) of the row and column variables. The null hypothesis is that the row and column variables are independent, while the alternative hypothesis is that the variables are associated. The usual formulation of 'no association' is that the cell probabilities in any given column are common across the rows, or $p(1,1) = p(2,1)$ and $p(1,2) = p(2,2)$, with more such equalities if the table has more than two rows and/or more than two columns. But if one cell is a structural zero and the corresponding cell in the same column is not, then the usual null hypothesis makes no sense. Even under the null hypothesis, the cell probabilities cannot be the same owing to the zero count cells, so the row and column variables are not independent even under the null hypothesis. This is not as big a problem as it may first appear, however, because further thought reveals that interest would not lie in the usual null hypothesis anyway. In fact, there is no reason to even insist on the usual two-way structure at all.

The three possible outcomes in the problem of primary and secondary pneumonia infections of calves can be displayed alternatively as a one-way layout, or a 1×3 table with categories for no infection, only a primary infection, or a combination of a primary and a secondary infection. This new variable is the information-preserving composite endpoint [2]. Agresti considered testing if the probability of the primary infection is the same as the conditional probability of a secondary infection given that the calf got the primary infection [1]; that is, the null hypothesis can be written as $\mathrm{H}_0 : \pi_{1+} = \pi_{11}/\pi_{1+}$. Tang and Tang [3] developed several exact unconditional methods on the basis of the above null hypothesis. The one-way layout can be used for the other examples as well, but two-way structures may be used with different parameterizations.

Instead of number of children in a household and number of smoking children in a household, for example, one could cross-classify by number of nonsmoking children and number of smoking children. This would avoid the structural zero. Likewise, structural zeros could be avoided by cross-classifying by the number of felonies committed on a given day in a given area without an arrest and the number of these felonies for which at least one suspect was arrested and charged with the felony in question. Finally, the number of nonserious infections experienced and the number of serious infections experienced for patients in a study could be tabulated with two-way structure and no structural zeros.

References

[1] Agresti, A. (2002). *Categorical Data Analysis*, 2nd edition, Wiley – Interscience.

[2] Berger, V.W. (2002). Improving the information content of categorical clinical trial endpoints, *Controlled Clinical Trials* **23**, 502–514.

[3] Tang, N.S. & Tang, M.L. (2002). Exact unconditional inference for risk ratio in a correlated 2×2 table with structural zero, *Biometrics* **58**, 972–980.

[4] Yu, C. & Zelterman, D. (2002). Statistical inference for familial disease clusters, *Biometrics* **58**, 481–491.

Vance W. Berger and Jialu Zhang

Subjective Probability and Human Judgement

Peter Ayton

Volume 4, pp. 1960–1967

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Subjective Probability and Human Judgement

How good are people at judging probabilities? One early benchmark used for comparison was Bayes' theorem (*see* **Bayesian Belief Networks**). Bayes' theorem defines mathematically how probabilities should be combined and can be used as a normative theory of the way in which subjective probabilities representing degrees of belief attached to the truth of hypotheses should be revised in the light of new information. Bayes' theorem states that the posterior odds of the hypothesis being correct in the light of new information is a product of two elements: the *prior odds* of the hypothesis being correct before the information is observed and the *likelihood ratio* of the information, given that the hypothesis is correct or incorrect (*see* **Bayesian Statistics**).

In the 1960s, Ward Edwards and his colleagues conducted a number of studies using the book-bag and poker-chip paradigm. A typical experiment would involve two opaque bags. Each bag contained 100 colored poker-chips in different, but stated, proportions of red to blue. One – bag A contains 70 red chips and 30 blue, while the second – bag B contains 30 red chips and 70 blue. The experimenter first chooses one bag at random and then draws a series of chips from it. After each draw, the poker-chip is replaced and the bag well shaken before the next chip is drawn. The subject's task is to say how confident he/she is – in probability terms – that the chosen bag is bag A, containing predominantly red chips, or bag B, containing predominantly blue chips. As the bag was drawn randomly from two bags, our prior odds that we have bag A (or bag B) are 0.5/0.5. If we draw (say) a red chip, we know the likelihood of this is 0.7 if we have bag A and 0.3 if we have bag B. We thus multiply $0.5/0.5 \times 0.07/0.03$ to discover the posterior odds $(0.35/0.15 = 0.7/0.3)$. The posterior odds computed after the first draw then become the prior odds for computing the impact of the second draw and the process repeated subsequently.

A crucial aspect of the logic of these studies is that the experimenter is able to say what the correct subjective probabilities should be for the participants by the simple expedient of calculating them using Bayes' theorem. All of the information required as inputs to Bayes' theorem is explicit and unambiguous. Ironically, though this meant that the *subjectivity* of probability was not a part of the studies, in the sense that the experimenters assumed that they could objectively compute that the correct answer – which they would be able to assume – should be the same for all the participants faced with the same evidence.

The fact that the experimenter assumes he is able to calculate what the subjective probabilities *should* be for all of the participants was absolutely necessary if one was to be able to judge judgment by this method. However, it is also an indication of the artificiality of the task – and is at the root of the difficulties that were to emerge with interpreting the participants' behavior. The experiments conducted with this procedure produced a good deal of evidence that human judgment under these conditions is not well described by Bayes' theorem. Although participants' opinion revisions were proportional to the values calculated from Bayes' rule, they did not revise their opinions sufficiently in the light of the evidence, a phenomenon that was labeled *conservatism*. The clear suggestion was that human judgment was to this extent poor, although there was some debate as to the precise reason for this. It might be due to a failure to understand the impact of the evidence or to an inability to aggregate the assessments according to Bayes' theorem. Aside from any theoretical interest in these possibilities, there were practical implications of this debate. If people are good at assessing probabilities, but poor at combining them (as Edwards [5] suggested), then perhaps they could be helped; a relatively simple remedy would be to design a support system that took the human assessments and combined them using Bayes' theorem. However, if they were poor at assessing the component probabilities, then there would not be much point in devising systems to help them aggregate these.

Before any firm conclusions were reached as to the cause of conservatism, however, the research exploring the phenomenon fizzled out. The reasons for this seem to be twofold. One cause was the emergence of the heuristics and biases research and, in particular, the discovery of what Kahneman and **Tversky** [19] called *base-rate neglect*. Base-rate neglect is the exact opposite of conservatism – according to this account of judgment, people, far from being conservative about opinion revision, disregard prior odds and are *only* influenced by the likelihood ratio. Before this development occurred, however, there was growing disquiet as to the validity of the book-bag experimental method as a basis for judging real-world judgment.

A number of studies had shown considerable variability in the amount of conservatism manifested according to various, quite subtle differences in the task set to participants. For example, the *diagnosticity* of the data seemed an important variable. Imagine, instead of our two bags with a 70/30 split in the proportions of blue and red poker-chips, the bags contained 49 red and 51 blue or 49 blue and 51 red chips. Clearly, two consecutive draws of a blue chip would not be very diagnostic as to which of the two bags we were sampling from. Experiments have shown that the more diagnostic the information, the more conservative is the subject. When information is very weakly diagnostic, as in our example, human probability revision, rather than being conservative, can be too extreme [29].

DuCharme and Peterson [4] argued that the fact that the information was restricted to one of two different possibilities (red chip or blue chip) meant that there were very few possible revisions that could be made. In the real world, information leading to revision of opinion does not have discrete values, but may more fairly be described as varying along a continuum. In an experimental study, DuCharme and Peterson used a hypothesis test consisting of the population of male heights and the population of female heights. The participants' task was to decide which population was being sampled from, based on the information given by randomly sampling heights from one of the populations. Using this task, DuCharme and Peterson found conservatism greatly reduced to half the level found in the more artificial tasks. They concluded that this was due to their participants' greater familiarity with the data-generating process underlying their task.

The argument concerning the validity of the conclusions from the book-bag and poker-chip paradigm was taken further by Winkler and Murphy [35]. Their article entitled 'Experiments in the laboratory and the real world' argued that the standard task differed in several crucial aspects from the real world. For example, the bits of evidence that are presented to experimental participants are conditionally independent; knowing one piece of information does not change the impact of the other. Producing one red chip from the urn and then replacing it does not affect the likelihood of drawing another red chip. However, in real-world probability revision, this assumption often does not make sense.

For example, consider a problem posed by medical diagnosis. Loss of appetite is a symptom which, used in conjunction with other symptoms, can be useful for identifying the cause of certain illnesses. However, if I know that a patient is nauseous, I know that they are more likely (than in the absence of nausea) to experience loss of appetite. These two pieces of information, therefore, are not conditionally independent and so, when making my diagnosis, I should not revise my opinion on seeing the loss of appetite symptom as much as I might, before knowing about the nausea symptom, to diagnose diseases indicated by loss of appetite.

Winkler and Murphy argued that in many real-world situations lack of conditional independence of the information would render much of it redundant. In the book-bag task, participants may have been behaving much as they do in more familiar situations involving redundant information sources. Winkler and Murphy considered a range of other artificialities with this task and concluded that 'conservatism may be an artifact caused by dissimilarities between the laboratory and the real world'.

Heuristics and Biases

From the early 1970s, Kahneman and Tversky provided a formidable series of demonstrations of human judgmental error and linked these to the operation of a set of mental heuristics – mental rules of thumb – that they proposed the brain uses to simplify the process of judgment. For example, Tversky and Kahneman [30] claimed that human judgment is overconfident, ignores base rates, is insufficiently regressive, is influenced by arbitrary anchors, induces illusory correlations, and misconceives randomness. These foibles, they argued, indicated that the underlying judgment process was not normative (i.e., it did not compute probabilities using any kind of mental approximation to Bayes' theorem), but instead used simpler rules that were easier for the brain to implement quickly.

The idea, spelled out in [18], is that, due to limited mental processing capacity, strategies of simplification are required to reduce the complexity of judgment tasks and make them tractable for the kind of mind that people happen to have. Accordingly, the principal reason for interest in judgmental biases was not merely that participants made errors, but that it supported the notion that people made use of relatively simple, but error-prone, heuristics for making judgments.

One such heuristic is *representativeness*. This heuristic determines how likely it is that an event is a member of a category according to how similar or typical the event is to the category. For example, people may judge the likelihood that a given individual is employed as a librarian by the extent to which the individual resembles a typical librarian. This may seem a reasonable strategy, but it neglects consideration of the relative prevalence of librarians. Tversky and Kahneman found that when base rates of different categories vary, judgments of the occupations of described people were correspondingly biased due to base-rate neglect. People using the representativeness heuristic for forecasting were employing a form of stereotyping in which similarity dominates other cues as a basis for judgment and decision-making.

In Kahneman and Tversky's [19] experiments demonstrating neglect of base rates, participants were found to ignore information concerning the prior probabilities of the hypotheses. For example, in one study, participants were presented with the following brief personal description of an individual called Jack:

> *Jack is a 45-year old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles.*

Half the participants were told that the description had been randomly drawn from a sample of 70 engineers and 30 lawyers, while the other half were told that the description was drawn from a sample of 30 engineers and 70 lawyers. Both groups were asked to estimate the probability that Jack was an engineer (or a lawyer). The mean estimates of the two groups of participants were only very slightly different (50 vs 55%). On the basis of this result and others, Kahneman and Tversky concluded that prior probabilities are largely ignored when individuating information was made available.

Although participants used the base rates when told to suppose that they had no information whatsoever about the individual (a 'null description'), when a description designed to be totally uninformative with regard to the profession of an individual called Dick was presented, complete neglect of the base rates resulted.

> *Dick is a 30-year-old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues.*

When confronted with this description, participants in both base rate groups gave median estimates of 50%. Kahneman and Tversky concluded that when no specific evidence is given, the base rates were properly utilized; but when worthless information is given, base rates were neglected.

Judgment by representativeness was also invoked by Tversky and Kahneman [32] to explain the *conjunction fallacy* whereby a conjunction of events is judged more likely than one of its constituents. This is a violation of a perfectly simple principle of probability logic: If A includes B, then the probability of B cannot exceed A. Nevertheless, participants who read a description of a woman called Linda who had a history of interest in liberal causes gave a higher likelihood to the possibility that she was a feminist bank clerk than to the possibility that she was a bank clerk – thereby violating the conjunction rule. Although it may seem unlikely that someone who had interests in liberal causes would be a bank clerk, but a bit more likely that she were a feminist bank clerk, all feminist bank clerks are of course bank clerks.

Another heuristic used for probabilistic judgment is *availability*. This heuristic is invoked when people estimate likelihood or relative frequency by the ease with which instances can be brought to mind. Instances of frequent events are typically easier to recall than instances of less frequent events, so availability will often be a valid cue for estimates of likelihood. However, availability is affected by factors other than likelihood. For example, recent events and emotionally salient events are easier to recollect. It is a common experience that the perceived riskiness of air travel rises in the immediate wake of an air disaster. Applications for earthquake insurance in California are apparently higher in the immediate wake of a major quake. Judgments made on the basis of availability then are vulnerable to bias whenever availability and likelihood are uncorrelated.

The *anchor and adjust* heuristic is used when people make estimates by starting from an initial value that is adjusted to yield a final value. The claim is that adjustment is typically insufficient. For instance, one experimental task required participants to estimate various quantities stated in percentages (e.g., the percentage of African countries in the UN). Participants communicated their answers by using a spinner wheel showing numbers between 0 and 100. For each question, the wheel was spun and then participants were first asked whether the true answer was above or below this arbitrary value. They then gave their estimate of the actual value. Estimates were found to be considerably influenced by the initial (entirely random) starting point (cf. [34]).

The research into heuristics and biases provided a methodology, a very vivid explanatory framework and a strong suggestion that judgment is not as good as it might be. However, the idea that all of this should be taken for granted was denied by the proponents of the research some time ago. For example, Kahneman and Tversky [20] made clear that the main goal of the research was to understand the processes that produce both valid and invalid judgments. However, it soon became apparent that: 'although errors of judgment are but a method by which some cognitive processes are studied, the method has become a significant part of the message' [20, p. 494]. So, how should we regard human judgment?

There has been an enormous amount of discussion of Tversky and Kahneman's findings and claims. Researchers in the heuristics and biases tradition have sometimes generated shock and astonishment that people seem so bad at reasoning with probability despite the fact that we all live in an uncertain world. Not surprisingly, and as a consequence, the claims have been challenged. The basis of the challenges has varied. Some have questioned whether these demonstrations of biases in judgment apply merely to student samples or also to experts operating in their domain of expertise. Another argument is that the nature of the tasks set to participants gives a misleading perspective of their competence. A third argument is that the standards for the assessment of judgment are inappropriate.

Criticisms of Heuristics and Biases Research

Research following Tversky and Kahneman's original demonstration of base-rate neglect established that base rates might be attended to more (though usually not sufficiently) if they were perceived as relevant [1], had a causal role [31], or were 'vivid' rather than 'pallid' in their impact on the decision-maker [27]. However, Gigerenzer, Hell, and Blank [10] have argued that the real reason for variations in base-rate neglect has nothing to do with any of these factors *per se*, but because the different tasks may, to varying degrees, encourage the subject to represent the problem as a Bayesian revision problem. They claimed that there are few inferences in real life that correspond directly to Bayesian revision where a known base-rate is revised on the basis of new information. Just because the experimenter assumes that he has defined a Bayesian revision problem does not imply that the subject will see it the same way. In particular, the participants may not take the base rate asserted by the experimenter as their subjective prior probability. In Kahneman and Tversky's original experiments, the descriptions were not actually randomly sampled (as the participants were told), but especially selected to be 'representative' of the professions. To the extent that the participants suspected that this was the case then they would be entitled to ignore the offered base rate and replace it with one of their own perception.

In an experiment, Gigerenzer et al. [10] found that when they let the participants experience the sampling themselves, base-rate neglect 'disappeared'. In the experiment, their participants could examine 10 pieces of paper, each marked lawyer or engineer in accord to the base rates. Participants then drew one of the pieces of paper from an urn and it was unfolded so they could read a description of an individual without being able to see the mark defining it as being of a lawyer or engineer. In these circumstances, participants clearly used the base rates in a proper fashion. However, in a replication of the verbal presentation where base rates were asserted, rather than sampled, Kahneman and Tversky's base-rate neglect was replicated.

In response to this, Kahneman and Tversky [21] argued that a fair summary of the research would be that explicitly presented base rates are generally underweighted, but not ignored. They have also pointed out that in Gigerenzer et al.'s experiment [10], participants who sampled the information themselves still produced judgments that deviated from the Bayesian solution in the direction predicted by representativeness. Plainly, then representativeness is useful for predicting judgments. However, to the extent that base rates are not entirely ignored (as argued in an extensive review of the literature by Koehler [23]), the heuristic rationale for representativeness is limited. Recall that the original explanation for base-rate neglect was the operation of a simple heuristic that reduced the need for integration of multiple bits of information. If judgments in these experiments reflect base rates – even to a limited extent – it is hard to account for by the operation of the representativeness heuristic.

Tversky and Kahneman [32] reported evidence that violations of the conjunction rule largely disappeared when participants were requested to assess the relative frequency of events rather than the probability of a single event. Thus, instead of being asked about likelihood for a particular individual, participants were requested to assess how many people in a survey of 100 adult males had had heart attacks and then were asked to assess the number of those who were both over 55 years old *and* had had heart attacks. Only 25% of participants violated the conjunction rule by giving higher values to the latter than to the former. When asked about likelihoods for single events, it is typically the vast majority of participants who violate the rule. This difference in performance between frequency and single-event versions of the conjunction problem has been replicated several times since (cf. [8]).

Gigerenzer (e.g., [8], [9] has suggested that people are naturally adapted to reasoning with information in the form of frequencies and that the conjunction fallacy 'disappears' if reasoning is in the form of frequencies for this reason. This suggests that the difficulties that people experience in solving probability problems can be reduced if the problems require participants to assess relative frequency for a class of events rather than the probability of a single event. Thus, it follows that if judgments were elicited with frequency formats there would be no biases. Kahneman and Tversky [21] disagree and argue that the frequency format serves to provide participants with a powerful cue to the relation of inclusion between sets that are explicitly compared, or evaluated in immediate succession. When the structure of the conjunction is made more apparent, then participants who appreciate the constraint supplied by the rule will be less likely to violate it. According to their account, salient cues to set inclusion, not the frequency information *per se*, prompted participants to adjust their judgment.

To test this explanation, Kahneman and Tversky [21] reported a new variation of the conjunction problem experiment where participants made judgments of frequencies, but the cues to set inclusion were removed. They presented participants with the description of Linda and then asked their participants to suppose that there were 1000 women who fit the description. They then asked one group of participants to estimate how many of them would be bank tellers; a second independent group of participants were asked how many were bank tellers and active feminists; a third group made evaluations for both categories. As predicted, those participants who evaluated both categories mostly conformed to the conjunction rule. However, in a between-groups comparison of the other two groups, the estimates for 'bank tellers and active feminists' were found to be significantly higher than the estimates for bank tellers. Kahneman and Tversky argue that these results show that participants use the representativeness heuristic to generate their judgments and then edit their responses to respect class inclusion where they detect cues to that relation. Thus, they concluded that the key variable controling adherence to the conjunction rule is not the relative frequency format *per se*, but the opportunity to detect the relation of class inclusion.

Other authors have investigated the impact of frequency information [7, 12, 13, 24] and concluded that it is not the frequency information *per se*, but the perceived relations between the entities that is affected by different versions of the problem, though this is rejected by Hoffrage, Gigerenzer, Krauss, and Martignon [15].

We need to understand more of the reasons underlying the limiting conditions of cognitive biases – how it is that seemingly inconsequential changes in the format of information can so radically alter the quality of judgment. Biases that can be cured so simply cannot be held to reveal fundamental characteristics of the processes of judgment. Gigerenzer's group has recently developed an alternative program of research studying the *efficacy* of simple heuristics – rather than their association with biases (*see* **Heuristics: Fast and Frugal**). We consider the changing and disputed interpretations given to another claimed judgmental bias next.

Overconfidence

In the 1970s and 1980s, a considerable amount of evidence was marshaled for the view that people suffer from an overconfidence bias. Typical laboratory studies of calibration ask participants to answer question such as

'Which is the world's longest canal?' (a) Panama
(b) Suez

Participants are informed that one of the answers is correct and are then required to indicate the answer that they think is correct and state how confident they are on a probability scale ranging from 50 to 100% (as one of the answers is always correct, 50% is the probability of guessing correctly). To be well calibrated, assessed probability should equal percentage correct over a number of assessments of equal probability. For example, if you assign a probability of 70% to each of 10 predictions, then you should get 7 of those predictions correct. Typically, however, people tend to give overconfident responses – their average confidence is higher than their proportion of correct answers. For a full review of this aspect of probabilistic judgment, see [25] and [14].

Overconfidence of judgments made under uncertainty is commonly found in calibration studies and has been recorded in the judgments of experts. For example, Christensen-Szalanski and Bushyhead [3] explored the validity of the probabilities given by physicians to diagnoses of pneumonia. They found that the probabilities were poorly calibrated and very overconfident; the proportion of patients who turned out to have pneumonia was far less than the probability statements implied. These authors had previously established that the physicians' estimates of the probability of a patient having pneumonia were significantly correlated with their decision to give a patient a chest X ray and to assign a pneumonia diagnosis.

Wagenaar and Keren [33] found overconfidence in lawyers' attempts to anticipate the outcome of court trials in which they represented one side. As they point out, it is inconceivable that the lawyers do not pay attention to the outcomes of trials in which they have participated, so why do they not learn to make well-calibrated judgments? Nonetheless, it is possible that the circumstances in which the lawyers, and other experts, make their judgments, and the circumstances in which they receive feedback, combine to impede the proper monitoring of feedback necessary for the development of well-calibrated judgments. A consideration of the reports of well-calibrated experts supports this notion; they all appear to be cases where some explicit unambiguous quantification of uncertainty is initially made and the outcome feedback is prompt and unambiguous.

The most commonly cited example of well-calibrated judgments is weather forecasters' estimates of the likelihood of precipitation [26], but there are a few other cases. Keren [22] found highly experienced tournament bridge players (but not experienced nontournament players) made well-calibrated forecasts of the likelihood that a contract, reached during the bidding phase, would be made, and Phillips [28] reports well-calibrated forecasts of horse races by bookmakers. In each of these three cases, the judgments made by the experts are precise numerical statements and the outcome feedback is unambiguous and received promptly and so can be easily compared with the initial forecast. Under these circumstances, the experts are unlikely to be insensitive to the experience of being surprised; there is very little scope for neglecting, or denying, any mismatch between forecast and outcome.

However, 'ecological' theorists (cf. [25]) claim that overconfidence is an artifact of the artificial experimental tasks and the nonrepresentative sampling of stimulus materials. Gigerenzer et al. [11] and Juslin [16] claim that individuals are well adapted to their environments and do not make biased judgments. Overconfidence is observed because the typical general knowledge quiz used in most experiments contains a disproportionate number of misleading items. These authors have found that when knowledge items are randomly sampled, the overconfidence phenomenon disappears. For example, Gigerenzer et al. [11] presented their participants with items generated with random pairs of the German cities with more than 100 000 inhabitants and asked them to select the biggest and indicate their confidence they had done so correctly. With this randomly sampled set of items, there was no overconfidence.

Moreover, with conventional general knowledge quizzes, participants are aware of how well they are likely to perform overall. Gigerenzer et al. [11] found that participants are really quite accurate at indicating *the proportion of items* that they have correctly answered. Such quizzes are representative of general knowledge quizzes experienced in the past. Thus, even when they appear overconfident with their answers to the individual items, participants are not overconfident about their performance on the same items as a set. Note that this observation is consistent with Gigerenzer's claim that though people may be poor at representing and so reasoning with probabilities about single events they can effectively infer probabilities when represented as frequencies.

Juslin et al. [17] report a **meta-analysis** comparing 35 studies, where items were randomly selected from a defined domain, with 95 studies where items were selected by experimenters. While overconfidence was evident for selected items, it was close to zero for randomly sampled items, which suggests that overconfidence is not simply a ubiquitous cognitive bias. This analysis suggests that the appearance of overconfidence may be an illusion created by research and not a cognitive failure by respondents.

Moreover, in cases of judgments of repeated events (weather forecasters, horse race bookmakers, tournament bridge players), experts make well-calibrated forecasts. In these cases, respondents might be identifying relative frequencies for sets of similar events rather than judging likelihood for individual events. And, if we compare studies of the calibration of probability assessments concerning individual events (e.g., [36]) with those where subjective assessments have been made for repetitive predictions of events [26], we observe that relatively poor calibration has been observed in the former, whereas relatively good calibration has been observed in the latter.

Another idea relevant to the interpretation of the evidence of overconfidence comes from Erev, Wallsten, and Budescu [6], who have suggested that overconfidence may, to some degree, reflect an underlying stochastic component of judgment. Any degree of error variance in judgment would create a regression that appears as overconfidence in the typical calibration analysis of judgment. When any two variables are not perfectly correlated – and confidence and accuracy are not perfectly correlated – there will be a regression effect. So, it is that a sample of the (adult) sons of very tall fathers will, on average, be shorter than their fathers, and, *at the same time,* a sample of the fathers of very tall sons will, on average, be shorter than their sons.

Exploring this idea, Budescu, Erev, and Wallsten [2] presented a generalization of the results from the Erev et al. [6] article, which shows that overconfidence and its apparent opposite, underconfidence, can be observed *simultaneously* in one study, depending upon whether probabilistic judgments are analyzed by conditionalizing accuracy as a function of confidence (the usual method showing overconfidence) or vice versa.

Conclusions

Although there has been a substantial amassing of evidence for the view that humans are inept at dealing with uncertainty using judged – subjective – probability, we also find evidence for a counterargument. It seems that disparities with basic requirements of probability theory can be observed when people are asked to make judgments of probability as a measure of propensity or strength of belief. The counterargument proposes that people may be very much better at reasoning under uncertainty than this research suggests when they are presented with tasks in a manner that permits them to conceive of probability in frequentist terms. This debate is currently unresolved and highly contentious. Nevertheless, for those with the hope of using subjective probabilities as inputs into decision support systems, we hope we have gone some way toward demonstrating that human judgments of uncertainty are worth considering as a valuable resource, rather than as objects to be regarded with suspicion or disdain.

References

[1] Bar-Hillel, M. (1980). The base-rate fallacy in probability judgements, *Acta Psychologica* **44**, 211–233.

[2] Budescu, D.V., Erev, I. & Wallsten, T.S. (1997). On the importance of random error in the study of probability judgment: Part I: new theoretical developments, *Journal of Behavioral Decision Making* **10**, 157–171.

[3] Christensen-Szalanski, J.J.J. & Bushyhead, J.B. (1981). Physicians use of probabilistic information in a real clinical setting, *Journal of Experimental Psychology, Human Perception and Performance* **7**, 928–935.

[4] DuCharme, W.M. & Peterson, C.R. (1968). Intuitive inference about normally distributed populations, *Journal of Experimental Psychology* **78**, 269–275.

[5] Edwards, W. (1968). Conservatism in human information processing, in *Formal Representation of Human Judgment*, B. Kleinmuntz, ed., Wiley, New York.

[6] Erev, I., Wallsten, T.S. & Budescu, D.V. (1994). Simultaneous over-and underconfidence: the role of error in judgment processes, *Psychological Review* **101**, 519–528.

[7] Evans, J.St.B.T., Handley, S.J., Perham, N., Over, D.E. & Thompson, V.A. (2000). Frequency versus probability formats in statistical word problems, *Cognition* **77**, 197–213.

[8] Gigerenzer, G. (1994). Why the distinction between single event probabilities and frequencies is important for Psychology and vice-versa, in *Subjective Probability*, G. Wright & P. Ayton, eds, Wiley, Chichester.

[9] Gigerenzer, G. (1996). On narrow norms and vague heuristics: a rebuttal to Kahneman and Tversky, *Psychological Review* **103**, 592–596.

[10] Gigerenzer, G., Hell, W. & Blank, H. (1988). Presentation and content: the use of base rates as a continuous variable, *Journal of Experimental Psychology: Human Perception and Performance* **14**, 513–525.

[11] Gigerenzer, G., Hoffrage, U. & Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence, *Psychological Review* **98**, 506–528.

[12] Girotto, V. & Gonzales, M. (2001). Solving probabilistic and statistical problems: a matter of information structure and question form, *Cognition* **78**, 247–276.

[13] Girotto, V. & Gonzales, M. (2002). Chances and frequencies in probabilistic reasoning: rejoinder to Hoffrage, Gigerenzer, Krauss, and Martignon, *Cognition* **84**, 353–359.

[14] Harvey, N. (1998). Confidence in judgment, *Trends in Cognitive Sciences* **1**, 78–82.

[15] Hoffrage, U., Gigerenzer, G., Krauss, S. & Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not, *Cognition* **84**, 343–352.

[16] Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items, *Organizational Behavior and Human Decision Processes* **57**, 226–246.

[17] Juslin, P., Winman, A. & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: a critical examination of the hard-easy effect, *Psychological Review* **107**, 384–396.

[18] Kahneman, D., Slovic, P. & Tversky, A., eds (1982). *Judgement Under Uncertainty: Heuristics and Biases*, Cambridge University Press.

[19] Kahneman, D. & Tversky, A. (1973). On the psychology of prediction, *Psychological Review* **80**, 237–251.

[20] Kahneman, D. & Tversky, A. (1982). On the study of statistical intuitions, in *Judgement Under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic & A. Tversky, eds, Cambridge University Press.

[21] Kahneman, D. & Tversky, A. (1996). On the reality of cognitive illusions: a reply to Gigerenzer's critique, *Psychological Review* **103**, 582–591.

[22] Keren, G.B. (1987). Facing uncertainty in the game of bridge: a calibration study, *Organizational Behavior and Human Decision Processes* **39**, 98–114.

[23] Koehler, J.J. (1995). The base-rate fallacy reconsidered - descriptive, normative, and methodological challenges, *Behavioral and Brain Sciences* **19**, 1–55.

[24] Macchi, L. (2000). Partitive formulation of information in probabilistic problems: Beyond heuristics and frequency format explanations, *Organizational Behavior and Human Decision Processes* **82**, 217–236.

[25] McClelland, A.G.R. & Bolger, F. (1994). The calibration of subjective probabilities: theories and models 1980–1994, in *Subjective Probability*, G. Wright & P. Ayton, eds, Wiley, New York.

[26] Murphy, A.H. & Winkler, R.L. (1984). Probability forecasting in meteorology, *Journal of the American Statistical Association* **79**, 489–500.

[27] Nisbett, R. & Ross, L. (1980). *Human Inference: Strategies and Shortcomings*, Prentice Hall, Englewood Cliffs.

[28] Phillips, L.D. (1987). On the adequacy of judgmental probability forecasts, in *Judgmental Forecasting*, G. Wright & P. Ayton, eds, Wiley, Chichester.

[29] Phillips, L.D. & Edwards, W. (1966). Conservatism in simple probability inference tasks, *Journal of Experimental Psychology* **72**, 346–357.

[30] Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases, *Science* **185**, 1124–1131.

[31] Tversky, A. & Kahneman, D. (1982). Evidential impact of base rates, in *Judgement Under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic & A. Tversky, eds, Cambridge University Press.

[32] Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment, *Psychological Review* **90**, 293–315.

[33] Wagenaar, W.A. & Keren, G.B. (1986). Does the expert know? The reliability of predictions and confidence ratings of experts, in *Intelligent Decision Support In Process Environments*, E. Hollnagel, G. Mancini & D.D. Woods, eds, Springer-Verlag, Berin.

[34] Wilson, T.D., Houston, C.E., Etling, K.M. & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents, *Journal of Experimental Psychology: General* **125**, 387–402.

[35] Winkler, R.L. & Murphy, A.M. (1973). Experiments in the laboratory and the real world, *Organizational Behavior and Human Performance* **20**, 252–270.

[36] Wright, G. & Ayton, P. (1992). Judgmental probability forecasting in the immediate and medium term, *Organizational Behavior and Human Decision Processes* **51**, 344–363.

Further Reading

Savage, L. (1954). *The Foundations of Statistics*, Wiley, New York.

(*See also* **Heuristics: Fast and Frugal**)

PETER AYTON

Summary Measure Analysis of Longitudinal Data

Brian S. Everitt

Volume 4, pp. 1968–1969

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Summary Measure Analysis of Longitudinal Data

There are a variety of approaches to the analysis of **longitudinal data**, including **linear mixed effects models** and **generalized estimating equations**. But many investigators may prefer (initially at least) to use a less complex procedure. One that may fit the bill is summary measure analysis, the essential feature of which is the reduction of the repeated response measurements available on each individual in the study, to a single number that is considered to capture an essential feature of the individual's response over time. In this way, the multivariate nature of the repeated observations is transformed to univariate. The approach has been in use for many years – see [3].

Choosing a Summary Measure

The most important consideration when applying a summary measure analysis is the choice of a suitable summary measure, a choice that needs to be made before any data are collected. The measure chosen needs to be relevant to the particular questions of interest in the study and in the broader scientific context in which the study takes place. A wide range of summary measures have been proposed as we see in Table 1. In [1], it is suggested that the average response over time is often likely to be the most relevant, particularly in *intervention studies* such as **clinical trials**.

Having chosen a suitable summary measure, the analysis of the longitudinal data becomes relatively straightforward. If two groups are being compared and normality of the summary measure is thought to be a valid assumption, then an independent samples *t* Test can be used to test for a group difference, or (preferably) a **confidence interval** for this difference can be constructed in the usual way. A one-way **analysis of variance** can be applied when there are more than two groups if again the necessary assumptions of normality and homogeneity hold. If the distributional properties of the selected summary measure are such that normality seems difficult to justify, then nonparametric analogues of these procedures might be used (*see* **Catalogue of Parametric Tests**; **Distribution-free Inference, an Overview**).

An Example of Summary Measure Analysis

The summary measure approach can be illustrated using the data shown in Table 2 that arise from a study of alcohol dependence. Two groups of subjects, one with severe dependence and one with moderate dependence on alcohol, had their salsolinol excretion levels (in millimoles) recorded on four consecutive days. (Salsolinol is an alkaloid with a structure similar to heroin.)

Using the mean of the four measurements available for each subject as the summary measure followed by the application of a *t* Test and the construction of a **confidence interval** leads to the results

Table 1 Possible summary measures (taken from [2])

Type of data	Question of interest	Summary measure
Peaked	Is overall value of outcome variable the same in different groups?	Overall mean (equal time intervals) or area under curve (unequal intervals)
Peaked	Is maximum (minimum) response different between groups?	Maximum (minimum) value
Peaked	Is time to maximum (minimum) response different between groups?	Time to maximum (minimum) response
Growth	Is rate of change of outcome different between groups?	Regression coefficient
Growth	Is eventual value of outcome different between groups?	Final value of outcome or difference between last and first values or percentage change between first and last values
Growth	Is response in one group delayed relative to the other?	Time to reach a particular value (e.g., a fixed percentage of baseline)

Table 2 Salsolinol excretion data

	Day			
Subject	1	2	3	4
Group 1 (Moderate dependence)				
1	0.33	0.70	2.33	3.20
2	5.30	0.90	1.80	0.70
3	2.50	2.10	1.12	1.01
4	0.98	0.32	3.91	0.66
5	0.39	0.69	0.73	3.86
6	0.31	6.34	0.63	3.86
Group 2 (Severe dependence)				
7	0.64	0.70	1.00	1.40
8	0.73	1.85	3.60	2.60
9	0.70	4.20	7.30	5.40
10	0.40	1.60	1.40	7.10
11	2.50	1.30	0.70	0.70
12	7.80	1.20	2.60	1.80
13	1.90	1.30	4.40	2.80
14	0.50	0.40	1.10	8.10

Table 3 Results from using the mean as a summary measure for the data in Table 2

	Moderate	Severe
Mean	1.80	2.49
sd	0.60	1.09
n	6	8
$t = -1.40$, df = 12, $p = 0.19$, 95% CI: [−1.77, 0.39]		

shown in Table 3. There is no evidence of a group difference in salsolinol excretion levels.

A possible alternative to the use of the mean as summary measure is the maximum excretion level recorded over the four days. Testing the null hypothesis that the measure has the same median in both populations using a Mann-Whitney (*see* **Wilcoxon–Mann–Whitney Test**) test results in a test statistic of 36 and associated P value of 0.28. Again, there is no evidence of a difference in salsolinol excretion levels in the two groups.

Problems with the Summary Measure Approach

In some situations, it may be impossible to identify a suitable summary measure and where interest centers on assessing details of how a response changes over time and how this change differs between groups, then summary measure analysis has little to offer. The summary measure approach to the analysis of longitudinal data can accommodate **missing data** by simply using only the available observations in its calculation, but the implicit assumption is that values are missing completely at random. Consequently, in, say, clinical trials in which a substantial number of participants drop out before the scheduled end of the trial, the summary measure approach is probably not to be recommended.

References

[1] Frison, L. & Pocock, S.J. (1992). Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design, *Statistics in Medicine* **11**, 1685–1704.

[2] Matthews, J.N.S., Altman, D.G., Campbell, M.J. & Royston, P. (1989). Analysis of serial measurements in medical research, *British Medical Journal* **300**, 230–235.

[3] Oldham, P.D. (1962). A note on the analysis of repeated measurements of the same subjects, *Journal of Chronic Disorders* **15**, 969–977.

(*See also* **Repeated Measures Analysis of Variance**)

BRIAN S. EVERITT

Survey Questionnaire Design

Roger Tourangeau

Volume 4, pp. 1969–1977

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Survey Questionnaire Design

Introduction

Survey statisticians have long distinguished two major sources of error in survey estimates – sampling error and measurement error. Sampling error arises because the survey does not collect data from the entire population and the characteristics of the sample may not perfectly match those of the population from which it was drawn (*see* **Survey Sampling Procedures**). Measurement error arises because the information collected in the survey differs from the true values for the variables of interest. The discrepancies between survey reports and true values can arise because the survey questions measure the wrong thing or because they measure the right thing but do it imperfectly. For example, the survey designers may want to measure unemployment, and, in fact, most developed countries conduct regular surveys to monitor employment and unemployment rates. Measuring unemployment can be tricky. How does one classify workers with a job but on extended sick leave? A major problem is asking the right questions, the questions that are needed to classify each respondent correctly. Another major problem is inaccurate reporting. Even if the questions represent the concept of interest, respondents may still not report the right information because they misunderstand the questions or they do not know all the relevant facts. For example, they may not know about the job search activities of other family members. The goal of survey questionnaire design is simple – it is to reduce such measurement errors to a minimum, subject to whatever cost constraints apply to the survey.

There are two basic methods for attaining this goal. First, questionnaire designers attempt to write questions that follow well-established principles for survey questions. Texts with guidelines for writing survey questions have been around for at least fifty years, appearing at about the same time as the first texts on survey sampling (for an early example, see [9]). Initially, these texts offered guidelines that were based on the experiences of the authors, but over the years a large base of methodological research has accumulated and this work has provided an empirical foundation for the questionnaire design guidelines (*see*, e.g. [14]). In addition, over the last twenty years or so, survey researchers have drawn more systematically on research in cognitive psychology to understand how respondents answer questions in surveys. This work, summarized in [15] and [17], has provided a theoretical grounding for the principles of questionnaire design. Traditionally, survey researchers have thought of writing survey questions as more on an art than a science, but, increasingly, because of the empirical and theoretical advances of the last twenty years, survey researchers have begun referring to the science of asking questions [13].

Aside from writing good questions in the first place, the other strategy for minimizing measurement error is to test survey questions and cull out or improve questions that do not seem to yield accurate information. There are several tools questionnaire designers use in developing and testing survey questions. These include both pretesting methods such as cognitive interviews or pilot tests used before the survey questionnaire is fielded, and methods that can be applied as the survey is carried out such as recontacting some of the respondents and asking some questions a second time.

The Questionnaire Design Process

Questionnaire design encompasses four major activities. The first step often consists of library research, in which the questionnaire designers look for existing items with desirable measurement properties. There are few general compilations of existing survey items. As a result, survey researchers generally rely on subject matter experts and earlier surveys on the same topic as sources for existing questions.

The next step is to assemble the questions into a draft questionnaire; the draft is usually a blend of both existing and newly written items. The initial draft often undergoes some sort of expert review. Questionnaire design experts may review the questionnaire to make sure that the questions adhere to questionnaire design principles and that they are easy for the interviewers to administer and for the respondents to understand and answer. Subject matter experts may review the draft to make sure that the survey will yield all the information needed in the analysis and that the questions correspond to the concepts of interest. For instance, the concept of unemployment involves both not having a job and wanting one; the survey questions must adequately cover both aspects of the concept. Subject matter experts are typically in the best position to decide whether the survey questions as drafted will meet the analytical requirements of the survey. When the questionnaire includes questions about a new topic, the questionnaire designers may also conduct one or more focus groups, a procedure described in more detail below, to discover how members of the survey population think about the topic of interest and the words and phrases they use in talking about it. Respondents are more likely to answer the questions accurately when the questions match their experiences and circumstances, and when they use familiar terminology.

Rarely does a survey questionnaire go into the field without having been pretested in some way. Thus, the third step in the questionnaire design process typically involves testing, evaluating, and revising the questionnaire prior to conducting the survey. Two types of pretesting are commonly used. The first is called *cognitive interviewing*. Cognitive interviews are generally conducted in a centralized laboratory setting rather than in the field. The purpose of these interviews is to discover how respondents answer the questions and whether they encounter any cognitive difficulties in formulating their answers. The cognitive interviewers administer a draft of the survey questions. They may encourage the respondents to think out loud as they answer the questions or they may administer follow-up probes designed to explore potential problems with the draft survey items. The second type of pretest is a pilot test, or a small-scale version of the main study. Pretest interviewers may interview 50 to 100 respondents. The size of the pretest sample often reflects the size and budget of the main survey. The survey designers may evaluate this trial run of the questionnaire by drawing on several sources of information. Often, they examine the pilot test responses, looking for items with low variances or high rates of missing data. In addition, the questionnaire designers may conduct a 'debriefing' with the pilot test interviewers, eliciting their input on such matters as the items that seemed to cause problems for the interviewers or the respondents. Pilot tests sometimes incorporate experiments comparing two or more methods for asking the questions and, in such cases, the evaluation of the pretest results will include an assessment of the experimental results. Or the pilot test may include recording or monitoring the interviews. The point of such monitoring is to detect items that interviewers often do not read as written or items that elicit frequent requests for clarification from the respondents. On the basis of the results of cognitive or pilot test interviews, the draft questionnaire may be revised, often substantially. The pilot test may reveal other information about the questionnaire, such as the average time needed to administer the questions, which may also feed into the evaluation and revision of the questions.

The final step is to administer questions in the real survey. The evaluation of the questions does not necessarily stop when the questionnaire is fielded, because the survey itself may collect data that are useful for evaluating the questions. For example, some education surveys collect data from both the students and their parents, and the analysis can assess the degree that the information from the two sources agrees upon. Low levels of agreement between sources would suggest high levels of measurement error in one or both sources. Similarly, surveys on health care may collect information both from the patient and from medical records, allowing an assessment of the accuracy of the survey responses. Many surveys also recontact some of the respondents to make sure the interviewers have not fabricated the data; these 'validation' interviews may readminister some of the original questions, allowing an assessment of the reliability of the questions. Thus, the final step in the development of the questionnaire is sometimes an after-the-fact assessment of the questions, based on the results of the survey.

Tools for Testing and Evaluating Survey Questions

Our overview of the questionnaire design process mentioned a number of tools survey researchers use in developing and testing survey questions. This section describes these tools – expert reviews, focus groups, cognitive testing, pilot tests, and split-ballot experiments – in more detail.

Expert Reviews. Expert reviews refer to two distinct activities. One type of review is carried out by substantive experts or even the eventual analysts of the survey data. The purpose of these substantive reviews is to ensure that the questionnaire collects all the information needed to meet the analytic objectives of the survey. The other type of review features questionnaire design experts, who review the wording of the questions, the response format and the particular response options offered, the order of the questions, the instructions to interviewers for administering the questionnaire, and the navigational instructions (e.g., 'If yes, please go to Section B'). Empirical evaluations [10] suggest that questionnaire design experts often point out a large number of problems with draft questionnaires.

Sometimes, the experts employ formal checklists of potential problems with questions. Several checklists are available. Lessler and Forsyth [8], for example, present a list of 25 types of potential problems with questions. Such checklists are generally derived from a cognitive analysis of the survey response process and the problems that can arise during that process (see, e.g., [15] and [17]). The checklists often distinguish various problems in comprehension of the questions, the recall of information needed to answer the questions, the use of judgment and estimation strategies, and the reporting of the answer. One set of researchers [6] has even developed a computer program that diagnoses 12 major problems with survey questions, most of them involving comprehension issues, thus providing an automated, if preliminary, expert appraisal. To illustrate the types of problems included in the checklists, here are the 12 detected by Graesser and his colleagues' program:

1. complex syntax,
2. working memory overload,
3. vague or ambiguous noun phrase,
4. unfamiliar technical term,
5. vague or imprecise predicate or relative term,
6. misleading or incorrect presupposition,
7. unclear question category,
8. amalgamation of more than one question category,
9. mismatch between the question category and the answer option,
10. difficulty in accessing (that is, recalling) information,
11. respondent unlikely to know answer, and
12. unclear question purpose.

Focus Groups. Before they start writing the survey questions, questionnaire designers often listen to volunteers discussing the topic of the survey. These focus group discussions typically include 6 to 10 members of the survey population and a moderator who leads the discussion. Questionnaire designers often use focus groups in the early stages of the questionnaire design process to learn more about the survey topic and to discover how the members of the survey population think and talk about it (*see* **Focus Group Techniques**).

Suppose, for example, one is developing a questionnaire on medical care. It is useful to know what kinds of doctors and medical plans respondents actually use and it is also useful to know whether they are aware of the differences between HMOs, other types of managed care plans, and fee-for-service plans. In addition, it is helpful to know what terminology they use in describing each sort of plan and how they describe different types of medical visits. The more that the questions fit the situations of the respondents, the easier it will be for the respondents to answer them. Similarly, the more closely the questions mirror the terminology that the respondents use in everyday life, the more likely it is that the respondents will understand the questions as intended.

Focus groups can be an efficient method for getting information from several people in a short period of time, but the method does have several limitations. Those who take part in focus groups are typically volunteers and they may or may not accurately represent the survey population. In addition, the number of participants in a focus group may be a poor guide to the amount of information produced by the discussion. No matter how good the moderator is, the discussion often reflects the views of the most articulate participants. In addition, the discussion may veer off onto some tangent, reflecting the group dynamic rather than the considered views of the participants. Finally, the conclusions from a focus group discussion are often simply the impressions of the observers; they may be unreliable and subject to the biases of those conducting the discussions.

Cognitive Interviewing. Cognitive interviewing is a family of methods designed to reveal the strategies that respondents use in answering survey questions. It is descended from a technique called 'protocol analysis', invented by Herbert Simon and his colleagues (*see*, e.g., [5]). Its purpose is to explore how people deal with higher-level cognitive problems, like solving chess problems or proving algebraic theorems. Simon asked his subjects to think aloud as they worked on such problems and recorded what they said. These verbalizations were the 'protocols' that Simon and his colleagues used in testing their hypotheses about problem solving. The term 'cognitive interviewing' is used somewhat more broadly to cover a range of procedures, including:

1. concurrent protocols, in which respondents verbalize their thoughts while they answer a question;
2. retrospective protocols, in which they describe how they arrived at their answers after they provide them;
3. confidence ratings, in which they rate their confidence in their answers;
4. definitions of key terms, in which respondents are asked to define terms in the questions;
5. paraphrasing, in which respondents restate the question in their own words, and
6. follow-up probes, in which respondents answer questions designed to reveal their response strategies.

This list is adopted from a longer one found in [7]. Like focus groups, cognitive interviews are typically conducted with paid volunteers. The questionnaires for cognitive interviews include both draft survey questions and prescripted probes designed to reveal how the respondents understood the questions and arrived at their answers. The interviewers may also ask respondents to think aloud as they answer some or all of the questions. The interviewers generally are not field interviewers but have received special training in cognitive interviewing.

Although cognitive interviewing has become a very popular technique, it is not very well standardized. Different organizations emphasize different methods and no two interviewers conduct cognitive interviews in exactly the same way. Cognitive interviews share two of the main drawbacks with focus groups. First, the samples of respondents are typically volunteers so the results may not be representative of the survey population. Second, the conclusions from the interviews are often based on the impressions of the interviewers rather than objective data such as the frequency with which specific problems with an item are encountered.

Pilot Tests. Pilot tests or field tests of a questionnaire are mini versions of the actual survey conducted by field interviewers under realistic survey conditions. The pilot tests use the same mode of data collection as the main survey. For instance, if the main survey is done over the telephone, then the pilot test is done by the telephone as well. Pilot tests have some important advantages over focus groups and cognitive interviews; they often use probability samples of the survey population and they are done using the same procedures as the main survey. As a result, they can provide information about the data collection and sampling procedures – Are they practical? Do the interviews take longer than planned to complete? – as well as information about the draft questionnaire.

Pilot tests typically yield two main types of information about the survey questionnaire. One type consists of the feedback from the pilot test interviewers. The reactions of the interviewers are often obtained in an interviewer debriefing, where some or all of the field test interviewers meet for a discussion of their experiences with the questionnaire during the pilot study. The questionnaire designers attend these debriefing sessions to hear the interviewers present their views about questions that do not seem to be working and other problems they experienced during the field test. The second type of information from field test is the data – the survey responses themselves. Analysis of the pretest data may produce various signs diagnostic of questionnaire problems, such as items with high rates of missing data, out-of-range values, or inconsistencies with other questions. Such items may be dropped or rewritten.

Sometimes pilot tests gather additional quantitative data that is useful for evaluating the questions. These data are derived from monitoring or recording the pilot test interviews and then coding the interchanges between the interviewers and respondents. Several schemes have been developed for systematically coding these interactions. Fowler and Cannell [4] have proposed and applied a simple scheme for assessing survey questions. They argue that coders should record for each question whether the interviewer read the question exactly as worded, with minor changes, or with major changes that altered the meaning of the questions. In addition, the coders should record whether the respondent interrupted the interviewer before he or she had finished reading the question, asked for clarification of the question, and gave an adequate answer, a don't know or refusal response, or an answer that required further probing from the interviewer. Once the interviews have been coded, the questionnaire designers can examine a variety of statistics for each item, including the percentage of times the interviewers departed from the verbatim text in administering the item, the percentage of respondents who asked for clarification of the question, and the percentage of times the interviewer had to ask additional questions to obtain an acceptable answer. All three behaviors are considered signs of poorly written questions.

Relative to expert reviews, focus groups, and cognitive testing, field tests, especially when supplemented by behavior coding, yields data that are objective, quantitative, and replicable.

Split-ballot Experiments. A final method used in developing and testing questionnaires is the split-ballot experiment. In a split-ballot experiment, random subsamples of the respondents receive different versions of the questions or different methods of data collection, for example, self-administered questions versus questions administered by interviewers, or both. Such experiments are generally conducted as part of the development of questionnaires and procedures for large-scale surveys, and they may be embedded in a pilot study for such surveys. A discussion of the design issues raised by such experiments can be found in [16]; examples can be found in [3] and [16].

Split-ballot experiments have the great virtue that they can clearly show what features of the questions or data collection procedures affect the answers. For example, the experiment can compare different wordings of the questions or different question orders. But the fact that two versions of the questionnaire produce different results does not necessarily resolve the question of which version is the right one to use. Experiments can produce more definitive results when they also collect some external validation data that can be used to measure the accuracy of the responses. For example, in a study of medical care, the researchers can compare the survey responses to medical records, thereby providing some basis for deciding which version of the questions yielded the more accurate information. In other cases, the results of a methodological experiment may be unambiguous even in the absence of external information. Sometimes there is a strong *a priori* reason for thinking that reporting errors follow a particular pattern or direction. For example, respondents are likely to underreport embarrassing or illegal behaviors, like illicit drug use. Thus, a strong reason for thinking that a shift in a specific direction (such as higher levels of reported drug use) represents an increase in accuracy.

The major drawback to methodological experiments is their complexity and expense. Detecting even moderate differences between experimental groups may require substantial numbers of respondents. In addition, such experiments add to the burden on the survey designers, requiring them to develop multiple questionnaires or data collection protocols rather than just one. The additional time and expense of this effort may exceed the budget for questionnaire development or delay the schedule for the main survey too long. For these reasons, split-ballot experiments are not a routine part of the questionnaire design process.

Combining the Tools. For any particular survey, the questionnaire design effort is likely to employ several of these tools. The early stages of the process are likely to rely on relatively fast and inexpensive methods such as expert reviews. Depending on how much of the questionnaire asks about a new topic, the questionnaire designers may also conduct one or more focus groups. If the survey is fielding a substantial number of new items, the researchers are likely to conduct one or more rounds of cognitive testing and a pilot test prior to the main survey. Typically, the researchers analyze the pilot study data and carry out a debriefing of the pilot test interviewers. They may supplement this with the coding and analysis of data on the exchanges between respondents and interviewers. If there are major unresolved questions about how the draft questions should be organized or worded, the survey designers may conduct an experiment to settle them.

The particular combination of methods used in developing and testing the survey questionnaire for a given survey will reflect several considerations, including the relative strengths and weaknesses of the different methods, the amount of time and money available for the questionnaire design effort, and the issues that concern the survey designers the most. The different questionnaire design tools yield different kinds of information. Cognitive interviews, for example, provide information about respondents' cognitive difficulties with the questions but are less useful for deciding how easy it will be for interviewers to administer the questions in the field. Expert reviews are a good, all-purpose tool but they yield educated guesses rather than objective data on how the questions are likely to work. Pilot tests are essential if there are concerns about the length of the questionnaire or other practical issues that only can be addressed by a dry run of the questionnaire and data collection procedures under realistic field conditions. If the questionnaire requires the use of complicated response aids or new software for administering the questions, a field test and interviewer debriefing are likely to be deemed essential. And, of course, decisions about what tools to use and how many rounds of the testing to carry out are likely to reflect the overall survey budget, the amount of prior experience with this or similar questionnaires, and other factors related to cost and schedule.

Standards for Evaluating Questions

Although the goal for questionnaire design is straightforward in principle – to minimize survey error and cost – as a practical matter, the researchers may have to examine various indirect measures of cost or quality. In theory, the most relevant standards for judging the questions are their reliability and validity; but such direct measures of error are often not available and the researchers are forced to fall back on such indirect indicators of measurement error as the results of cognitive interviews. This section takes a more systematic look at the criteria questionnaire designers use in judging survey questions.

Content Standards. One important nonstatistical criterion that the questionnaire must meet is whether it covers all the topics of interest and yields all the variables needed in the analysis. It does not matter much whether the questions elicit accurate information if it is not the right information. Although it might seem a simple matter to make sure the survey includes the questions needed to meet the analytical objectives, there is often disagreement about the best strategy for measuring a given concept and there are always limits on the time or space available in a questionnaire. Thus, there may be deliberate compromises between full coverage of a particular topic of interest and cost. To keep the questionnaire to a manageable length, the designers may include a subset of the items from a standard battery in place of the full battery or they may explore certain topics superficially rather than in the depth they would prefer.

Statistical Standards: Validity and reliability. Of course, the most fundamental standards for a question are whether it yields consistent and accurate information. Reliability and validity (*see* **Validity Theory and Applications**) are the chief statistical measures of these properties.

The simplest mathematical model for a survey response treats it as consisting of components – a true score and an error:

$$Y_{it} = \mu_i + \varepsilon_{it}, \tag{1}$$

in which Y_{it} refers to the reported value for respondent i on occasion t, μ_i refers to the true score for that respondent, and ε_{it} to the error for the respondent on occasion t (*see* **Measurement: Overview**). The true score is the actual value for the respondent on the variable of interest and the error is just the discrepancy between the true score and the reported value. The idea of a true score makes more intuitive sense when the variable involves some readily verifiable fact or behavior – say, the number of times the respondent visited a doctor in the past month. Still, many survey researchers find the concept useful even for subjective variables, for example, how much the respondent favors or opposes some policy. In such cases, the true score is defined as the mean across the hypothetical population of measures for the concept that are on the same scale (*see*, e.g., [1]).

Several assumptions are often made about the errors. The simplest model assumes first that, for any given respondent, the expected value of the errors is zero and, second, that the correlation between the errors for any two respondents or between those for the same respondent on any two occasions is zero. The validity of the item is usually defined as the correlation between Y_{it} and μ_i. The reliability is defined as the correlation between Y_{it} and $Y_{it'}$, where t and t' represent two different occasions. Under the simplest model, it is easy to show that validity (V) is just:

$$\begin{aligned} V &= \frac{\mathrm{Cov}(Y,\mu)}{[\mathrm{Var}(Y)\mathrm{Var}(\mu)]^{1/2}} \\ &= \frac{\mathrm{Var}(\mu)}{[\mathrm{Var}(Y)\mathrm{Var}(\mu)]^{1/2}} \\ &= \frac{\mathrm{Var}(\mu)^{1/2}}{\mathrm{Var}(Y)^{1/2}} \end{aligned} \tag{2}$$

in which $\mathrm{Cov}(Y,\mu)$ is the covariance between the observed values and the true scores and $\mathrm{Var}(Y)$ and

Var(μ) are their variances. Under this model, the validity is just the square of the reliability.

As a practical matter, the validity of a survey item is often estimated by measuring the correlation between the survey reports and some external 'gold standard,' such as administrative records or some other measure of the variable of interest that is assumed to be error-free or nearly error-free. Reliability is estimated in one of two ways. For simple variables derived from a single item, the item may be administered to the respondent a second time in a reinterview. Rather than assessing the reliability by calculating the correlation between the two responses, survey researchers often calculate the gross discrepancy rate – the proportion of respondents classified differently in the original interview and the reinterview. This approach is particularly common when the survey report yields a simple categorical variable such as whether the respondent is employed or not. The gross discrepancy rate is a measure of *un*reliability rather than reliability. For variables derived from multi-item batteries, the average correlation among the items in the battery or some related index, such as Cronbach's alpha [2], is typically used to assess reliability.

The model summarized in (1) is often unrealistically simple and more sophisticated models relax one or more of its assumptions. For example, it is sometimes reasonable to assume that errors for two different respondents are correlated when the same interviewer collects the data from both. Other models allow the true scores and observed values to be on different scales of measurement or allow the observed scores to reflect the impact of other underlying variables besides the true score on the variable of interest. These other variables affecting the observed score might be other substantive constructs or measurement factors, such as the format of the questions (see [12] for an example).

Cognitive Standards. Most questionnaire design efforts do not yield direct estimates of the reliability or validity of the key survey items. As noted, the typical procedure for estimating validity is to compare the survey responses to some external measure of the same variable, and this requires additional data collection for each survey item to be validated. Even obtaining an estimate of reliability typically requires recontacting some or all of the original respondents and administering the items a second time. The additional data collection may exceed time or budget constraints. As a result, many survey design efforts rely on cognitive testing or other indirect methods to assess the measurement properties of a given survey item. The assumptions of this approach are that if respondents consistently have trouble understanding a question or remembering the information needed to answer it, then the question is unlikely to yield accurate answers. The evidence that respondents have difficulty comprehending the question, retrieving the necessary information, making the requisite judgments, and so on often comes from cognitive interviews. Alternatively, questionnaire design experts may flag the question as likely to produce problems for the respondents or evidence of such problems may arise from the behavior observed in the pilot interviews. For example, a high percentage of respondents may ask for clarification of the question or give inadequate answers. Whatever the basis for their judgments, the developers of survey questionnaires are likely to assess whether the draft questions seem to pose a reasonable cognitive challenge to the respondents or are too hard for respondents to understand or answer accurately.

Practical Standards. Surveys are often large-scale efforts and may involve thousands of respondents and hundreds of interviewers. Thus, a final test for a survey item or survey questionnaire is whether it can actually be administered in the field in a standardized way and at a reasonable cost. Interviewers may have difficulty reading long questions without stumbling; or they may misread questions involving unfamiliar vocabulary. Any instructions the interviewers are supposed to follow should be clear. Both individual items and the questionnaire as a whole should be easy for the interviewers to administer. The better the questionnaire design, the less training the interviewers will need. In addition, it is important to determine whether the time actually needed to complete the interview is consistent with the budget for the survey. These practical considerations are often the main reasons for conducting a pilot test of the questionnaire and debriefing the pilot interviewers.

Increasingly, survey questionnaires take the form of computer programs. The reliance on electronic questionnaires raises additional practical issues – Is the software user-friendly? Does it administer the items as the authors intended? Do interviewers or respondents make systematic errors in interacting with the program? The questionnaire design process may include an additional effort – usability testing – to address these practical questions.

Conclusion

Designing survey questionnaires is a complex activity, blending data collection and statistical analysis with subjective impressions and expert opinions. Well-validated principles for writing survey questions are gradually emerging and questionnaire designers can now consult a substantial body of evidence about how to ask questions. For a summary of recent developments, see [11]. Still, for any particular survey, the questionnaire designers often have to rely on low-cost indirect indicators of measurement error in writing specific questions. In addition, the questionnaire that is ultimately fielded is likely to represent multiple compromises that balance statistical considerations, such as reliability and validity, against practical considerations, such as length, usability, and cost. Survey questionnaire design remains a mix of art and science, a blend of practicality and principle.

References

[1] Biemer, P. & Stokes, L. (1991). Approaches to the modeling of measurement error, in *Measurement Errors in Surveys*, P. Biemer, R.M. Groves, L. Lyberg, N. Mathiowetz & S. Sudman, eds, John Wiley & Sons, New York.

[2] Cronbach, L. (1951). Coefficient alpha and the internal structure of tests, *Psychiatrika* **16**, 297–334.

[3] Fowler, F.J. (2004). Getting beyond pretesting and cognitive interviews: the case for more experimental pilot studies, in *Questionnaire Development Evaluation and Testing Methods*, S. Presser, J. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin & E. Singer, eds, John Wiley & Sons, New York.

[4] Fowler, F.J. & Cannell, C.F. (1996). Using behavioral coding to identify cognitive problems with survey questions, in *Answering Questions*, N.A. Schwarz & S. Sudman, eds, Jossey-Bass, San Francisco.

[5] Ericsson, K.A. & Simon, H.A. (1980). Verbal reports as data, *Psychological Review* **87**, 215–257.

[6] Graesser, A.C., Kennedy, T., Wiemer-Hastings, P. & Ottati, V. (1999). The use of computational cognitive models to improve questions on surveys and questionnaires, in *Cognition in Survey Research*, M.G. Sir ken, D.J. Herrmann, S. Schechter, N. Schwarz, J. Tanur & R. Tourangeau, eds, John Wiley & Sons, New York.

[7] Jobe, J.B. & Mingay, D.J. (1989). Cognitive research improves questionnaires, *American Journal of Public Health* **79**, 1053–1055.

[8] Lessler, J.T. & Forsyth, B.H. (1996). A coding system for appraising questionnaires, in *Answering Questions*, N.A. Schwarz & S. Sudman, eds, Jossey-Bass, San Francisco.

[9] Payne, S. (1951). *The Art of Asking Questions*, Princeton University Press, Princeton.

[10] Presser, S. & Blair, J. (1994). Survey pretesting: do different methods produce different results?, in *Sociological Methodology*, P.V. Marsden, ed., American Sociological Association, Washington.

[11] Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J. & Singer, E. (2004). *Questionnaire Development Evaluation and Testing Methods*, John Wiley & Sons, New York.

[12] Saris, W.E. & Andrews, F.M. (1991). Evaluation of measurement instruments using a structural modeling approach, in *Measurement Errors in Surveys*, P. Biemer, R.M. Groves, L. Lyberg, N. Mathiowetz & S. Sudman, eds, John Wiley & Sons, New York.

[13] Schaeffer, N.C. & Presser, S. (2003). The science of asking questions, *Annual Review of Sociology* **29**, 65–88.

[14] Sudman, S. & Bradburn, N. (1982). *Asking Questions*, Jossey-Bass, San Francisco.

[15] Sudman, S., Bradburn, N. & Schwarz, N. (1996). *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*, Jossey-Bass.

[16] Tourangeau, R. (2004),. Design considerations for questionnaire testing and evaluation, in S. Presser, J. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin & E. Singer, eds, *Questionnaire Development Evaluation and Testing Methods*, John Wiley, New York.

[17] Tourangeau, R., Rips, L.J. & Rasinski, K. (2000). *The Psychology of Survey Responses*, Cambridge University Press, New York.

(*See also* **Telephone Surveys**)

ROGER TOURANGEAU

Survey Sampling Procedures

Kris K. Moore

Volume 4, pp. 1977–1980

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Survey Sampling Procedures

Survey sampling provides a method of obtaining information about a population based on a sample selected from the population. A population is the totality of elements or individuals about which one wants to obtain information. A sample is a portion of the population that has been selected randomly. A sample characteristic is used to estimate the population characteristic. The cost of surveying all the elements in a population can be high, and the time for surveying each member of the population can be enormous. Sampling can provide a savings in cost and time.

Parameters or characteristics of the population are estimated from statistics that are characteristics of the sample. The estimates may differ by only ±2.5% from the actual population values. For example, in 2004, public opinion polls in the United States obtained information about approximately 300 million people from a sample of 1500 people. If in the sample of 1500, say, 600 favor the president's performance, the sample statistic is $600/1500 = 0.4$. The statistic 0.4 or 40% differs by only 2.5% from the population parameter 19 out of 20 times a sample of this size is taken. Thus, this sample would indicate that the interval 37.5 to 42.5% estimates the lower and upper bound of an interval that contains the percentage who favor the president's performance. About 95 out of 100 times a sample of this size (1500) would produce an interval that contains the population value.

Instrument of Survey Sampling

A questionnaire is usually designed to obtain information about a population from sampled values (*see* **Survey Questionnaire Design**). The questionnaire should be as brief as possible, preferably not more than two or three pages. The questionnaire should have a sponsor that is well known and respected by the sampled individuals. For example, if the population is a group of teachers, the teachers will be more likely to respond if the survey is sponsored by a recognizable and respected teacher organization. Also, money incentives such as 50¢ or a $1.00 included with the mailing improve the response rate on mailed questionnaires.

The questionnaire should begin with easy-to-answer questions; more difficult questions should be placed toward the end of the questionnaire. Demographic questions should be placed toward the beginning of the questionnaire because these questions are easy to answer. Care should be taken in constructing questions so that the questions are not offensive. For example rather than asking, 'How old are you' a choice of appropriate categories is less offensive. For example, 'Are you 20 or younger?' 'Are you 20 to 30?' and so on is less offensive. A few open-ended questions should be included to allow the respondent to clearly describe his or her position on issues not included in the body of the questionnaire.

Questionnaires should be tested in a small pilot survey before the survey is implemented to be sure the desired information is selected and the questions are easily understood. The pilot survey should include about 30 respondents.

Type of Question

Again, some open-ended questions should be used to allow the respondent freedom in defining his or her concerns. The Likert scale of measurement should be used on the majority of the questions. For example,

Strongly disagree	Mildly disagree	Disagree	Neutral	Agree	Mildly agree	Strongly agree
1	2	3	4	5	6	7

The Likert scale is easily converted to a form that computers can process.

The Frame

The sample in survey sampling is randomly selected from a frame or list of elements in the population. A random selection means each element of the population has an equal chance of being selected. Numbers can be assigned to each member of the population $1, 2, \ldots, N$, where N is the total number in the population. Then random numbers can be used to select the sample elements.

Often a list of the population elements does not exist. Other methods will work for this type of problem and are discussed in the Selection Methods section.

Collection of the Sample Information

The basic methods of obtaining responses in survey sampling are personal interview, mail questionnaire, telephone interview, or electronic responses via computers. Personal interviews are accurate but very expensive and difficult if the population includes a large geographic area. The person conducting the interview must be careful to be neutral and must not solicit preferred responses.

Mail questionnaires (*see* **Mail Surveys**) are inexpensive but the response rate is usually low, sometimes less than 10%. Incentives such as enclosing a dollar or offering the chance to win a prize increase response rates to 20 to 30%. Additional responses can be obtained by second and third mailings to those who failed to respond. Responses from first, second, and third mailings should be compared to see if trends are evident from one mailing to the other. For example, people who feel strongly about an issue may be more likely to respond to the first mailing than people who are neutral about an issue.

Telephone interviews are becoming more difficult to obtain because the general public has grown tired of tele-marketing and the aggravation of telephone calls at inconvenient times. In the United States, approximately 95% of the general population has telephones and of these, 10 to 15% of telephone numbers are unlisted. Random digit dialing allows unlisted numbers to be included in the sample by using the prefix for an area to be sampled, for example 756-XXXX. The XXXX is a four digit random number that is selected from a list of random numbers. This procedure randomly selects telephone numbers in 756 exchange.

Electronic methods of sampling are the most recently developed procedures (*see* **Internet Research Methods**). Internet users are sampled and incentives are used to produce a high response rate. Samples are easily selected at low cost via the computer.

Selection Methods

The simple random sample in which each element has an equal chance of selection is the most frequently used selection method when a frame or list of elements exists (see [2]). A systematic sample in which every kth element is selected is often easier to obtain than a random sample. If N is the number in the population and n the sample size, then $k = N/n$, where k is rounded to the nearest whole number. The starting point 1 to k is randomly selected. If the sampled elements are increasing in magnitude, for example, inventory ordered by value from low-cost items to high-cost items, a systematic sample is better than a random sample. If the elements to be sampled are periodic, for example sales Monday through Saturday, then a random sample is better than a systematic sample because a systematic sample could select the same day each time. If the population is completely random, then systematic and random sample produce equivalent results.

If a population can be divided into groups of similar elements called strata, then a stratified random sample is appropriate (*see* **Stratification**). Random samples are selected from each stratum, which insures that the diversity of the population is represented in the sample and an estimate is also obtained for each stratum.

If no frame exists, a cluster sample is possible. For example, to estimate the number of deer on 100 acres, the 100 acres can be divided into one-acre plots on a map. Then a random sample of the one-acre plots can be selected and the number of deer counted for each selected plot. Suppose five acres are selected and a total of 10 deer are found. Then the estimate for the 100 acres would be 2 deer per acre and 200 for the 100 acres.

Sample Size

An approximate estimate of the sample size can be determined from the following two equations (see [1]). If you are estimating an average value for a quantitative variable, (1) can be used.

$$n = \left(\frac{2\sigma}{B}\right)^2, \tag{1}$$

where n is the sample size, σ is the population standard deviation, and B is the bound on the error. The bound on the error is the maximum differences between the true value and the stated value with probability. 9544. An estimate of σ may be obtained in three ways: (a) estimate σ from historical studies, (b) estimate σ from (range of values /6) $\cong \sigma$, and

(c) obtain a pilot sample of 30 or more elements and estimate σ by calculating $\hat{\sigma}$ (the sample standard deviation), (see 2).

For example, if you wanted to estimate miles per gallon, mpg, for a large population of automobiles within 2 mpg, then B would equal 2. A pilot sample can be taken and $\hat{\sigma}$ is used to estimate σ where $\hat{\sigma}$ is found from equation

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}, \tag{2}$$

where X_i is the sample value and $\bar{X}$ is the sample mean. For the above example, suppose $\sigma \cong 4$. Then using (1)

$$n = \left(\frac{2(4)}{2}\right)^2 = 16. \tag{3}$$

A sample of size 16 would produce a sample mean mpg that differs from the population mean mpg by 2 mpg or less.

If a population proportion (percentage) is to be estimated, the sample size is found from (4).

$$n = \left(\frac{2}{B}\right)^2 p(1-p), \tag{4}$$

where B is the bound on the error of the estimate, n is the sample size, and, p is the population proportion. The population proportion can be estimated three ways: (a) use historical values of p, (b) obtain a pilot sample of 30 or more elements and estimate p, and (c) use $p = .5$ because this produces the widest interval.

An example of determining the sample size necessary to estimate the population proportion is given below when no information exists about the value of p. Suppose an estimate of the proportion of voters that favor a candidate is to be estimated within ±3% points. If there is no information from previous research, we select $p = .5$. Then $B = .03$ and n is determined from (4)

$$n = \left(\frac{2}{.03}\right)^2 .5.5 = 1111.11 \text{ or } 1112. \tag{5}$$

Consider another example. If a low percentage is to be estimated like the proportion of defective light bulbs that is known to be 5% or less, then use $p = .05$ to estimate the proportion of defectives. If we want to estimate the percentage within ±1%, we use $B = .01$. Then from (4)

$$n = \left(\frac{2}{.01}\right)^2 .05(.95) = 1900. \tag{6}$$

In summary, survey sampling procedures allow estimates of characteristics of populations (parameters) from characteristics of a sample (statistics). The procedure of survey sampling saves time and cost, and the accuracy of estimates is relatively high.

References

[1] Keller, G. & Warrack, B. (2003). *Statistics for Management and Economics*, 6th edition, Thomson Learning Academic Resource Center.

[2] Scheaffer, R.L., Mendenhall, I.I.I.W. & Ott, L.O. (1996). *Elementary Survey Sampling*, 5th edition, Duxbury Press.

Kris K. Moore

Survival Analysis

Sabine Landau

Volume 4, pp. 1980–1987

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Survival Analysis

In many studies the main outcome is the time from a well-defined *time origin* to the occurrence of a particular event or *end-point*. If the end-point is the death of a patient the resulting data are literally *survival times*. However, other end-points are possible, for example, the time to relief of symptoms or to recurrence of a particular condition, or simply to the completion of an experimental task. Such observations are often referred to as *time to event data* although the generic term *survival data* is commonly used to indicate any time to event data.

Standard statistical methodology is not usually appropriate for survival data, for two main reasons:

1. The distribution of survival time in general is likely to be *positively skewed* and so assuming normality for an analysis (as done for example, by a t Test or a regression) is probably not reasonable.
2. More critical than doubts about normality, however, is the presence of *censored* observations, where the survival time of an individual is referred to as censored when the end-point of interest has not yet been reached (more precisely *right-censored*). For true survival times this might be because the data from a study are analyzed at a time point when some participants are still alive. Another reason for censored event times is that an individual might have been *lost to follow-up* for reasons unrelated to the event of interest, for example, due to moving to a location which cannot be traced. When censoring occurs all that is known is that the actual, but unknown, survival time is larger than the censored survival time.

Specialized statistical techniques that have been developed to analyze such censored and possibly skewed outcomes are known as *survival analysis*. An important assumption made in standard survival analysis is that the censoring is *noninformative*, that is that the actual survival time of an individual is independent of any mechanism that causes that individual's survival time to be censored. For simplicity this description also concentrates on techniques for continuous survival times - for the analysis of discrete survival times see [3, 6].

A Survival Data Example

As an example, consider the data in Table 1 that contain the times heroin addicts remained in a clinic for methadone maintenance treatment [2]. The study ($n = 238$) recorded the duration spent in the clinic, and whether the recorded time corresponds to the time the patient leaves the programme or the end of the observation period (for reasons other than the patient deciding that the treatment is 'complete'). In this study the time of origin is the date on which the addicts first attended the methadone maintenance clinic and the end-point is methadone treatment cessation (whether by patient's or doctor's choice). The durations of patients who were lost during the follow-up process are regarded as right-censored. The 'status' variable in Table 1 takes the value unity if methadone treatment was stopped, and zero if the patient was lost to follow-up. In addition a number of prognostic variables were recorded;

1. one of two clinics,
2. presence of a prison record
3. and maximum methadone dose prescribed.

The main aim of the study was to identify predictors of the length of the methadone maintenance period.

Survival Analysis Concepts

To describe survival two functions of time are of central interest. The *survival function* $S(t)$ is defined as the probability that an individual's survival time, T, is greater than or equal to time t, that is,

$$S(t) = \text{Prob}(T \geq t) \qquad (1)$$

The graph of $S(t)$ against t is known as the *survival curve*. The survival curve can be thought of as a particular way of displaying the frequency distribution of the event times, rather than by say a histogram.

In the analysis of survival data it is often of some interest to assess which periods have the highest and which the lowest chance of death (or whatever the event of interest happens to be), amongst those people at risk at the time. The appropriate quantity for such risks is the *hazard function*, $h(t)$, defined as the (scaled) probability that an individual experiences the event in a small time interval δt, given that

Table 1 Durations of heroin addicts remaining in a methadone treatment programme (only five patients from each clinic are shown)

Patient ID	Clinic	Status	Time (days)	Prison record (1 = 'present', 0 = 'absent')	Maximum methadone (mg/day)
1	1	1	428	0	50
2	1	1	275	1	55
3	1	1	262	0	55
4	1	1	183	0	30
5	1	1	259	1	65
...					
103	2	1	708	1	60
104	2	0	713	0	50
105	2	0	146	0	50
106	2	1	450	0	55
109	2	0	555	0	80
...					

the individual has not experienced the event up to the beginning of the interval. The hazard function therefore represents the *instantaneous event rate* for an individual at risk at time t. It is a measure of how likely an individual is to experience an event as a function of the age of the individual. The hazard function may remain constant, increase or decrease with time, or take some more complex form. The hazard function of death in human beings, for example, has a 'bath tub' shape. It is relatively high immediately after birth, declines rapidly in the early years and then remains pretty much constant until beginning to rise during late middle age.

In formal, mathematical terms, the hazard function is defined as the following limiting value

$$h(t) = \lim_{\delta t \to 0} \left[\frac{\text{Prob}(t \le T < t + \delta t | T \ge t)}{\delta t} \right] \quad (2)$$

The conditional probability is expressed as a probability per unit time and therefore converted into a rate by dividing by the size of the time interval, δt.

A further function that features widely in survival analysis is the *integrated* or *cumulative hazard function*, $H(t)$, defined as

$$H(t) = \int_0^t h(u)\,\mathrm{d}u \quad (3)$$

The hazard function is mathematically related to the survivor functions. Hence, once a hazard function is specified so is the survivor function and *vice versa.*

Nonparametric Procedures

An initial step in the analysis of a set of survival data is the numerical or graphical description of the survival times. However, owing to the censoring this is not readily achieved using conventional descriptive methods such as boxplots and summary statistics. Instead survival data are conveniently summarized through estimates of the survivor or hazard function obtained from the sample.

When there are no censored observations in the sample of survival times, the survival function can be estimated by the *empirical survivor function*

$$\hat{S}(t) = \frac{\text{Number of individuals with event times} \ge t}{\text{Number of individuals in the data set}} \quad (4)$$

Since every subject is 'alive' at the beginning of the study and no-one is observed to survive longer than the largest of the observed survival times then

$$\hat{S}(0) = 1 \text{ and } \hat{S}(t) = 0 \text{ for } \quad t > t_{\max} \quad (5)$$

Furthermore the estimated survivor function is assumed constant between two adjacent death times so that a plot of $\hat{S}(t)$ against t is a step function that decreases immediately after each 'death'. However, this simple method cannot been used when there are censored observations since the method does not allow for information provided by an individual whose survival time is censored before time t to be used in the computing of the estimate at t.

The most commonly used method for estimating the survival function for survival data containing censored observations is the *Kaplan–Meier* or *product-limit-estimator* [8]. The essence of this approach is the use of a product of a series of conditional probabilities. This involves ordering the $r$sample event times from the smallest to the largest such that

$$t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(r)} \tag{6}$$

Then the survival curve is estimated from the formula

$$\hat{S}(t) = \prod_{j|t_{(j)} \leq t} \left(1 - \frac{d_j}{r_j}\right) \tag{7}$$

where r_j is the number of individuals at risk at $t_{(j)}$ and d_j is the number experiencing the event of interest at $t_{(j)}$. (Individuals censored at $t_{(j)}$ are included in r_j.) For example, the estimated survivor function at the second event time $t_{(2)}$ is equal to the estimated probability of not experiencing the event at time $t_{(1)}$ times the estimated probability, given that the individual is still at risk at time $t_{(2)}$, of not experiencing it at time $t_{(2)}$.

The Kaplan-Meier estimators of the survivor curves for the two methadone clinics are displayed in Figure 1. The survivor curves are step functions with decrease at the time points when patients ceased methadone treatment. The censored observations in the data are indicated by the 'cross' marks on the curves.

The variance of the Kaplan-Meier estimator of the survival curve can itself be estimated from *Greenwood's formula* and once the standard error has been determined point-wise symmetric confidence intervals can be found by assuming a normal distribution on the original scale or asymmetric intervals can be constructed after transforming $\hat{S}(t)$ to a value on the continuous scale, for details see [3, 6].

A *Kaplan-Meier type estimator* of the hazard function is given by the proportion of individuals experiencing an event in an interval per unit time, given that they are at risk at the beginning of the interval, that is

$$\tilde{h}(t) = \frac{d_j}{r_j\left(t_{(j+1)} - t_{(j)}\right)} \tag{8}$$

Integration leads to the *Nelson-Aalen* or *Altshuler's* estimator of the cumulative hazard function, $\tilde{H}(t)$, and employing the theoretical relationship between the survivor function and the cumulative hazard function to the Nelson-Aalen estimator of the survivor function. Finally, it needs to be noted that relevant functions can be estimated using the so-called *life-table* or *Actuarial estimator*. This approach is, however, sensitive to the choice of intervals used in its construction and therefore not generally recommended for continuous survival data (readers are referred to [3, 6]).

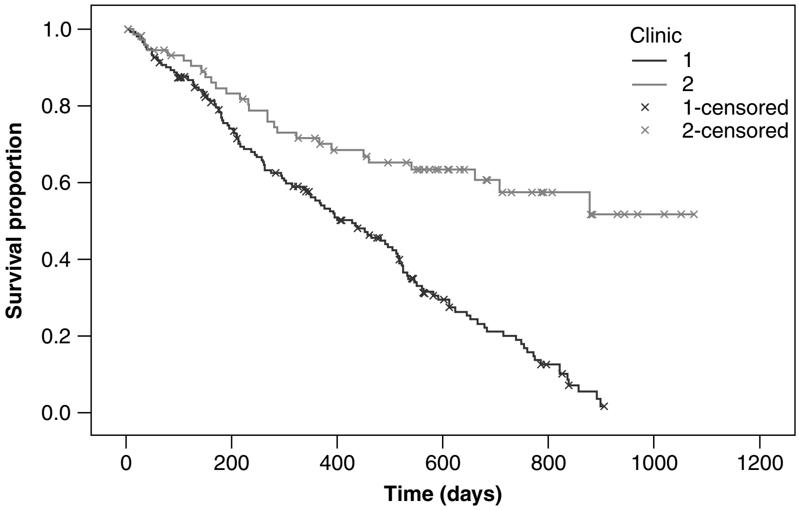


Figure 1 Kaplan-Meier survivor curves for the heroin addicts data

Standard errors and confidence intervals can be constructed for all three functions although the estimated hazard function is generally considered 'too noisy' for practical use. The Nelson-Aalen estimator is typically used to describe the cumulative hazard function while the Kaplan-Meier estimator is used for the survival function.

Since the distribution of survival times tends to be positively skewed the median is the preferred summary measure of location. The *median event time* is the time beyond which 50% of the individuals in the population under study are expected to 'survive', and, once the survivor function has been estimated by $\hat{S}(t)$, can be estimated by the smallest observed survival time, t_{50}, for which the value of the estimated survivor function is less than 0.5. The estimated median survival time can be read from the survival curve by finding the smallest value on the x-axis for which the survival proportions reaches less than 0.5. Figure 1 shows that the median methadone treatment duration in clinic 1 group can be estimated as 428 days while an estimate is not available for clinic 2 since more than 50% of the patients continued treatment throughout the study period. A similar procedure can be used to estimate other percentiles of the distribution of the survival times and approximate confidence intervals can be found once the variance of the estimated percentile has been derived from the variance of the estimator of the survivor function.

In addition to comparing survivor functions graphically a more formal statistical test for a group difference is often required. In the absence of censoring a nonparametric test, like the Mann-Whitney test could be used (*see* **Distribution-free Inference, an Overview**). In the presence of censoring the *log-rank* or *Mantel-Haenszel test* [9] is the most commonly used nonparametric test. It tests the null hypothesis that the population survival functions $S_1(t), S_2(t), \ldots, S_k(t)$ are the same in k groups.

Briefly, the test is based on computing the expected number of events for each observed event time in the data set, assuming that the chances of the event, given that subjects are at risk, are the same in the groups. The total number of expected events is then computed for each group by adding the expected number of events for each event time. The test finally compares the observed number of events in each group with the expected number of events using a chi-squared test with $k - 1$ degrees of freedom, *see* [3, 6].

The log-rank test statistic, X^2, weights contributions from all failure times equally. Several alternative test statistics have been proposed that give differential weights to the failure times. For example, the *generalized Wilcoxon test* (or *Breslow test*) uses weights equal to the number at risk. For the heroin addicts data in Table 1 the log-rank test ($X^2 = 27.9$ on 1 degree of freedom, $p < 0.0001$) detects a significant clinic difference in favor of longer treatment durations in clinic 2. The Wilcoxon test puts relatively more weight on differences between the survival curves at earlier times but also reaches significance ($X^2 = 11.6$ on 1 degree of freedom, $p = 0.0007$).

Modeling Survival Times

Modeling survival times is useful especially when there are several explanatory variables of interest. For example the methadone treatment durations of the heroin addicts might be affected by the prognostic variables maximum methadone dose and prison record as well as the clinic attended. The main approaches used for Modeling the effects of covariates on survival can be divided roughly into two classes – *models for the hazard function* and *models for the survival times themselves*. In essence these models act as analogies of **multiple linear regression** for survival times containing censored observations, for which regression itself is clearly not suitable.

Proportional Hazards Models

The main technique is due to Cox [4] and known as the *proportional hazards model* or, more simply, *Cox's regression*. The approach stipulates a model for the hazard function. Central to the procedure is the assumption that the hazard functions for two individuals at any point in time are proportional, the so-called *proportional hazards assumption*. In other words, if an individual has a risk of the event at some initial time point that is twice as high as another individual, then at all later times the risk of the event remains twice as high. Cox's model is made up of an unspecified *baseline hazard function*, $h_0(t)$, which is then multiplied by a suitable function of an individual's explanatory variable values, to give the individual's hazard function. Formally, for a set of p explanatory variables, $x_1, x_2, \ldots, x_p$, the model is

$$h(t) = h_0(t) \exp\left(\sum_{i=1}^{p} \beta_i x_i\right) \tag{9}$$

where the terms $\beta_1, \ldots, \beta_p$ are the parameters of the model which have to be estimated from sample data. Under this model the *hazard* or *incidence rate ratio*, h_{12}, for two individuals, with covariate values $x_{11}, x_{12}, \ldots, x_{1p}$ and $x_{21}, x_{22}, \ldots, x_{2p}$

$$h_{12} = \frac{h_1(t)}{h_2(t)} = \exp\left[\sum_{i=1}^{p} \beta_i (x_{1i} - x_{2i})\right] \tag{10}$$

does not depend on t. The interpretation of the parameter β_i is that $\exp(\beta_i)$ gives the incidence rate change associated with an increase of one unit in x_i, all other explanatory variables remaining constant. Specifically, in the simple case of comparing hazards between two groups, $\exp(\beta)$, measures the hazard ratio between the two groups. The effect of the covariates is assumed multiplicative.

Cox's regression is considered a *semiparametric* procedure because the baseline hazard function, $h_0(t)$, and by implication the probability distribution of the survival times does not have to be specified. The baseline hazard is left unspecified; a different parameter is essentially included for each unique survival time. These parameters can be thought of as nuisance parameters whose purpose is merely to control the parameters of interest for any changes in the hazard over time. Cox's regression model can also be extended to allow the baseline hazard function to vary with the levels of a stratification variable. Such a *stratified proportional hazards model* is useful in situations where the stratifier is thought to affect the hazard function but the effect itself is not of primary interest.

A Cox regression can be used to model the methadone treatment times from Table 1. The model uses prison record and methadone dose as explanatory variables whereas the variable clinic, whose effect was not of interest, merely needed to be taken account of and did not fulfill the proportional hazards assumption, was used as a stratifier. The estimated regression coefficients are shown in Table 2. The coefficient of the prison record indicator variable is 0.389 with a standard error of 0.17. This translates into a hazard ratio of $\exp(0.389) = 1.475$ with a 95% confidence interval ranging from 1.059 to 2.054. In other words a prison record is estimated to increase the hazard of immediate treatment cessation by 47.5%. Similarly the hazard of treatment cessation was estimated to be reduced by 3.5% for every extra mg/day of methadone prescribed.

Statistical software packages typically report three different tests for testing regression coefficients, the *likelihood ratio* (*LR*) *test* (*see* **Maximum Likelihood Estimation**), the *score test* (which for Cox's proportional hazards model is equivalent to the log-rank test) and the *Wald test*. The test statistic of each of the tests can be compared with a chi-squared distribution to derive a P value. The three tests are asymptotically equivalent but differ in finite samples. The likelihood ratio test is generally considered the most reliable and the Wald test the least. Here presence of a prison record tests statistically significant after adjusting for clinic and methadone dose (LR test: $X^2 = 5.2$ on 1 degree of freedom, $p = 0.022$) and so does methadone dose after adjusting for clinic and prison record (LR test: $X^2 = 30.0$ on 1 degree of freedom, $p < 0.0001$).

Cox's model does not require specification of the probability distribution of the survival times. The hazard function is not restricted to a specific form and as a result the semiparametric model has flexibility and is widely used. However, if the assumption of

Table 2 Parameter estimates from Cox regression of treatment duration on maximum methadone dose prescribed and presence of a prison record stratified by clinic

	Effect estimate			95% CI for $\exp(\beta)$	
Predictor variable	Regression coefficient ($\hat{\beta}$)	Standard error $\left(\sqrt{var(\hat{\beta})}\right)$	Hazard ratio ($\exp(\hat{\beta})$)	Lower limit	Upper limit
Prison record	0.389	0.17	1.475	1.059	2.054
Maximum methadone dose	−0.035	0.006	0.965	0.953	0.978

a particular probability distribution for the data is valid, inferences based on such an assumption are more precise. For example estimates of hazard ratios or median survival times will have smaller standard errors. A *fully parametric proportional hazards model* makes the same assumptions as Cox's regression but in addition also assumes that the baseline hazard function, $h_0(t)$, can be parameterized according to a specific model for the distribution of the survival times. Survival time distributions that can be used for this purpose, that is that have the *proportional hazards property*, are principally the *Exponential, Weibull,* and *Gompertz* distributions (*see* **Catalogue of Probability Density Functions**). Different distributions imply different shapes of the hazard function, and in practice the distribution that best describes the functional form of the observed hazard function is chosen – for details *see* [3, 6].

Models for Direct Effects on Survival Times

A family of fully parametric models that assume direct multiplicative effects of covariates on survival times and hence do not rely on proportional hazards are *accelerated failure time models.* A wider range of survival time distributions possesses the *accelerated failure time property*, principally the *Exponential, Weibull, log-logistic, generalized gamma,* or *lognormal* distributions. In addition this family of parametric models includes distributions (e.g., the log-logistic distribution) that model unimodal hazard functions while all distributions suitable for the proportional hazards model imply hazard functions that increase or decrease monotonically. The latter property might be limiting, for example, for Modeling the hazard of dying after a complicated operation that peaks in the postoperative period.

The general accelerated failure time model for the effects of p explanatory variables, $x_1, x_2, \ldots, x_p$, can be represented as a **log-linear model** for survival time, T, namely,

$$\ln(T) = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_i + \text{error} \qquad (11)$$

where $\alpha_1, \ldots, \alpha_p$ are the unknown coefficients of the explanatory variables and α_0 an intercept parameter. The parameter α_i reflects the effect that the ith covariate has on log-survival time with positive values indicating that the survival time increases with increasing values of the covariate and *vice versa.* In terms of the original time scale the model implies that the explanatory variables measured on an individual act multiplicatively, and so affect the speed of progression to the event of interest.

The interpretation of the parameter α_i then is that $\exp(\alpha_i)$ gives the factor by which any survival time percentile (e.g., the median survival time) changes per unit increase in x_i, all other explanatory variables remaining constant. Expressed differently, the probability, that an individual with covariate value $x_i + 1$ survives beyond t, is equal to the probability, that an individual with value x_i survives beyond $\exp(-\alpha_i)t$. Hence $\exp(-\alpha_i)$ determines the change in the speed with which individuals proceed along the time scale and the coefficient is known as the *acceleration factor* of the ith covariate.

Software packages typically use the log-linear formulation. The regression coefficients from fitting a log-logistic accelerated failure time model to the methadone treatment durations using prison record and methadone dose as explanatory variables and clinic as a stratifier are shown in Table 3. The negative regression coefficient for prison suggests that the treatment durations tend to be shorter for those with a prison record. The positive regression coefficient for dose suggests that treatment durations tend to be prolonged for those on larger methadone doses.

Table 3 Parameter estimates from log-logistic accelerated failure time model of treatment duration on maximum methadone dose prescribed and presence of a prison record stratified by clinic

	Effect estimate			95% CI for $\exp(\alpha)$	
Predictor variable	Regression coefficient ($\hat{\alpha}$)	Standard error ($\sqrt{var(\hat{\alpha})}$)	Acceleration factor ($\exp(-\hat{\alpha})$)	Lower limit	Upper limit
Prison record	−0.328	0.140	1.388	1.054	1.827
Maximum methadone dose	0.0315	0.0055	0.969	0.959	0.979

The estimated acceleration factor for an individual with a prison record compared with one without such a record is $\exp(0.328) = 1.388$, that is, a prison record is estimated to accelerate the progression to treatment cessation by a factor of about 1.4. Both explanatory variables, prison record (LR test: $X^2 = 5.4$ on 1 degree of freedom, $p = 0.021$) and maximum methadone dose prescribed (LR test: $X^2 = 31.9$ on 1 degree of freedom, $p < 0.0001$) are found to have statistically significant effects on treatment duration according to the log-logistic accelerated failure time model.

Summary

Survival analysis is a powerful tool for analyzing time to event data. The classical techniques Kaplan-Meier estimation, proportional hazards, and accelerated failure time Modeling are implemented in most general purpose statistical packages with the S-PLUS and R packages having particularly extensive facilities for fitting and checking nonstandard Cox models, *see* [10]. The area is complex and one of active current research. For additional topics such as other forms of censoring and *truncation* (delayed entry), *recurrent events*, models for *competing risks, multistate models* for different transition rates, and *frailty models* to include random effects the reader is referred to [1, 5, 7, 10].

References

[1] Andersen, P.K., ed. (2002). *Multistate Models, Statistical Methods in Medical Research 11*, Arnold Publishing, London.

[2] Caplethorn, J. (1991). Methadone dose and retention of patients in maintenance treatment, *Medical Journal of Australia* **154**, 195–199.

[3] Collett, D. (2003). *Modelling Survival Data in Medical Research*, 2nd Edition, Chapman & Hall/CRC, London.

[4] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society: Series B* **74**, 187–220.

[5] Crowder, K.J. (2001). *Classical Competing Risks*, Chapman & Hall/CRC, Boca Raton.

[6] Hosmer, D.W. & Lemeshow, S. (1999). *Applied Survival Analysis*, Wiley, New York.

[7] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer, New York.

[8] Kaplan, E.L. & Meier, P. (1958). Nonparametric estimation for incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

[9] Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariance test procedures (with discussion), *Journal of the Royal Statistical Society: Series A* **135**, 185–206.

[10] Therneau, T.M. & Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*, Springer, New York.

SABINE LANDAU

Symmetry: Distribution Free Tests for

CLIFFORD E. LUNNEBORG

Volume 4, pp. 1987–1988

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Symmetry: Distribution Free Tests for

Nonparametric tests for the median of a distribution and the related estimation of confidence intervals for that parameter assume that the distribution sampled is symmetric about the median. To avoid obtaining misleading results, a preliminary test for distribution symmetry is advisable.

Tests for Symmetry of Distribution

The four tests of symmetry selected for mention here have been shown to have reasonable power to detect asymmetric distributions and are either widely used or easy to apply.

The Gupta test proposed in 1967 [3] and [4] is based on the set of pair-wise comparisons of the sample data. For each pair of data points, the statistic δ_{ij} is assigned a value of 1 if $(x_i + x_j)$ is greater than twice the sample median or 0 otherwise. The sum of these δ_{ij}s will be large if the underlying distribution is skewed to the right, small if it is skewed to the left, and intermediate in value if the distribution is symmetric. After centering and standardization, the test statistic is asymptotically distributed as the standard normal random variable. The approach is explained in detail in [4], and has been built into the SC statistical package (`www.mole-software.demon.co.uk`) as the *Gupta* procedure.

The Randles et al. *test*, published in 1980 [5] and [6], is based on the set of triplets of data points. For each unique set of three data points, the statistic ζ_{ijk} is assigned the value 1/3 if the mean of the three data points, (x_i, x_j, x_k), is greater than their median, 0 if the mean and median are equal, and $-1/3$ if the median is larger than the mean. The sum of these ζ_{ijk}s will be large and positive if the underlying distribution is skewed to the right, large and negative if the distribution is skewed to the left, and small if the distribution is symmetric. After standardization, this sum enjoys, at least asymptotically, a normal sampling distribution under the null, symmetric hypothesis. Details are given in [5].

The Boos test, described in 1982 [1], is based on the set of absolute differences between the $n(n+1)/2$ **Walsh averages** and their median, the one-sample **Hodges–Lehmann estimator**. When summed, large values are suggestive of an underlying asymmetric distribution. Critical asymptotic values for a scaling of the sum of absolute deviations are given in [1].

The Cabilio & Masaro test, presented in 1996 [2], features simplicity of computation. The test statistic,

$$S_K = \frac{[\sqrt{n}(Mn - Mdn)]}{Sd}, \tag{1}$$

requires for its computation only four sample quantities – size, mean, median, and standard deviation. The test's authors recommend comparing the value of the test statistic, S_K, against the critical values in

Table 1 Empirical quantiles of the distribution of S_K (normal samples)

n	Q_{90}	Q_{95}	$Q_{97.5}$	n	Q_{90}	Q_{95}	$Q_{97.5}$
5	0.88	1.07	1.21	6	0.71	0.88	1.02
7	0.91	1.13	1.30	8	0.77	0.96	1.13
9	0.92	1.15	1.35	10	0.81	1.01	1.19
11	0.93	1.17	1.37	12	0.83	1.05	1.24
13	0.93	1.18	1.39	14	0.85	1.08	1.27
15	0.94	1.19	1.41	16	0.86	1.10	1.30
17	0.94	1.19	1.41	18	0.87	1.11	1.32
19	0.94	1.20	1.42	20	0.88	1.12	1.33
21	0.95	1.20	1.43	22	0.89	1.13	1.35
23	0.95	1.21	1.43	24	0.89	1.14	1.36
25	0.95	1.21	1.44	26	0.90	1.15	1.37
27	0.95	1.21	1.44	28	0.90	1.15	1.37
29	0.95	1.21	1.44	30	0.91	1.16	1.38
∞	0.97	1.24	1.48				

Table 1, reproduced here from [2] with the permission of the authors.

These critical values were obtained by calibrating the test against samples from a particular symmetric distribution, the normal. However, the nominal significance level of the test appears to hold for other symmetric distributions, with the exception of the Cauchy and uniform distributions (*see* **Catalogue of Probability Density Functions**) [2].

Comments

The power of the four tests of distribution symmetry have been evaluated and compared, [2] and [6]. Briefly, the Randles et al. test dominates the Gupta test [6], while the Boos and Cabilio–Masaro tests appear to be superior to Randles et al. [2]. Although the Boos test may have a slight power advantage over the Cabilio–Masaro procedure, the ease of application of the latter suggests that the assumption of symmetry might more frequently be checked than it has been in the past. It should be noted, however, that tests of symmetry appear to have fairly low power to detect asymmetric distributions when the sample size is smaller than 20 [6].

References

[1] Boos, D.D. (1982). A test for asymmetry associated with the Hodges-Lehmann estimator, *Journal of the American Statistical Association* **77**, 647–651.

[2] Cabilio, P. & Masaro, J. (1996). A simple test of symmetry about an unknown median, *The Canadian Journal of Statistics* **24**, 349–361.

[3] Gupta, M.K. (1967). An asymptotically nonparametric test of symmetry, *Annals of Mathematical Statistics* **38**, 849–866.

[4] Hollander, M. & Wolfe, D.A. (1973). *Nonparametric Statistical Methods*, Wiley, New York.

[5] Hollander, M. & Wolfe, D.A. (1999). *Nonparametric Statistical Methods*, 2nd Edition, Wiley, New York.

[6] Randles, H., Fligner, M.A., Pollicello, G. & Wolfe, D.A. (1980). An asymptotically distribution-free test for symmetry versus asymmetry, *Journal of the American Statistical Association* **75**, 168–172.

CLIFFORD E. LUNNEBORG

Symmetry Plot

Sandy Lovie

Volume 4, pp. 1989–1990

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Symmetry Plot

This plot does exactly what it says on the tin, that is, it provides a graphical test of whether a sample is symmetrically distributed about a measure of location; in this case, the median. Having such information about a sample is useful in that just about all tests of significance assume that the parent population from which the sample came is at least symmetrical about some location parameter and, in effect, that the sample should not markedly violate this condition either. A further linked role for the plot is its use in evaluating transformations to *achieve* symmetry, particularly following schemes like the *ladder of powers* advocated for **Exploratory Data Analysis** (see, for example [2]).

The plot itself is built up by first ordering the data and calculating a median, if necessary, by interpolation for even numbered samples. Secondly, each reading in the sample is subtracted from the median, thus reexpressing all the sample values as (signed and ordered) distances from the median. Then, these distances are expressed as unsigned values, whilst still keeping separate those ordered distances that lie above the median from those below it. Next, as with the **empirical quantile-quantile (EQQ) plot**, the ordered values above and below the median are paired in increasing order of size, and then plotted on a conventional **scatterplot**. Here the zero/zero origin represents the median itself, with the ascending dots representing the ordered pairs, where the lowest one represents the two smallest distances above and below the median, the next higher dot the next smallest pair, and so on. Also, as with the EQQ plot, a 45° comparison line is placed on the plot to represent perfect symmetry about the median (note that the *x* and *y* axes of the scatterplot are equal in all respects, hence the angle of the line). All judgements as to the symmetry of the sample are therefore made relative to this line. The statistics package Minitab adds a simple histogram of the data to its version of the plot, thus aiding in its interpretation (*see* **Software for Statistical Analyses**).

The three illustrative plots below are from Minitab and use the *Pulse* data set. Figure 1 is a symmetry plot of the raw data for 35 human pulse readings after exercise (running on the spot for one minute). Here, the histogram shows a marked skew to the left, which shows up on the full plot as the data, both lying below the comparison line and increasingly divergent from it, as one moves to the right. However, if the data had been skewed to the right, then the plotted data would have appeared above the comparison line.

The next two plots draw on transforms from the ladder of powers to improve the symmetry of the data. The first applies a $\log_{10}$ transform to the data, while the second uses a reciprocal $(1/x)$ transform (*see* **Transformation**). Notice that the log transform in Figure 2 improves the symmetry of the histogram a little, which shows up in the symmetry plot as data, which are now somewhat closer to the comparison line and less divergent from it than in the raw plot.

However, this is improved on even further in Figure 3, where the histogram is even more symmetric, and the data of the symmetry plot is much closer and much less divergent than in the log transformed plot.

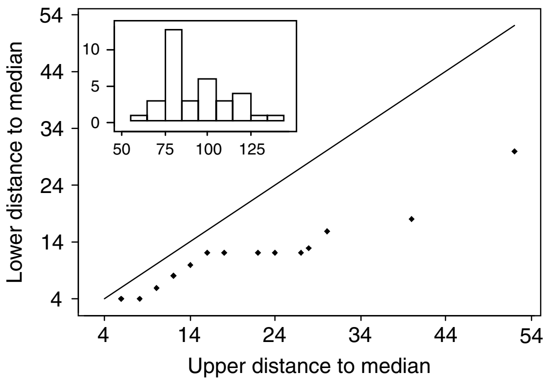


Figure 1 Symmetry plot of raw 'pulse after exercise' data

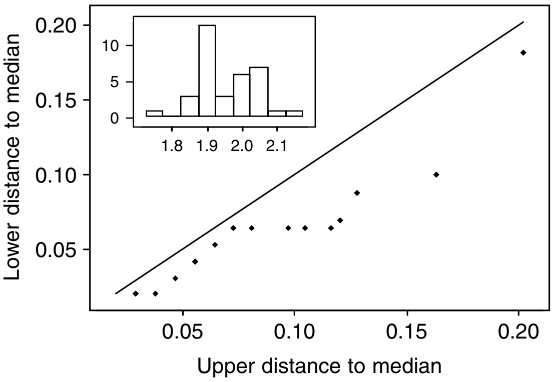


Figure 2 Symmetry plot of the log transformed 'pulse after exercise' data

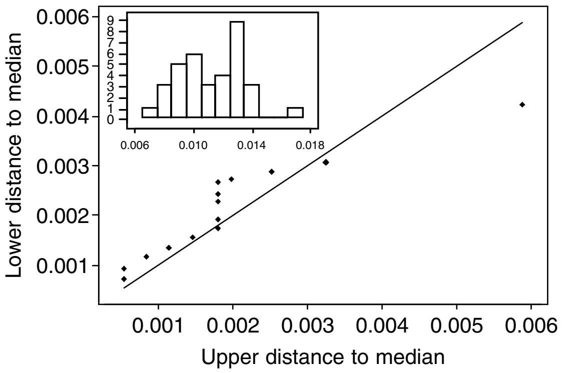


Figure 3 Symmetry plot of the reciprocal transformed 'pulse after exercise' data

Interestingly, a somewhat more complex transform lying between the two chosen here from the ladder of powers, the *reciprocal/square root* $(1/\sqrt{x})$, generates a symmetry plot (not included here), which reproduces many of the aspects of Figure 3 for the bulk of the data on the left-hand side of the plot, and also draws the somewhat anomalous data point on the far right, much closer to the comparison line. Although analyses of the resulting data might be more difficult to interpret, this transform is probably the one to choose if you want your (transformed) results to conform to the symmetry assumption behind all those tests of significance!

More information on symmetry plots can be found in [1], pages 29 to 32.

References

[1] Chambers, J.M., Cleveland, W.S., Kleiner, B. & Tukey, P.A. (1983). *Graphical Methods for Data Analysis*, Duxbury, Boston.

[2] Tukey, J.W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading.

SANDY LOVIE

Tau-Equivalent and Congeneric Measurements

Jos M.F. Ten Berge

Volume 4, pp. 1991–1994

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Tau-Equivalent and Congeneric Measurements

It is a well-known fact that the reliability of a test, defined as the ratio of true score to observed score variance, cannot generally be determined from a single test administration, but requires the use of a parallel test. More often than not, parallel tests are not available. In such cases, two approaches are popular to obtain indirect information on the reliability of the test: either lower bounds to reliability can be used, or one may resort to hypotheses about the nature of the test parts.

Evaluating lower bounds to the reliability, such as Guttman's λ_3 [6], better known as coefficient alpha [4] has gained wide popularity. A lower bound that is nearly always better than alpha is Guttman's λ_4. It is the highest alpha that can be obtained by splitting up the items in two parts (not necessarily of equal numbers) and treating those two parts as novel 'items'. Jackson and Agunwamba [8] proposed the greatest lower bound (glb) to reliability. It exceeds all conceivable lower bounds by using the available information implied by the observed covariance matrix exhaustively. A computational method for the glb has been proposed by Bentler & Woodward [3], also see [19]. Computation of the glb has been implemented in EQS 6.

When lower bounds are high enough, the reliability has been shown adequate by implication. However, when lower bounds are low, they are of limited value. Also, some lower bounds to reliability involve a considerable degree of sampling bias. To avoid these problems, it is tempting to look to alternative approaches, by introducing hypotheses on the nature of the test parts, from which the reliability can be determined at once. Two of such hypotheses are well-known in **classical test theory**.

Tau Equivalence

The first hypothesis is that of (essentially) tau-equivalent tests. Test parts $X_1, \ldots, X_k$ are essentially tau-equivalent when for $i, j = 1, \ldots, k$,

$$T_j = T_i + a_{ij}. \tag{1}$$

This implies that the true scores of the test parts are equal up to an additive constant. When the additive constants are zero, the test parts are said to be tau-equivalent. Novick and Lewis [14] have shown that coefficient alpha is the reliability (instead of merely a lower bound to it) if and only if the test parts are essentially tau-equivalent.

Unfortunately, the condition for essential tau-equivalence to hold in practice is prohibitive: All covariances between test parts must be equal. This will only be observed when $k = 2$ or with contrived data. Moreover, the condition of equal covariances is necessary, but not sufficient for essential tau-equivalence. For instance, let Y_1, Y_2 and Y_3 be three uncorrelated variables with zero means and unit variances, and let the test parts be $X_1 = Y_2 + Y_3$, $X_2 = Y_1 + Y_3$, and $X_3 = Y_1 + Y_2$. Then, the covariance between any two test parts is 1, but the test parts are far from being essentially tau-equivalent, which shows that having equal covariances is necessary but not sufficient for essential tau-equivalence. Because the necessary condition will never be satisfied in practice, coefficient alpha is best thought of as a lower bound (underestimate) to the reliability of a test.

Congeneric Tests

A weaker, and far more popular hypothesis, is that of a congeneric test, consisting of k test parts satisfying

$$T_j = c_{ij}T_i + a_{ij}, \tag{2}$$

which means that the test parts have perfectly correlated true scores. Equivalently, the test parts are assumed to fit a one-factor model. Essential tau equivalence is more restricted because it requires that the weights c_{ij} in (2) are unity. For the case $k = 3$, Kristof has derived a closed-form expression for the reliability of the test, based on this hypothesis. It is always at least as high as alpha, and typically better [10]. Specifically, there are cases where it coincides with or even exceeds the greatest lower bound to reliability [20].

For $k > 3$, generalized versions of Kristof's coefficients have been proposed. For instance, Gilmer and Feldt [5] offered coefficients that could be evaluated without having access to powerful computers. They were fully aware that these coefficients would be supplanted by common factor analysis (*see* **History of Factor Analysis: A Psychological Perspective**) based coefficients by the time large computers would be generally available. Nowadays, even the smallest of personal computers can evaluate the reliability of a test in the framework of common factor analysis, assuming that the one-factor hypothesis is true. For instance, McDonald [13], also see Jöreskog [9] for a similar method, proposed estimating the loadings on a single factor and evaluating the reliability as the ratio of the squared sum of loadings to the test variance. When $k = 3$, this yields Kristof's coefficient. Coefficients like Kristof's and McDonald's have been considered useful alternatives to lower bounds like glb, because they aim to estimate, rather than underestimate, reliability, and they lack any reputation of sampling bias. However, much like the hypothesis of essential tau-equivalence, the one-factor hypothesis is problematic.

The Hypothesis of Congeneric Tests is Untenable for $k > 3$, and Undecided Otherwise

The hypothesis of congeneric tests relies on the existence of communalities to be placed in the diagonal cells of the item covariance matrix, in order to reduce the rank of that matrix to one. The conditions under which this is possible have been known for a long time. **Spearman** [17] already noted that unidimensionality is impossible (except in contrived cases) when $k > 3$. Accordingly, when $k > 3$, factor analysis with only one common factor will never give perfect fit. More generally, Wilson and Worcester [23], Guttman [7], and Bekker and De Leeuw [1] have argued that rank reduction of a covariance matrix by communalities does not carry a long way. Shapiro [15] has proven that the minimal reduced rank that can possibly be achieved will be at or above the Ledermann bound [11] almost surely. It means that the minimal reduced rank is almost surely at or above 1 when $k = 3$, at or above 2 when $k = 4$, at or above 3 when $k = 5$ or 6, and so on. The notion of 'almost surely' reflects the fact that, although covariance matrices that do have lower reduced rank are easily constructed, they will never be observed in practice. It follows that the hypothesis of congeneric tests is nearly as unrealistic as that of essential tau-equivalence. It may be true only when there are three or fewer items.

Even when reduction to rank 1 is possible, this is not sufficient for the hypothesis to be true: We merely have a necessary condition that is satisfied. The example of X_1, X_2, and X_3 with three uncorrelated underlying factors Y_1, Y_2, and Y_3, given above in the context of tau-equivalence, may again be used to demonstrate this: There are three underlying factors, yet communalities that do reduce the rank to 1 do exist (being 1, 1, and 1). The bottom line is that the hypothesis of congeneric tests cannot be rejected, but still may be false when $k = 3$ or less, and it has to be rejected when $k > 3$. 'Model-based coefficients are not useful if the models are not consistent with the empirical data' [2]. Reliability coefficients based on the single-factor hypothesis are indeed a case where this applies.

Sampling Bias

Lower bounds to reliability do not rest on any assumption other than that error scores of the test parts correlate only with themselves and with the observed scores they belong with. On the other hand, lower bounds do have a reputation for sampling bias. Whereas coefficient alpha tends to slightly underestimate the population alpha [21, 24], Guttman's λ_4, and the greatest lower bound in particular, may grossly overestimate the population value when computed in small samples. For instance, when $k = 10$ and the population glb is 0.68, its average sample estimate may be as high as 0.77 in samples of size 100 [16].

It may seem that one-factor-based coefficients have a strong advantage here. But this is not true. When $k = 3$, Kristof's coefficient often coincides with glb, and for $k > 3$, numerical values of McDonald's coefficient are typically very close to the glb. In fact, McDonald's coefficient demonstrates the same sampling bias as glb in Monte Carlo studies [20]. Because McDonald's and other factor analysis–based coefficients behave very similarly to the glb and have the same bias problem, and, in addition, rely on a single-factor hypothesis, which is either undecided or false, the glb is to be preferred.

Bias Correction of the glb

Although the glb seems superior to single-factor-based coefficients of reliability, when the test is hypothesized to be unidimensional, this does not mean that the glb must be evaluated routinely for the single administration of an arbitrary test. The glb has gained little popularity, mainly because of the sampling bias problem. Bias correction methods are under construction [12, 22], but still have not reached the level of accuracy required for practical applications. The problem is especially bad when the number of items is large relative to sample size. Until these bias problems are over, alpha will prevail as the lower bound to reliability.

Reliability versus Unidimensionality

Reliability is often confused with unidimensionality. A test can be congeneric, a property of the true score parts of the items, yet have large error variances, a property of the error parts of the items, and the reverse is also possible. Assessing the degree of unidimensionality is a matter of assessing how closely the single factor fits in common factor analysis. Ten Berge and Sočan [20] have proposed a method of expressing unidimensionality as the percentage of common variance explained by a single factor in factor analysis, using the so-called Minimum Rank Factor Method of Ten Berge and Kiers [18]. However, this is a matter of taste and others prefer goodness-of-fit measures derived from maximum-likelihood factor analysis.

References

[1] Bekker, P.A. & De Leeuw, J. (1987). The rank of reduced dispersion matrices, *Psychometrika* **52**, 125–135.

[2] Bentler, P.M. (2003). Should coefficient alpha be replaced by model-based reliability coefficients? *Paper Presented at the 2003 Annual Meeting of the Psychometric Society*, Sardinia.

[3] Bentler, P.M. & Woodward, J.A. (1980). Inequalities among lower bounds to reliability: with applications to test construction and factor analysis, *Psychometrika* **45**, 249–267.

[4] Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests, *Psychometrika* **16**, 297–334.

[5] Gilmer, J.S. & Feldt, L.S. (1983). Reliability estimation for a test with parts of unknown lengths, *Psychometrika* **48**, 99–111.

[6] Guttman, L. (1945). A basis for analyzing test-retest reliability, *Psychometrika* **10**, 255–282.

[7] Guttman, L. (1958). To what extent can communalities reduce rank, *Psychometrika* **23**, 297–308.

[8] Jackson, P.H. & Agunwamba, C.C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I. Algebraic lower bounds, *Psychometrika* **42**, 567–578.

[9] Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric tests, *Psychometrika* **36**, 109–133.

[10] Kristof, W. (1974). Estimation of reliability and true score variance from a split of the test into three arbitrary parts, *Psychometrika* **39**, 245–249.

[11] Ledermann, W. (1937). On the rank of reduced correlation matrices in multiple factor analysis, *Psychometrika* **2**, 85–93.

[12] Li, L. & Bentler, P.M. (2004). *The greatest lower bound to reliability: Corrected and resampling estimators*. Paper presented at the 82nd symposium of the Behaviormetric Society, Tokyo.

[13] McDonald, R.P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis, *British Journal of Mathematical and Statistical Psychology* **23**, 1–21.

[14] Novick, M.R. & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements, *Psychometrika* **32**, 1–13.

[15] Shapiro, A. (1982). Rank reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis, *Psychometrika* **47**, 187–199.

[16] Shapiro, A. & Ten Berge, J.M.F. (2000). The asymptotic bias of minimum trace factor analysis, with applications to the greatest lower bound to reliability, *Psychometrika* **65**, 413–425.

[17] Spearman, C.E. (1927). *The Abilities of Man*, McMillan, London.

[18] Ten Berge, J.M.F. & Kiers, H.A.L. (1991). A numerical approach to the exact and the approximate minimum rank of a covariance matrix, *Psychometrika* **56**, 309–315.

[19] Ten Berge, J.M.F., Snijders, T.A.B. & Zegers, F.E. (1981). Computational aspects of the greatest lower bound to reliability and constrained minimum trace factor analysis, *Psychometrika* **46**, 357–366.

[20] Ten Berge, J.M.F. & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality, *Psychometrika* **69**, 611–623.

[21] Van Zijl, J.M., Neudecker, H. & Nel, D.G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha, *Psychometrika* **65**, 271–280.

[22] Verhelst, N.D. (1998). Estimating the reliability of a test from a single test administration, (Measurement and Research Department Report No. 98-2), Arnhem.

[23] Wilson, E.B. & Worcester, J. (1939). The resolution of six tests into three general factors, *Proceedings of the National Academy of Sciences* **25**, 73–79.

[24] Yuan, K.-H. & Bentler, P.M. (2002). On robustness of the normal-theory based asymptotic distributions of three reliability coefficient estimates, *Psychometrika* **67**, 251–259.

Jos M.F. Ten Berge

Teaching Statistics to Psychologists

George Spilich

Volume 4, pp. 1994–1997

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Teaching Statistics to Psychologists

In the introduction to his 1962 classic statistical text, Winer [7] describes the role of the statistician in a research project as similar to that of an architect; that is, determining whether the efficacy of a new drug is superior to that of competing products is similar to designing a building with a particular purpose in mind and in each case, there is more than one possible solution. However, some solutions are more elegant than others and the particulars of the situation, whether they are actual patient data or the size and placement of the building site, place boundaries on what can and cannot be accomplished.

What is the best way to teach the science and art of designing and conducting data analysis to today's graduate students in psychology? Perhaps history can be our guide, and so we begin by first asking what graduate education in statistics was like when today's senior faculty were students, then ask what is current common practice, and finally ask what the future might hold for tomorrow's graduate students. I propose the following:

- Graduate training in statistics has greatly changed over the last few decades in ways that are both helpful and harmful to students attempting to master statistical methodology.
- The major factor in this change is the development of computerized statistical packages (*see* **Software for Statistical Analyses**) which, when used in graduate education, cause students trained in **experimental design** to be more broadly but less thoroughly trained.
- 'Point and click' statistical programs allow individuals without professional training access to procedures they do not understand. The availability of statistical packages to individuals without professional statistical training might lead to a guild environment where psychologists could play an important role.

The Recent Past

Perusal of 'classic' texts such as Lindquist [5], Kirk [4], McNemar [6], Winer [7], Guilford and Fruchter [1], Hayes [2], and Keppel [3] suggests that 30 to 35 years ago (or just one generation ago in terms of the approximate length of an academic career), psychology graduate education emphasized descriptive measures, correlational techniques, and especially multiple regression (*see* **Multiple Linear Regression**) and the **analysis of variance** including *post hoc* tests and trend analysis (*see* **Multiple Comparison Procedures**). Statistical techniques were taught via hand calculation methods using small data sets. Computationally demanding methods such as **factor analysis**, time series analysis, **discriminant** and cluster analysis (*see* **Cluster Analysis: Overview**) as well as residual analysis (*see* **Residuals**) in multiple regression and multivariate analysis of variance (*see* **Multivariate Analysis: Overview**) were not typically presented to all students. These advanced techniques were generally presented to students whose professional success would depend upon mastery of those particular skills. For instance, a graduate student preparing for a career investigating personality, intelligence, or social psychology was much more likely to receive thorough training in factor analytic techniques (*see* **History of Factor Analysis: A Psychological Perspective**; **Factor Analysis: Confirmatory**) than a student studying classical or operant conditioning. There is little need for understanding the difference between varimax and orthomax rotations in factor analysis if one is pointing to a career observing rats in operant chambers. Statistical training of that era was subdiscipline specific.

For all students of this bygone era, training heavily emphasized experimental design. Graduate students of 30–35 years ago were taught that the best way to fix a problem in the data was not to get into trouble in the first place, and so the wise course of action was to employ one of several standard experimental designs. Texts such as Lindquist [5] and Kirk [4] instructed several decades' worth of students in the pros and cons of various experimental designs as well as the proper course of analysis for each design.

Current State

What has changed in graduate training since then and why? The answer is that computer-based statistical packages have become commonplace and altered the face of graduate statistical education in psychology. In times past, data analysis was so time consuming and labor intensive that it had to be planned carefully, prior to conducting the actual study. In that era, one could not easily recover from design errors through statistical control, especially when the size of the data set was large. Today, studies involving neuroimaging techniques such as fMRI and EEG result in tens of thousands and even millions of data points per subject, all potentially requiring a baseline correction, an appropriate transformation, and, possibly, covariates. The sheer hand labor of such an analysis without computer assistance is beyond imagination.

Currently, most graduate programs in psychology expect their students to gain competency in some statistical package, although there is a considerable amount of variability in how that goal is met. A survey of 60 masters and doctoral programs in psychology at colleges and universities in the United States provides a glimpse of the statistical packages used in graduate education. Individual faculty members who indicated that they were responsible for graduate training in statistics were asked if their department had a standard statistical package for graduate training. All 60 respondents replied that some statistical package was a part of graduate training and their responses are presented in Table 1.

These same faculty members were asked how graduate students gained mastery of a statistical package; their responses are presented in Table 2. While most graduate programs appear to have a formal course dedicated to teaching a particular statistical package, some programs integrate the statistical package into laboratory courses or teach it as part of an advisor's research program. One rationale for teaching the use of a statistical package during research training instead of in a formal course is that with the former, students are likely to learn material appropriate to their immediate professional life. A contrasting view is that in a formal course that cuts across the domain boundaries of psychology, students are exposed to a wide variety of statistical techniques that may be useful in the future. Because many professionals start their career in one field but end up elsewhere in academia or in industry, breadth of skill may be preferable to depth, at least for the intermediate level of graduate training.

Table 1 Standard statistical packages in graduate programs ($N = 60$)

Statistical program of choice	Percent respondents
SPSS	45
SAS	15
Minitab	5
Systat	5
No standard	25
Other	5

Table 2 Methods that graduate programs use to ensure student mastery of statistical packages ($N = 60$)

How is a statistical package taught?	Percent respondents
Through a dedicated course	55
No course; through lab and research	25
Both course and lab/research	15
No standard method	5

Statistical packages also provide graduate students with tools that were not typically a part of the common curriculum a generation ago. Respondents to the survey indicated that cluster analysis, discriminant analysis, factor analysis, binary logistic regression, and categorical regression are processes generally presented to graduate students because they have access to a statistical package. Furthermore, a wide assortment of two- and three-dimensional graphs and other visual techniques for exploring relationships are now presented to graduate students.

What is the Impact of Statistical Packages on Today's Graduate Education?

Providing graduate students with access to statistical packages has dramatically changed the breadth of their education. To assay the magnitude of the change, faculty in the survey were asked what topics they would keep or drop if their current statistical package was no longer available for graduate education. Faculty replied that anova, extensive discussion of *post hoc* analysis, orthogonal contrasts, power analysis and eta-squared (*see* **Effect Size Measures**) would still be covered but with very simple examples. Currently, they expect their students to fully master such techniques for a wide variety of settings. They also indicated that a variety of regression techniques would still be taught as would the fundamentals of factor analysis.

However, the survey respondents agreed almost unanimously that coverage of statistical techniques that are computationally intensive such as manova, cluster analysis, discriminant analysis, and complex factor analytic techniques (*see* **Factorial Designs**; **Repeated Measures Analysis of Variance**) would be reduced or eliminated if students did not have access to a statistical package. The comments of these professional educators suggest that today's graduate students are exposed to a wider range of statistical techniques and are considerably more adept at data manipulations and massaging than their counterparts of 30–35 years ago.

What is the Future?

The evolution of statistical programs suggests that they will become 'smarter' and will interact with the user by suggesting analyses for particular data sets. As the programs become easier to use and available to a wider audience, will biostatisticians still be needed? Perhaps the answer lies in Winer's [7] description of statisticians as architects. Today, one can easily purchase a computer program that helps design a kitchen, office, or home. Have these programs eliminated the need for architects? The answer is 'not at all', except for the simplest applications. When homeowners can use a software product to play with the design of an addition or a house, they are more likely to conceive of a project that requires the skill of a professional. The same may happen in biostatistics. Just as certification as an architect is necessary for a practitioner's license, we may see the emergence of a 'guild' that certifies professional statisticians after sufficient coursework and some demonstration of technical ability. Statistical packages can impart skill, but they cannot impart wisdom. The challenge for faculty who are training tomorrow's graduate students in biostatistics is to ensure that we impart more than knowledge of statistics; we need to teach the wisdom that comes from experience.

Conclusion

It would be easy to assume that the impact of statistical packages upon graduate education is merely to increase the range of techniques that are taught. The real sea change is in how the analytic process itself is approached conceptually. Before the advent of statistical packages, a student was taught to carefully plan analyses prior to actual computation, and this plan guided as well as constrained the design. In contrast, today's students are taught to break the data into various subsets and then to examine it from all angles. Such sifting and winnowing is so time consuming if performed by hand that extensive exploratory data analyses that are common today were all but impossible in the earlier era. Students now have the possibility of seeing more in the data than was possible a generation ago but at the cost of a deeper understanding of how the view was derived.

References

[1] Guilford, J.P. & Fruchter, B. (1973). *Fundamental Statistics in Psychology and Education*, 5th Edition. McGraw-Hill, New York.

[2] Hayes, W.L. (1973). *Statistics for the Social Sciences*, 2nd Edition, Harcourt Brace, New York.

[3] Keppel, G. (1973). *Design and Analysis: A Researcher's Handbook*, Prentice Hall, Englewood Cliffs.

[4] Kirk, R.E. (1968). *Experimental Design: Procedures for the Behavioral Sciences*, Brooks-Cole, Belmont.

[5] Lindquist, E.F. (1953). *Design and Analysis of Experiments in Psychology and Education*, Houghton-Mifflin, Boston.

[6] McNemar, Q. (1969). *Psychological Statistics*, 4th Edition, Wiley, New York.

[7] Winer, B.J. (1962). *Statistical Principles in Experimental Design*, McGraw-Hill, New York.

George Spilich

Teaching Statistics: Sources

Bernard C. Beins

Volume 4, pp. 1997–1999

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Teaching Statistics: Sources

Teachers of statistics are aware of the dismal reputation of their discipline among students, but they work to improve it by displaying delight in the topic and by using clever demonstrations and activities. Fortunately, there is a population of useful, interesting material available. The books and periodicals reviewed here include sources that stand alone as fascinating reading for teachers and students, as well as books that appeal to teachers who want to enliven their classes. The comments here proceed from books to periodicals, with more general material preceding the more technical.

The 1954 classic by Huff [4], *How to lie with statistics,* is an entertaining depiction of statistics in everyday life. Its main limitation is that its content is, at times, quite outdated. Few of us would now be impressed with college graduates who earn a beginning salary of $10 000. Although the examples do not always lend themselves to updating, they illustrate universal and timeless pitfalls. Instructors can generate current examples of these pitfalls.

A recent philosophical descendent of Huff's book is Best's [1] *Damned Lies and Statistics*, which provides many examples of dubious statistics and how they originated. The book opens with what Best considers the worst possible social statistic: the 'fact' that the number of children gunned down in the United States has doubled every year since 1950. The book documents the history of this number. It also includes sources of bad statistics, 'mutant' statistics that arise when original values are misinterpreted or distorted, and inappropriate comparisons using superficially plausible statistics. Best illustrates how statistics that are important to social issues can take on a life of their own and confuse rather than illuminate those issues. Students will enjoy this book and instructors can use it to enhance classroom presentations and discussions.

Paulos's [5] entertaining book, *Once upon a number: The hidden mathematical logic of stories*, takes the same approach as Best's. Both cite examples of events in everyday life that relate to social controversies and public policies. The anecdotes selected by Paulos illustrate the general principles of statistics and probabilities (*see* **Probability: An Introduction**) found in many facets of life. He offers compelling examples of how to think using statistics. One of his important messages is that generalizations from a single instance can be faulty because they may be highly idiosyncratic. He also identifies questions that statistics can answer and how to appropriately apply those statistics. For instance, he discusses the statistics associated with why we invariably have to wait longer in line than we think we should, a phenomenon of interest to those of us who always think we pick the wrong line. Paulos invokes probability to show that minority populations are statistically (and realistically) more prone to encountering racism than are people in majority populations. Because many of the illustrations involve people and their behavior, this book is particularly useful for statistics classes in the behavioral and social sciences. This volume uses the same clear approach that Paulos took in his previous books on numerical literacy.

Readers get a more technical, but still highly readable, presentation of statistics and probability in Holland's [3] volume, *What are the chances?* Examples range widely. They include how the random distribution of rare events such as cancer creates clusters in a few locations (neighborhoods or buildings), the probabilities of Prussian soldiers being kicked to death by their horses between 1875 and 1894, and probabilities linked to waiting times in queues and traffic jams. The examples are supported by formulas, tables, and graphs. Holland emphasizes that the meaning of the data comes from their interpretation, but that interpretations are fraught with their own difficulties. Sampling is emphasized as is measurement error, which is explained quite nicely. Readers generally familiar with statistical notation and the rudiments of normal distributions (*see* **Catalogue of Probability Density Functions**) and factorials (i.e., instructors who have completed a graduate level statistics course) will be able to fathom the points Holland makes. This book shows both the use and limitations of statistics.

Both the Paulos and the Holland books can serve to augment classroom discussions. In addition, both are likely to be accessible to competent and motivated students who have mastered the rudimentary statistical concepts.

A quite different book by Gelman and Nolan [2], *Teaching statistics: A bag of tricks*, focuses on activities and demonstrations. The authors' audience is statistics teachers in secondary schools and colleges. An introductory topics section contains demonstrations and activities that match the contents of virtually any beginning statistics course in the behavioral and social sciences. The numerous demonstrations and activities illustrate important statistical concepts and provide excellent exercises in critical thinking. Consequently, the instructor can use varied exercises in different classes. The in-class activities include a simple memory demonstration that can be used to illustrate confidence intervals and an evaluation of newspaper articles that report data. The authors present the important message that researchers are bound by ethical principles in their use of data. The second segment of Gelman and Nolan's book gives the logistics for the successful implementation of the demonstrations and activities. This part of the book will certainly be of interest to beginning teachers, but it contains helpful hints for veterans as well.

Statistics teachers can benefit from two of the handbooks of articles reprinted from the journal *Teaching of Psychology* [6] and [7]. The two volumes devoted to teaching statistics contain nearly five dozen articles. These books resemble that of Gelman and Nolan in that they all feature activities and demonstrations that instructors have used successfully in their classes. The entries in the handbooks were written by a diverse set of statistics teachers and cover a wide variety of topics. The examples are broad enough to be useful to instructors in all of the behavioral sciences. The topics of many articles overlap with those of Gelman and Nolan, but the entries in the handbooks provide pedagogical advice as well as activities and demonstrations. For instance, there are sections on topics such as developing student skills, evaluating successes in statistics, and presenting research results. Many of the activities and demonstrations in these handbooks are accompanied by at least a basic empirical evaluation of their effects on student learning.

In addition to books, several periodicals publish activities, demonstrations, and lecture enhancements The quarterly magazine *Chance* (not be confused with the gambling magazine with the same title) publishes articles on topics such as the misuse of statistics in the study of intelligence, **Bayesian statistics** in cancer research, teacher course evaluations, and the question of who wrote the 15th book of Oz. Feature articles illustrate the value of relying on statistical knowledge to help address the important issues in our lives. The articles are typically very engaging, although the authors do not dumb down the content. Sometimes the reading requires diligence because of the complexity of the statistical issues, but the articles are worth the work. Most of the time, readers with a modicum of knowledge of statistical ideas and notation will find the writing accessible. (One vintage article [from 1997] discussed the statistics associated with waiting in lines for various services. Statisticians' fascination with time spent waiting suggests that they spend an inordinate amount of time doing nothing, but that they are quite productive during those times.) *Chance* has regular columns on sport, visual presentation of data, book reviews, and other topics.

A periodical with a more narrow orientation is the *Journal of Statistics Education*, published by the American Statistical Association three times a year. This online journal is available free on the Internet. Its featured articles involve some aspect of pedagogy. A recent volume included articles on teaching power and sample size, using the Internet in teaching statistics, the use of analogies and heuristics in teaching statistics, and the use of student-specific datasets in the classroom. In addition, a column 'Teaching Bits: A Resource for Teachers of Statistics' offers brief excerpts of current events that are of relevance to teaching statistics. The journal also offers data sets that can be downloaded from the Internet.

The journal *Teaching of Psychology* publishes articles on teaching statistics. As the journal title suggests, the material is oriented toward the discipline of psychology, but there are regular articles on teaching statistics that are suitable for other behavioral sciences. This journal is the organ of the Society for the Teaching of Psychology. A parallel journal, *Teaching Sociology*, also publishes occasional articles on teaching statistics and research methods.

The periodical, *American Statistician*, includes the 'Teacher's Corner'. The entries associated with teaching are often quite technical and involve advanced topics in statistics. They are less likely to be relevant to students in the behavioral sciences.

Psychological Methods appears quarterly. The articles are usually fairly technical and are addressed to sophisticated researchers. Recent topics included new approaches to regression analyses (*see* **Regression Models**) **meta-analysis**, and **item response**

theory (IRT). The material in this journal is best suited for advanced students. Although the intent of most articles is not pedagogical, the articles can help instructors bring emerging statistical ideas to their classrooms.

Finally, the Open Directory Project (`http://www.dmoz.org/Science/Math/Statistics/`) is an Internet site that provides a wealth of information on teaching statistics. This site provides a compendium of useful web addresses on a vast array of topics, including statistics education. The web pages to which the Open Directory Project sends you range from light-hearted and humorous (the three most common misspellings of statistics) to the very serious (e.g., an interactive Internet environment for teaching undergraduate statistics). There is also an interesting link to a web page that evaluates the use of statistics in the media. The Open Directory Project site has links to virtually any topic being covered in an introductory level statistics class, as well as links to more advanced topics for higher level students.

References

[1] Best, J. (2001). *Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists*, University of California Press, Berkeley.

[2] Gelman, A. & Nolan, D. (2002). *Teaching Statistics: A Bag of Tricks*, Oxford University Press, Oxford.

[3] Holland, B.K. (2002). *What are the Chances? Voodoo Deaths, Office Gossip, and other Adventures in Probability*, Johns Hopkins University Press, Baltimore.

[4] Huff, D. (1954). *How to lie with Statistics*, Norton, New York.

[5] Paulos, J.A. (1998). *Once upon a Number: The Hidden Mathematical Logic of Stories*, Basic Books, New York.

[6] Ware, M.E. & Brewer, C.L. (1999). *Handbook for Teaching Statistics and Research Methods*, 2nd Edition, Erlbaum, Mahwaw.

[7] Ware, M.E. & Johnson, D.E. (2000). *Handbook of Demonstrations and Activities in the Teaching of Psychology, Vol. 1: Introductory, Statistics, Research Methods, and History*, 2nd Edition, Erlbaum, Mahwah.

BERNARD C. BEINS

Telephone Surveys

Robyn Bateman Driskell

Volume 4, pp. 1999–2001

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Telephone Surveys

In recent years, telephone surveys have become increasingly popular and almost commonplace. When the methodology of telephone surveys was developed in the 1970s, many assumed that telephone surveys would replace face-to-face surveys for the most part. While this did not occur, telephone surveying is the preferred approach in many cases [2]. Telephone survey methodologies have undergone dramatic changes and examination in the last 20 years. Telephone survey methodology is widely used. As a result of recent advances in telephone technology, the methodology is viewed as both valid and reliable. Telephone surveys are efficient and effective means of collecting data that can aid decision-making processes in both the public and private sectors [5].

When designing the questionnaire, it is often easier and safer to borrow questions that have been used and tested previously (*see* **Survey Questionnaire Design**). This allows for comparability of the questions across time and place. The *introductory spiel* is the standardized introduction read by an interviewer when contact is made with a possible eligible household or respondent. A carefully worded introduction is of utmost importance. A weak introductory spiel can lead to refusals and nonresponse error. During the introduction, the potential respondent decides whether to cooperate [3]. The credibility of the interviewer and the survey must be established in the introduction. The introduction must reduce the respondent's fears and skepticism by providing assurances of legitimacy. Lyon suggests [5] that the introduction should *always* include (a) the interviewer's full name; (b) the organization and/or its sponsor that is conducting the research; (c) the survey's general topics; (d) the procedure for selection; (e) a screening technique; and (f) an assurance of confidentiality. Following the introduction, the screening technique selects a particular individual within the household to be interviewed. The screening technique is designed to systematically select respondents by age and sex, so that every individual in each sampled household has an equal probability of being selected, thereby ensuring a representative sample.

Advantages of Telephone Surveys

Telephone surveys have numerous advantages. The most important advantage is the ability to maintain quality control throughout the data collection process [4]. A second advantage is cost efficiency. Telephone surveys are much less expensive than face-to-face interviews but more expensive than mail surveys.

The third major advantage of telephone surveys is their short turn around time. The speed with which information is gathered and processed is much faster than that of any other survey method. A telephone survey takes 10 to 20% less time than the same questions asked in a face-to-face interview [4]. When a central interviewing facility with several phone banks is used, it is possible for 10 callers to complete approximately 100 interviews in an evening. Hence, large nationwide studies can be completed in a short time. By polling only a few hundred or a few thousand persons, researcher can obtain accurate and statistically reliable information about tens of thousands or millions of persons in the population. This assumes, of course, that proper techniques are implemented to avoid survey errors in sampling, coverage, measurement, and nonresponse [2].

A fourth advantage is the ability to reach most homes as a result of methodological advances in random-digit dialing (RDD) and the proliferation of phones. It is estimated by the US census that approximately 97% of US households have a telephone. RDD has improved telephone survey methodology. RDD uses a computer to select a telephone sample by random generation of telephone numbers. There are several different techniques for generating an RDD sample. The most common technique begins with a list of working exchanges in the geographical area from which the sample is to be drawn. The last four digits are computer generated by a random procedure. RDD procedures have the advantage of including unlisted numbers that would be missed if numbers were drawn from a telephone directory. Telephone numbers also can be purchased from companies that create the sampling pool within a geographic area, including numbers for RDD surveying (*see* **Survey Sampling Procedures**).

A fifth advantage is that telephone interviewers do not have to enter high-crime neighborhoods or enter people's homes, as is required for face-to-face interviewers. In some cases, a respondent will be more honest in giving socially disapproved answers if they do not have to face the interviewer. Likewise, it is possible to probe into more sensitive areas over the phone than it is in face-to-face interviews [1]. The major differences between telephone interviewing and face-to-face interviewing is that the interviewer's voice is the principal source of interviewing bias. By not seeing the interviewer, respondents are free from the biases that would be triggered by appearance, mannerisms, gestures, and expressions. On the other hand, the interviewer's voice must project an image of a warm, pleasant person who is stimulating to talk to and who is interested in the respondent's views.

Other advantages found in [5] include the ability to probe, the ability to ask complex questions with complex skip patterns (with computer-assisted telephone-interviewing, CATI), the ability to use long questionnaires, the assurance that the desired respondent completes the questionnaire, and the ability to monitor the interviewing process.

CATI is a survey method in which a printed questionnaire is not used. Instead, the questions appear on the screen of a computer terminal, and the answers are entered directly into a computer via the keyboard or mouse. The major advantages of this procedure are that it allows for the use of a complex questionnaire design with intricate skip patterns. It also provides instant feedback to the interviewer if an impossible answer is entered, and it speeds up data processing by eliminating intermediate steps. The computer is programmed not only to present the next question after a response is entered but also to determine from the response exactly which question should be asked next. The computer branches automatically to the next question according to the filter instructions. CATI can randomly order the sequence of possible response categories and incorporate previous answers into the wording of subsequent items. The CATI software also can control the distribution of the sampling pool and dial the appropriate phone number for the interviewer. During the polling process, the supervisory staff are able to access the interviewer's completed calls, the duration of each call, response rates, and listen in to assure the accuracy of the script. When CATI is properly implemented, the quality of the data is improved and survey errors are often reduced [2, 4, 5].

Disadvantages of Telephone Surveys

Despite the numerous advantages, telephone surveys have limitations. During a telephone interview, it is not possible to use visual aids. The length of the survey is another concern. Owing to respondent fatigue, most telephone surveys should not be longer than 15 min. Face-to-face interviews can be longer, up to 20 to 30 min. In face-to-face interviews, the interviewer can assess body language and notice respondent fatigue. While complex skip patterns are easy to use with the CATI system, long, involved questions with many response categories are hard to follow in a telephone interview. The telephone methodology is hampered by the proliferation of marketing and sales calls veiled as 'research'. Many respondents distrust telephone surveys and want to know what you are selling. Also, a shortcoming of the telephone survey is the ease with which a potential respondent can hang up. It is quite easy to terminate the interview or make up some excuse for not participating at that time [1]. Another problem is the potential coverage error – every demographic category (i.e., sex, age, and gender) is not equally willing to answer the phone and complete a survey – although this can be avoided with an appropriate screening technique. Technological advances such as answering machines and caller ID have contributed to nonresponse rates as potential respondents screen incoming calls. Despite the several limitations of telephone survey methods, the advantages usually outweigh the disadvantages, resulting in an efficient and effective method for collecting data.

References

[1] Babbie, E. (1992). *The Practice of Social Research*, 6th Edition, Wadsworth Publishing Company, Belmont.

[2] Dillman, D.A. (2000). *Mail and Internet Surveys: The Tailored Design Method*, 2nd Edition, John Wiley, New York.

[3] Dijkstra, W. & Smit, J.H. (2002). Persuading reluctant recipients in telephone surveys, in *Survey Nonresponse*, R.M. Groves, D.A. Dillman, J.L. Eltinge, J.L. Little, & J.A. Roderick, eds, John Wiley, New York.

[4] Lavrakas, P.J. (1998). Methods for sampling and interviewing in telephone surveys, in *Handbook of Applied*

Social Research Methods, L. Brickmand & D.J. Rog, eds, Sage Publications, Thousand Oaks, pp. 429–472.

[5] Lyon, L. (1999). *The Community in Urban Society*, Waveland Press, Prospect Heights.

(*See also* **Randomized Response Technique**)

ROBYN BATEMAN DRISKELL

Test Bias Detection

Michal Beller

Volume 4, pp. 2001–2007

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Test Bias Detection

During the 1970s considerable attention was given to developing fair selection models in the context of college admissions and job entry. These models put heavy emphasis on predictive validity, and in one way or another they all address the possibility of differences in the predictor–criterion relationship for different groups of interest. With the exception of Cleary's regression model, most other models proposed were shown to be mutually contradictory in their goals and assumptions.

There is a growing consensus in the measurement community that bias refers to any construct-irrelevant source of variance that results in systematically higher or lower scores for identifiable groups of examinees. In regard to test use, the core meaning of fairness is *comparable validity*: A fair test is one that yields comparable and valid scores from person to person, group to group, and setting to setting. However, fairness, like validity, is not just a psychometric issue. It is also a social value, and therefore alternative views about its essential features will persist.

As the use of educational and psychological tests continues to grow, an increasing number of decisions that have profound effects on individuals' lives are being made based on test scores, despite the fact that most test publishers caution against the use of a single score for decision-making purposes. Tests are instruments that are designed to provide evidence from which inferences are drawn and on the basis from which such decisions are made. The degree to which this evidence is credible constitutes the validity of the test, and it should hold for all groups among the intended test-taking population (*see* **Validity Theory and Applications**).

Concerns of possible gender and/or ethnic biases in the use of various tests have drawn the attention of test users and test developers as well as of the public in general. A view of fairness, frequently used by the public, involves equality of testing outcomes. In the public debate on this issue the terms *test unfairness, cultural unfairness*, and *test bias* are often used interchangeably to refer to differences in test performance among subgroups of social interest. However, the idea that fairness requires overall performance or passing rates to be equal across groups is not the one generally accepted in the professional literature. In fact, Cole and Zieky [12] claim that: 'If the members of the measurement community currently agree on any aspect of fairness, it is that score differences are not proof of bias' (p. 375). This is because test outcomes may validly document group differences that are real and that may be reflective, in part, of unequal opportunity to learn and other possible reasons.

Bias refers to any construct[1] under-representation or construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers. Construct-irrelevant variance exists when the 'test contains excess reliable variance that is irrelevant to the interpreted construct' [30]. The effect of such irrelevant sources of variance on scores is referred to as *measurement bias*. Sources of irrelevant variance that result in systematically higher or lower scores for members of particular groups are of potential concern for both predictors and criteria. Determining whether measurement bias is present is often difficult, as this requires evaluating an observed score in relation to the unobservable construct of interest.

Fairness in Terms of Equitable Treatment of All Examinees

One way to reduce measurement bias is to assure equitable treatment of all examinees through test design and development practices intended from an early stage to prevent bias. Equity in terms of testing conditions, access to practice materials, performance feedback, retest opportunities, and providing reasonable accommodations for test-takers with disabilities when appropriate, are important aspects of fairness. Equitable treatment of all examinees is directly related to establishing construct validity and the fairness of the testing process [42]. It is useful to distinguish two kinds of comparability relevant to construct validity and fairness – *comparability of score interpretation* and *task comparability. Comparability of score interpretation* means that the properties of the test itself and its external relationships with other variables are comparable across groups and settings. Comparability of score interpretation is important in justifying uniform score use for different groups and in different circumstances. *Task comparability* means that the test task elicits the same cognitive processes across different groups and different circumstances.

Fairness of Accommodations. Within task comparability, two types of processes may be distinguished: those that are relevant to the construct measured and those that are irrelevant to the construct but nonetheless involved in task performance (possibly like reading aloud in mathematics). Comparability of construct-relevant processes across groups is necessary for validity. Ancillary or construct-irrelevant processes may, however, be modified without jeopardizing score interpretation. This provides a fair and legitimate basis for accommodating tests to the needs of students with disabilities and those who are English-language learners [40]. Thus, comparable validity and test fairness do not necessarily require identical task conditions, but rather common construct-relevant processes with ignorable construct-irrelevant or ancillary processes that may be different across individuals and groups. Such accommodations must be justified with evidence that score meanings have not been eroded in the process.

The general issue of improving the accessibility of a test must be considered within an assessment design framework that can help the assessment planner maximize validity within the context of specific assessment purposes, resources, and other constraints. Evidenced-centered assessment design (ECD), which frames an assessment as embodying an evidentiary argument, has been suggested as a promising approach in this regard [21].

Detecting Bias at the Item Level. Statistical procedures to identify test items that might be biased have existed for many years [3, 29]. One approach to examining measurement bias at the item level is to perform a **differential item functioning (DIF)** analysis, focusing on the way that comparable or *matched* people in different groups perform on each test item. DIF procedures are empirical ways to determine if the item performance of comparable subgroups is different. DIF occurs when a statistically significant difference in performance on an individual test item occurs across two or more groups of examinees, *after* the examinees have been matched on total test/subtest scores [20, 22]. DIF methodologies and earlier methods share a common characteristic in that they detect unfairness of a single item relative to the test as a whole, rather than detecting pervasive unfairness. Thus, in the extremely unlikely case that all items were biased to exactly the same degree against exactly the same groups, no items would be identified as unfair by current DIF methods [42].

DIF analysis is not appropriate in all testing situations. For example, it requires data from large samples and if only for this reason DIF analyses are more likely to be used in large-scale educational settings and are not likely to become a routine or expected part of the test development and validation process in employment settings.

DIF methods have been an integral part of test development procedures at several major educational test publishers since the 1980s [41]. Empirical research in domains where DIF analyses are common has rarely found sizable and replicable DIF effects [34]. However, it is worth noting that aggregations of DIF data have often led to generalizations about what kind of items to write and not to write. This in turn led to two outcomes: (a) fewer items with significant DIF values were found, because those with large DIF values were no longer being written; and (b) test fairness guidelines, which had previously been very distinct from empirical evaluations of test fairness, began to incorporate principles and procedures based on empirical (DIF) findings, rather than rely exclusively on the touchstone of social consensus on removing offensive or inappropriate test material.

Linked to the idea of measurement bias at the item level is the concept of an item sensitivity review. The fundamental purpose of fairness reviews is to implement the social value of not presenting test-takers with material that they are likely to find offensive or upsetting. In such a review, items are reviewed by individuals with diverse perspectives for language or content that might have differing meanings to members of various subgroups. Even though studies have noted a lack of correspondence between test items identified as possibly biased by statistical and by judgmental means [7, 37], major test publishers in the United States have instituted judgmental review processes designed to identify possibly unfair items and offensive, stereotyping, or alienating material (e.g., [16]).

Fairness Through the Test Design Process

It is important to realize that group differences can be affected intentionally or unintentionally by choices made in test design. Many choices made in the design of tests have implications for group differences. Such differences are seldom completely avoidable. Fair test design should, however, provide examinees comparable opportunity, insofar as possible, to demonstrate knowledge and skills they have acquired that are relevant to the purpose of the test [39]. Given that in many cases there are a number of different ways to predict success at most complex activities (such as in school or on a job), test developers should carefully select the relevant constructs and the ways to operationalize the measurement of those constructs to provide the best opportunity for all subgroups to demonstrate their knowledge.

As guidance to the test-construction process, *ETS Standards for Quality and Fairness* state that: 'Fairness requires that construct-irrelevant personal characteristics of test-takers have no appreciable effect on test results or their interpretation' (p. 17) [16]. More specifically, ETS standards recommends adopting the following guidelines: (a) treat people with respect in test materials; (b) minimize the effects of construct-irrelevant knowledge or skills; (c) avoid material that is unnecessarily controversial, inflammatory, offensive, or upsetting; (d) use appropriate terminology to refer to people; (e) avoid stereotypes; and (f) represent diversity in depictions of people.

Fairness as Lack of Predictive Bias

During the mid-1960s, concurrent with the Civil Rights Movement, measurement professionals began to pay increasing attention to score differences on educational and psychological tests among groups (often referred to as *adverse impact*). Considerable attention has been given to developing fair selection models in the context of college admissions and job entry. These models put heavy emphasis on predictive validity, and in one way or another they all address the possibility of differences in the predictor-criterion relationship for different groups of interest.

Fair Selection Models – mid-1960s and 1970s. Differential prediction (also called *predictive bias*; see AERA/APA/NCME *Standards*, 1999, for definition [2]) by race and gender in the use of assessment instruments for educational and personnel selection has been a long-standing concern. In trying to explicate the relation between prediction and selection, Willingham and Cole [39] and Cole [10] have pointed out that prediction has an obvious and important bearing on selection, but it is not the same thing – prediction involves an expected level of criterion performance given a particular test score; selection involves the use of that score in decision-making. Cleary's [8] model stipulated that no predictive bias exists if a common regression line can describe the predictive relationship in the two groups being compared. Other selection models [9, 14, 26, 38] were developed in the 1970s taking into account the fact that any imperfect predictor will fail to select members of a lower-scoring group in proportion to their rate of criterion success. The models all require setting different standards of acceptance for individuals in different groups to achieve group equity as defined by the model. Petersen and Novick [32] pointed out that the various models are fundamentally incompatible with one another at the level of their goals and assumptions and lead to contradictory recommendations, unless there is perfect prediction – and even given perfect prediction, these are competing selection models based on differing selection goals, and will thus always be mutually contradictory. As a result of the continuous public and professional discussion around test bias there was a growing recognition that fair selection is related to fundamental value differences. Several utility models were developed that go beyond the above selection models in that they require specific value positions to be articulated [13, 18, 32, 35]. In this way, social values are explicitly incorporated into the measurement involved in selection models. Such models were seldom utilized in practice, as they require data that are difficult to obtain. More importantly, perhaps, they require application of dichotomous definitions of success and failure that are themselves methodologically and conceptually rather arbitrary and problematic.

After the intense burst of fairness research in the late 1960s and early 1970s described above, Flaugher [17] noted that there was no generally accepted definition of the concept 'fair' with respect to testing. Willingham and Cole [39] concluded that the effort to determine which model, among the variety of such models proposed, best represented fair selection was perhaps the most important policy debate of the 1970s among measurement specialists.

Current View of Test Bias

The contemporary professional view of bias is that it is an aspect of validity. A commonly used definition of test bias is based on a lack of predictive bias. The AERA/APA/NCME Standards [2] defines predictive bias as 'the systematic under- or overprediction of criterion performance for people belonging to groups differentiated by characteristics not relevant to criterion performance' (p 179). This perspective (consistent with the Cleary model) views predictor use as unbiased if a common regression line can be used to describe the predictor-criterion relationship for all subgroups of interest. If the predictive relationship differs in terms of either slopes or intercepts, bias exists because systematic errors of prediction would be made on the basis of group membership [11, 19, 25, 27, 28, 31]. Whether or not subgroup differences on the predictor are found, predictive bias analyses should be undertaken when there are questions about whether a predictor and a criterion are related in the same way for relevant subgroups.

Several technical concerns need to be considered when trying to quantify predictive bias:

1. Analysis of predictive bias requires an unbiased criterion. It is possible for the same bias to exist in the predictor and the criterion, or it may be that there are different forms of bias in the criterion itself across groups, and that these biases might, however unlikely that may be, would cancel each other out.
2. The issue of statistical power to detect slope and intercept differences should be taken into account. Small total or subgroup sample sizes, unequal subgroup sample sizes, range restriction, and predictor or criterion unreliability are factors contributing to low power [1].
3. When discussing bias against a particular group in the admissions process, the entire set of variables used in selection should be taken into consideration, not just parts of it. The inclusion of additional variables may dramatically change the conclusions about predictive bias [26, 33].

Differential prediction by race has been widely investigated in the domain of cognitive ability. For White-African American and White-Hispanic comparisons, slope differences are rarely found. While intercept differences are not uncommon, they typically take the form of overprediction of minority group performance [4, 15, 23, 24, 36]. Similar results were found by Beller [5, 6] for the Psychometric Entrance Test (PET) used as one of two components for admissions to higher education in Israel. There were few indications of bias in using PET, and when found, they were in *favor* of the minority groups. In other words, the use of a common regression line overpredicted the criterion scores for the minority groups.

Studies investigating sex bias in tests (e.g., [40]) found negligible opposing effects of under- and overprediction of criterion scores when using general scholastic test scores and achievement scores, respectively. These opposing effects tend to offset each other, and it is therefore not surprising that an actual admission score, which often consists of both general scholastic test and achievement scores, is generally unbiased (e.g., [5]).

Conclusion

Much of the criticism of psychological tests is derived from the observation that ethnic and gender groups differ extensively in test performance. Criticism is generally stronger if the groups that show relatively poor performance are also socioeconomically disadvantaged. Much of the polemic concerning test bias confuses several issues: (a) differences in test performance among groups are often regarded, in and of themselves, as an indication of test bias, ignoring performance on the external criterion that the test is designed to predict. Often, groups that perform poorly on tests also tend to perform poorly on measures of the criterion. Furthermore, analysis of the relationship between tests and criteria often reveals similar regression lines for the various groups; and (b) the issue of test bias is often confused with the possibility of bias in the content of some individual items included in a test. Group differences in test performance are attributed to specific item content, and rather than eliminating problematic items in a systematic way (i.e., checking all items for differential performance), this confusion has, in some cases, led to suggestions of a wholesale rejection of reliable and valid test batteries.

Nevertheless, there is a growing recognition in the measurement community that, even though score differences alone are not proof of bias, score differences may not all be related to the measured constructs; and that even valid differences may be misinterpreted to the detriment of the lower-scoring groups when scores are used for important purposes such as selection. The fundamental question remains the degree to which the observed group differences reflect real, underlying psychological processes, and the degree to which group differences simply reflect the way the tests were constructed.

The current edition of the *Standards for Educational and Psychological Testing* [2] indicates that fairness 'is subject to different definitions and interpretations in different social and political circumstances'. With regard to test use, the core meaning of fairness is comparable validity: A fair test is one that yields comparable and valid scores from person to person, group to group, and setting to setting. However, fairness, like validity, is not just a psychometric issue. It also rests on social values. Thus alternative views about its essential features will persist.

Note

1. A construct is a set of knowledge, skills, abilities, or traits a test in intended to measure.

References

[1] Aguinis, H. & Stone-Romero, E.F. (1997). Methodological artifacts in moderated multiple regression and their effects of statistical power, *Journal of Applied Psychology* **82**, 192–206.

[2] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington.

[3] Angoff, W.H. & Ford, S.F. (1973). Item-race interaction on a test of scholastic aptitude, *Journal of Educational Measurement* **10**, 95–106.

[4] Bartlett, C.J., Bobko, P., Mosier, S.B. & Hanna, R. (1978). Testing for fairness with a moderated multiple regression strategy, *Personnel Psychology* **31**, 233–242.

[5] Beller, M. (1994). Psychometric and social issues in admissions to Israeli universities, *Educational Measurement: Issues and Practice* **13**(2), 12–20.

[6] Beller, M. (2001). Admission to higher education in Israel and the role of the psychometric entrance test: educational and political dilemmas, *Assessment in Education* **8**(3), 315–337.

[7] Bond, L. (1993). Comments on the O-Neill and McPeek paper, in *Differential Item Functioning*, P. Holland & H. Wainer, eds, Lawrence Erlbaum, Hillsdale.

[8] Cleary, T.A. (1968). Test bias: prediction of grades of negro and white students in integrated colleges, *Journal of Educational Measurement* **5**, 115–124.

[9] Cole, N.S. (1973). Bias in selection, *Journal of Educational Measurement* **10**, 237–255.

[10] Cole, Nancy. (1981). Bias in testing, *American Psychologist* **36**, 1067–1077.

[11] Cole, N.S. & Moss, P.A. (1989). Bias in test use, in *Educational Measurement*, 3rd Edition, R.L. Linn, ed., Macmillan Publishing, New York, pp. 201–219.

[12] Cole, N.S. & Zieky, M. (2001). The new faces of fairness, *Journal of Educational Measurement* **38**, 369–382.

[13] Cronbach, L.J. (1976). Equity in selection – where psychometrics and political philosophy meet, *Journal of Educational Measurement* **13**, 31–41.

[14] Darlington, R.B. (1971). Another look at "cultural fairness.", *Journal of Educational Measurement* **8**, 71–82.

[15] Dunbar, S. & Novick, M. (1988). On predicting success in training for men and women: examples from Marine Corps clerical specialties, *Journal of Applied Psychology* **73**, 545–550.

[16] Educational Testing Service. (2002). *ETS Standards for Quality and Fairness*, ETS, Princeton.

[17] Flaugher, R.L. (1978). The many definitions of test bias, *American Psychologist* **33**, 671–679.

[18] Gross, A.L. & Su, W. (1975). Defining a "fair" or "unbiased" selection model: a question of utilities, *Journal of Applied Psychology* **60**, 345–351.

[19] Gulliksen, H. & Wilks, S.S. (1950). Regression tests for several samples, *Psychometrika* **15**, 91–114.

[20] Hambleton, R.K. & Jones, R.W. (1994). Comparison of empirical and judgmental procedures for detecting differential item functioning, *Educational Research Quarterly* **18**, 21–36.

[21] Hansen, E.G. & Mislevy, R.J. (2004). Toward a unified validity framework for ensuring access to assessments by individuals with disabilities and English language learners, in *Paper Presented at the Annual Meeting of the National Council on Measurement in Education*, San Diego.

[22] Holland, P. & Wainer, H., eds (1993). *Differential Item Functioning*, Lawrence Erlbaum, Hillsdale.

[23] Houston, W.M. & Novick, M.R. (1987). Race-based differential prediction in air force technical training programs, *Journal of Educational Measurement* **24**, 309–320.

[24] Hunter, J.E., Schmidt, F.L. & Rauschenberger, J. (1984). Methodological and statistical issues in the study of bias in mental testing, in *Perspectives on Mental Testing*, C.R. Reynolds & R.T. Brown, eds, Plenum Press, New York, pp. 41–99.

[25] Lautenschlager, G.J. & Mendoza, J.L. (1986). A step-down hierarchical multiple regression analysis for examining hypotheses about test bias in prediction, *Applied Psychological Measurement* **10**, 133–139.

[26] Linn, R.L. (1973). Fair test use in selection, *Review of Educational Research* **43**, 139–164.

[27] Linn, R.L. (1978). Single group validity, differential validity and differential prediction, *Journal of Applied Psychology* **63**, 507–512.

[28] Linn, R.L. (1984). Selection bias: multiple meanings, *Journal of Educational Measurement* **8**, 71–82.

[29] Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*, Addison Wesley, Reading.

[30] Messick, S. (1989). Validity, in *Educational Measurement*, 3rd Edition, R.L. Linn, ed., Macmillan Publishing, New York, pp. 13–103.

[31] Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory*, 3rd Edition, McGraw-Hill, New York.

[32] Petersen, N.S. & Novick, M.R. (1976). An evaluation of some models for culture fair selection, *Journal of Educational Measurement* **13**, 3–29.

[33] Sackett, P.R., Laczo, R.M. & Lippe, Z.P. (2003). Differential prediction and the use of multiple predictors: the omitted variables problem, *Journal of Applied Psychology* **88**(6), 1046–1056.

[34] Sackett, P.R., Schmitt, N., Ellingson, J.E. & Kabin, M.B. (2001). High stakes testing in employment, credentialing, and higher education: prospects in a post-affirmative action world, *American Psychologist* **56**, 302–318.

[35] Sawyer, R.L., Cole, N.S. & Cole, J.W.L. (1976). Utilities and the issue of fairness in a decision theoretic model for selection, *Journal of Educational Measurement* **13**, 59–76.

[36] Schmidt, F.L., Pearlman, K. & Hunter, J.E. (1981). The validity and fairness of employment and educational tests for Hispanic Americans: a review and analysis, *Personnel Psychology* **33**, 705–724.

[37] Shepard, L.A. (1982). Definitions of bias, in *Handbook of Methods for Detecting Test Bias*, R.A. Berk, ed., Johns Hopkins University Press, Baltimore, pp. 9–30.

[38] Thorndike, R.L. (1971). Concepts of culture fairness, *Journal of Educational Measurement* **8**, 63–70.

[39] Willingham, W.W. & Cole, N.S. (1997). *Gender and Fair Assessment*, Lawrence Erlbaum, Mahwah.

[40] Willingham, W.W., Ragosta, M., Bennett, R.E., Braun, H., Rock, D.A. & Powers, D.E. (1988). *Testing Handicapped People*, Allyn & Bacon, Boston.

[41] Zieky, M. (1993). Practical questions in the use of DIF statistics in test development, in *Differential Item Functioning*, P. Holland & H. Wainer, eds, Lawrence Erlbaum, Hillsdale, pp. 337–347.

[42] Zieky, M., & Carlton, S. (2004). In search of international principles for fairness review of assessments, in *Paper Presented at the IAEA*, Philadelphia.

Further Reading

ETS Fairness Review Guidelines. (2003). Princeton, NJ: ETS. Retrieved February, 10, 2005, from `http://ftp.ets.org/pub/corp/overview.pdf`

MICHAL BELLER

Test Construction

Mark J. Gierl

Volume 4, pp. 2007–2011

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Test Construction

Item response theory (IRT) (*see* **Item Response Theory (IRT) Models for Polytomous Response Data**; **Item Response Theory (IRT) Models for Rating Scale Data**) provides an appealing conceptual framework for test construction due, in large part, to the item and test information function. The *item information function* provides a measure of how much psychometric information an item provides at a given ability level, θ. For dichotomously-scored items calibrated using the three-parameter logistic IRT model, the item information function for item i is calculated as:

$$I_i(\theta) = \frac{D^2 a_i^2 (1 - c_i)}{(c_i + \mathrm{e}^{Da_i(\theta - b_i)})(1 + \mathrm{e}^{-Da_i(\theta - b_i)})^2}, \qquad (1)$$

where $D = 1.7$, a_i is the item discrimination parameter, b_i is the item difficulty parameter, and c_i is the pseudo-chance parameter [9]. (To illustrate key concepts, the three-parameter logistic IRT model is used because it often provides the best fit to data from multiple-choice tests. The item and test information function for select polytomous item response models are described in Chapters 2 through 9 of van der Linden and Hambleton [12]). For any given θ, the amount of information increases with larger values of a_i and decreases with larger values of c_i. That is, item discrimination reflects the amount of information an item provides assuming the pseudo-chance level is relatively small.

The *test information function* is an extension of the item information function. The test information function is the sum of the item information functions at a given θ:

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta), \qquad (2)$$

where $I_i(\theta)$ is the item information and n is the number of test items. This function defines the relationship between ability and the psychometric information provided by a test. The more information each item contributes, the higher the test information function. The test information function is also related directly to measurement precision because the amount of information a test provides at a given θ is inversely proportional to the precision with which ability is estimated at that θ-value, meaning:

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}, \qquad (3)$$

where $SE(\theta)$ is the **standard error** of estimation. $SE(\theta)$ is the standard deviation of the asymptotically normal distribution for those examinees who have a maximum likelihood estimate of θ. The standard error of estimation can be used to compute a **confidence interval** for the corresponding θ-values across the score scale, which promotes a more accurate interpretation of the ability estimates. The standard error of estimation varies across ability level, unlike the standard error of measurement in classical test theory which is constant for all ability levels, because test information frequently varies across ability level.

Basic Approach to Test Construction Using Item and Test Information Functions

Both the item and test information functions are used in test construction. Lord [8] outlined the following four-step procedure, first suggested by Birnbaum [4], for designing a test using calibrated items from an item bank:

Step 1: Decide on the shape desired for the test information function. The desired function is called the *target information function*. Lord [8] called the *target information function* a target information *curve*.

Step 2: Select items with item information functions that will fill the hard-to-fill areas under the target information function.

Step 3: Cumulatively add up the item information functions, obtaining at all times the information function for the part-test composed of items already selected.

Step 4: Continue until the area under the target information function is filled up to a satisfactory approximation.

Steps 3 and 4 are easily understood, but steps 1 and 2 require more explanation.

The shape of the target information function, identified in Step 1, must be specified with the purpose of the test in mind. For example, a norm-referenced test designed to evaluate examinees across a broad

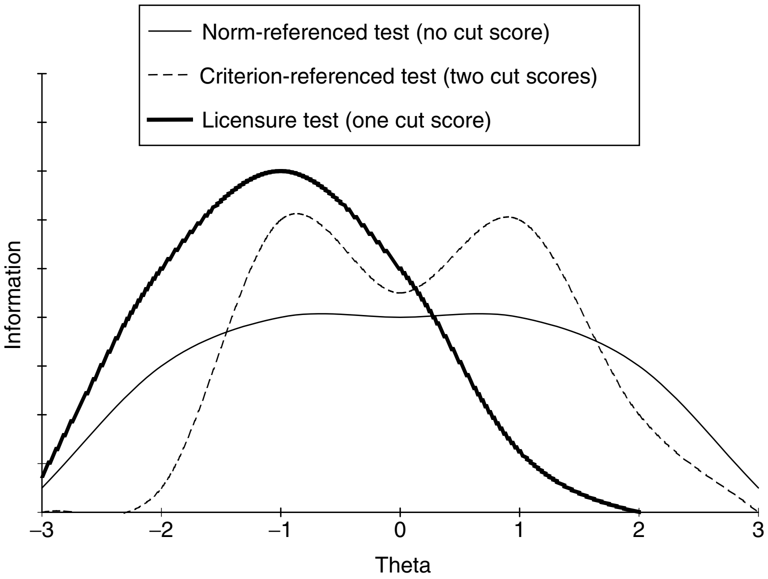


Figure 1 Target information function for three hypothetical tests designed for different purposes

range of ability levels would have a uniform target information function that spans much of the θ-scale. A criterion-referenced test designed to differentiate examinees at an 'acceptable standard' located at $\theta = -1.0$ and a 'standard of excellence' located at $\theta = 1.0$ would, by comparison, have a target information function with two information peaks near the θ cut scores associated with these two standards. A licensure test designed to identify minimally competent examinees, which could be operationally defined as a score above $\theta = -1.0$, would have a target information function with one peak near the θ cut score. These three hypothetical target information functions are illustrated in Figure 1.

Once the target information function is specified in step 1, then item selection in step 2 can be conducted using one of two statistically-based methods. The first item selection method is maximum information. It provides the maximum value of information for an item regardless of its location on the θ-scale. For the three-parameter logistic IRT model, maximum information is calculated as:

$$I_i(\theta)_{\text{MAX}} = \frac{D^2 a_i^2}{8(1-c_i)^2} \times \left[1 - 20c_i - 8c_i^2 + (1+8c_i)^{3/2}\right], \quad (4)$$

where $D = 1.7$, a_i is the discrimination parameter, and c_i is the pseudo-chance parameter [9]. Maximum information is often used in test construction because it provides a method for selecting the most discriminating items from an item bank. However, one problem that can arise when using the most discriminating items is estimation bias: Items with large expected a_i parameters are likely to overestimate their true a_i values because the correlation between the expected and true parameters is less than 1.0. Estimation bias is problematic when developing a test using the most discriminating items [i.e., items with the largest $I_i(\theta)_{\text{MAX}}$ values] because the a_i parameters will be inflated relative to their true values and, as a result, the test information function will be overestimated [6, 7]. This outcome could lead to overconfidence in the accuracy of the examinees' ability estimates, given the test information function at a given θ is inversely proportional to the precision of measurement at that ability level.

The second item selection method is theta maximum. It provides the location on the θ-scale where an item has the most information. For the three-parameter model, theta maximum is calculated as:

$$\theta_i(I)_{\text{MAX}} = b_i + \frac{1}{Da_i} \ln \frac{1 + \sqrt{1+8c_i}}{2}, \quad (5)$$

Table 1 Item parameters, maximum information, and theta maximum values for four example items

Item	a-parameter	b-parameter	c-parameter	$I_i(\Theta)_{\mathrm{MAX}}$	$\theta_i(I)_{\mathrm{MAX}}$
1	0.60	−1.10	0.20	0.18	−1.01
2	0.70	0.14	0.19	0.25	0.25
3	0.64	0.91	0.19	0.21	1.01
4	0.50	−2.87	0.19	0.13	−2.79

where $D = 1.7$, ln is the natural logarithm, a_i is the discrimination parameter, b_i is the difficulty parameter, and c_i is the pseudo-chance parameter [9]. Theta maximum is influenced primarily by the difficulty parameter because it reflects the location (i.e., b_i value) rather than the height (i.e., a_i value) of the item information function. Moreover, the b_i estimates tend to be more accurate than the a_i estimates. Therefore, theta maximum often contains less estimation bias than maximum information thereby producing a more consistent estimate of the test information function [5] which, in turn, yields a more reliable measure of θ.

A simple example helps illustrate the differences between $I_i(\theta)_{\mathrm{MAX}}$ and $\theta_i(I)_{\mathrm{MAX}}$. Table 1 contains the a-, b-, and c-parameter estimates for four items along with their $I_i(\theta)_{\mathrm{MAX}}$ and $\theta_i(I)_{\mathrm{MAX}}$ values. Figure 2 shows the information function for each item. If the goal was to select the most discriminating item from this set, then item 2 would be chosen because it has maximum information [i.e., $I_i(\theta)_{\mathrm{MAX}} = 0.25$]. If, on the other hand, the goal was to select the item that was most discriminating at $\theta = -1.0$, then item 1 would be chosen because it has maximum information around this point on the theta scale [i.e., $\theta_i(I)_{\mathrm{MAX}} = -1.01$]. Notice that item 1 is not the most discriminating item, overall, but it does yield the most information at $\theta = -1.0$, relative to the other three items.

Developments in Test Construction Using Item and Test Information Functions

Rarely are tests created using item selection methods based on statistical criteria alone. Rather, tests must conform to complex specifications that include content, length, format, item type, cognitive levels, reading level, and item exposure, in addition to statistical criteria. These complex specifications, when combined with a large bank of test items, can make the test construction task, as outlined by Lord [8], a formidable one because items must be selected to meet a statistical target information function while at the same time

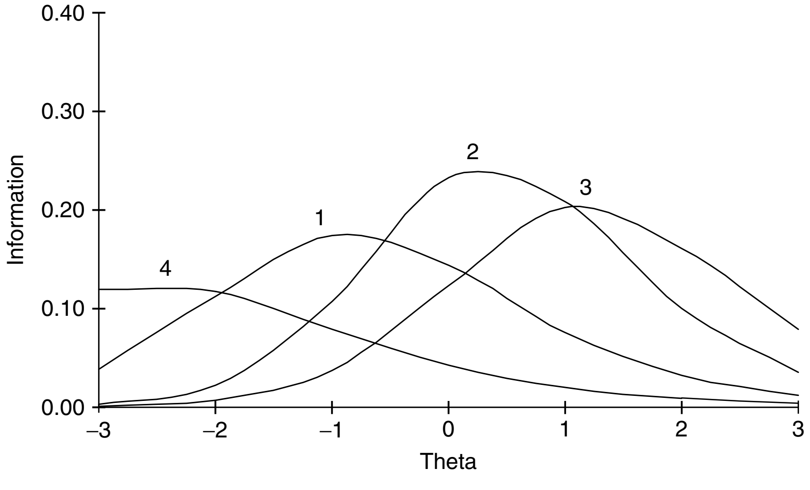


Figure 2 Information functions for four items

satisfying a large number of test specifications (e.g., content) and constraints (e.g., length). Optimal test assembly procedures have been developed to meet this challenge [11]. Optimal test assembly requires the optimization of a test attribute (e.g., target information function) using a unique combination of items from a bank. The goal in optimal test assembly is to identify the set of feasible item combinations in the bank given the test specifications and constraints. The assembly task itself is conducted using an computer algorithm or heuristic. Different approaches have been developed to automate the item selection process in order to optimize the test attribute including 0–1 linear programming [1], heuristic-based test assembly [10], network-flow programming [2], and optimal design [3]. These advanced test construction procedures, which often use item and test information functions, now incorporate the complex specifications and constraints characteristic of modern test design and make use of computer technology. But they also maintain the logic inherent to Lord's [8] procedure demonstrating why the basic four-step approach is fundamental for *understanding* and *appreciating* developments and key principles in optimal test assembly.

References

[1] Adema, J.J., Boekkooi-Timminga, E. & van der Linden, W.J. (1991). Achievement test construction using 0–1 linear programming, *European Journal of Operations Research* **55**, 103–111.

[2] Armstrong, R.D., Jones, D.H. & Wang, Z. (1995). Network optimization in constrained standardized test construction, in *Applications of Management Science: Network applications*, Vol. 8, K. Lawrence & G.R. Reeves, eds, JAI Press, Greenwich, pp. 189–122.

[3] Berger, M.P.F. (1998). Optimal design of tests with dichotomous and polytomous items, *Applied Psychological Measurement* **22**, 248–258.

[4] Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability, in *Statistical Theories of Mental Test Scores*, F.M. Lord & M.R. Novick, eds, Addison-Wesley, Reading, pp. 397–479.

[5] Gierl, M.J., Henderson, D., Jodoin, M. & Klinger, D. (2001). Minimizing the influence of item parameter estimation errors in test development: a comparison of three selection procedures, *Journal of Experimental Education* **69**, 261–279.

[6] Hambleton, R.K. & Jones, R.W. (1994). Item parameter estimation errors and their influence on test information functions, *Applied Measurement in Education* **7**, 171–186.

[7] Hambleton, R.K., Jones, R.W. & Rogers, H.J. (1993). Influence of item parameter estimation errors in test development, *Journal of Educational Measurement* **30**, 143–155.

[8] Lord, F.M. (1977). Practical applications of item characteristic curve theory, *Journal of Educational Measurement* **14**, 117–138.

[9] Lord, F.M. (1980). *Application of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum, Hillsdale.

[10] Luecht, R. (1998). Computer-assisted test assembly using optimization heuristics, *Applied Psychological Measurement* **22**, 224–236.

[11] van der Linden, W.J., ed. (1998). Optimal test assembly [special issue], *Applied Psychological Measurement* **22**(3), 195–211.

[12] van der Linden, W.J. & Hambleton, R.K., eds (1997). *Handbook of Modern Item Response Theory*, Springer, New York.

Mark J. Gierl

Test Construction: Automated

Bernard P. Veldkamp

Volume 4, pp. 2011–2014

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Test Construction: Automated

In large-scale educational measurement, thousands of candidates have to be tested and the results of these tests might have major impact on candidate's lives. Imagine, for example, the consequences of failing on an admission test for someone's professional career. Because of this, great care is given to the process of testing. Careful psychometric planning of high-stakes tests consists of several steps.

First, decisions are made about the kind of abilities that have to be measured, and about the characteristics of the test. These decisions result in a test blueprint. In this blueprint, the test length is specified and some rules for content balancing and/or other characteristics are defined.

The next step is to write items for the test. Instead of writing and pretesting items for each single test form over and over again, the concept of item banking was introduced. Items are written and pretested on a continuous basis and the item characteristics and statistics are stored in an item bank. In most item banks for large-scale testing, the item statistics are calculated on the basis of **item response theory** (IRT) models. In IRT measurement models, item parameters and person parameters are modeled separately [2]. Apart from sampling variation, the item parameters do not depend on the population or on the other items in the test. Because of this property, items that are calibrated can be used for different group of candidates that belong to the same population. (For a more detailed introduction to IRT, see [2] and [3].)

The final step is to select items from the bank and to compose a test. Optimal test construction deals with the problem of how to select those items that meet the specifications in the best way. All kinds of smart decision rules have been developed to select the items. The main objective for most tests is to maximize measurement precision. When IRT models are applied, measurement precision is determined by the amount of information in the test [3]. Birnbaum [1] presented a rather general approach for test construction. His algorithm consisted of the following steps.

1. Decide on the shape of the desired test information function.
2. Select items from the pool with information functions to fill areas under the target-information function.
3. After each item is added to the test, calculate the test information function.
4. Continue selecting items until the test information function approximates the desired shape.

However, Birnbaum's approach does not take all kinds of realistic test characteristics into account. It just focuses the amount of information, that is, the measurement precision of the test. If more and more test characteristics have to be added to the construction problem, the approach becomes hard to adapt. Optimal test construction is a generalization of Birnbaum's approach that does take these realistic characteristics into account [7]. In the mid-1980s, the first methods for optimal test construction were developed. The observation was made that test construction is just one example of a selection problem. Other well-known selection problems are flight-scheduling, work-scheduling, human resource planning, inventory management, and the traveler–salesman problem. In order to solve optimal-test-construction problems, methods to solve these selection problems had to be translated and applied in the area of test development.

Multiple Objectives

Like most real-world selection problems, optimal test construction is a rather complex problem. For example, a test blueprint might prefer the selection of those items that simultaneously maximize the information in the test, minimize the exposure of the items, optimize item bank usage, balance content, minimize the number of gender-biased items, provide most information for diagnostic purposes, and so on. Besides, in test blueprints, some of these objectives are usually favored above others. To make the problem even more complicated, most mathematical programming algorithms can only handle single-objective selection problems instead of multiple-objective ones.

Three different strategies have been developed to solve multiple-objective selection problems [10]. These strategies are classified as methods based on (a) prior, (b) progressive, and (c) posterior weighting of the importance of different objectives. It should be mentioned that the names of the groups of methods might be a little bit confusing because these names have different meaning in terminology of **Bayesian Statistics**.

For prior weighting, an inventory of preferences of objectives is made first. On the basis of this order, a sequence of single-objective selection problems is formulated and solved. For progressive methods, a number of solutions are presented to the test assembler. On the basis of his/her preference, a new single-objective selection problem is formulated, and this process is repeated until an acceptable solution is found. For posterior methods, all different kinds of priority orderings of objectives are taken into account. The solutions that belong to all different priority orderings are presented to the test assembler. From all these solutions, the preferred one is chosen. Methods in all groups do have in common that a multiple-objective problem is reformulated into one or a series of single-objective problems. They just differ in the way the priority of different objectives is implemented in the method.

The Methods

For optimal test construction, it is important that the methods are easy to interpret and easy to handle. Two methods from the class of prior weighting methods are generally used to solve optimal-test-construction problems. When the 0–1 Linear Programming (LP) [8] is applied, a target is formulated for the amount of information, the deviation from the target is minimized, and bounds on the other objectives are imposed. For the weighted deviation method [5], targets are formulated for all objectives, and a weighted deviation from these targets is minimized.

The general 0–1 LP model for optimal construction of a single test can be formulated as:

$$\min y \tag{1}$$

subject to:

$$\left|\sum_{i=1}^{I} I_i(\theta_k)x_i - T(\theta_k)\right| \leq y \quad \forall k, \quad \text{(target information)} \tag{2}$$

$$\sum_{i=1}^{I} a_c \cdot x_i \leq n_c \quad \forall c, \quad \text{(generic constraint)} \tag{3}$$

$$\sum_{i=1}^{I} x_i = n, \quad \text{(total test length)} \tag{4}$$

$$x_i \in \{0, 1\}. \quad \text{(decision variables)} \tag{5}$$

In (1) and (2), the deviation between the target-information curve and the information in the test is minimized for several points θ_k, $k = 1, \ldots, K$, on the ability scale. The generic constraint (3) denotes the possibility to include specifications for all kinds of item and test characteristics in the model. Constraints can be included to deal with item content, item type, the word count of the item, the time needed to answer the item, or even constraints to deal with inter-item dependencies. (For an overview of test-construction models, see [7].) The test length is defined in (4), and (5) is a technical constraint that makes sure that items are either in the test ($x_i = 1$) or not ($x_i = 0$).

When the weighted deviation model (WDM) is applied, the test blueprint is used to formulate goals for all item and test characteristics. These goals are seen as desirable properties. During the test-construction process, items that minimize the deviations from these goals are selected. If it is possible, the goals will be met. Otherwise, the deviations of the goal are as small as possible. For some characteristics, it is more important that the goals are met than for others. By assigning different weights to the characteristics, the impacts of the deviations differ. In this way, an attempt is being made to guarantee that the most important goals will be met when the item bank contains enough good-quality items.

The WDM model can be formulated as:

$$\min \sum_j w_j d_j \quad \text{(minimize weighted deviation)} \tag{6}$$

subject to:

$$\left|\sum_{i=1}^{I} I_i(\theta_k)x_i - T(\theta_k)\right| \leq d_k \quad \forall k, \quad \text{(target information)} \tag{7}$$

$$\sum_{i=1}^{I} a_c \cdot x_i - n_c \leq d_c \quad \forall c, \quad \text{(generic constraint)} \tag{8}$$

$$x_i \in \{0, 1\} \quad d_j \geq 0. \quad \text{(decision variables)} \tag{9}$$

Where the variables d_j denote the deviations and w_j denotes the weight of deviation j.

Both the 0–1 LP model and the WDM have been successfully applied to solve the optimal-test-construction problems. Which method to prefer

Table 1 Overview of different test-construction models

Parallel test forms	For security reasons, or when a test is administered in several testing windows, parallel tests have to be assembled. Several definitions of parallelness exist, but the concept of weakly parallel tests is most often applied. This means that the same set of constraints is met by the tests and the test information functions are identical. To assemble parallel tests, item variables x_i are replaced by variables $x_{ij} \in \{0, 1\}$ that indicate whether item i is selected for test j. The number of decision variables grows linearly, but the complexity of the problem grows exponentially.
Item sets	Some items in the pool may be grouped around a common stimulus. When a test is administered, all items for this stimulus have to be presented consecutively and sometimes a minimum or maximum number of items from this set need to be selected. To assemble a test with item sets, a model can be extended with decision variables $z_s \in \{0, 1\}$ that denote whether set s is selected for the test.
Classical test construction	Besides IRT, classical test theory (CTT) is still often applied to assemble tests. One of the drawbacks of CTT is that classical item parameters depend on the population and the other items in the test. When the assumption can be made that the population of the examinees hardly changes, optimal test construction might also be possible for classical test forms. A common objective function is to optimize Cronbach's alpha, a lower bound to the reliability of the test.
Test for multiple abilities	For many tests, several abilities are involved in answering the items correctly. In some cases, all abilities are intentional, but in other cases, some of them are considered nuisance. When only one dominant ability is present, the others might be ignored; otherwise, the easiest approach is to optimize Kullback–Leibler information instead of Fisher Information, because its multidimensional form still is a linear function of the items in the test.
Computerized adaptive testing	In Computerized adaptive testing, the items are selected sequentially during test administration. Difficulty of the items is adapted to the estimated ability level of the examinee. For each new item, a test assembly problem has to be solved. Since most test characteristics are defined at test level and item selection happens at item level, somehow these characteristics at test level have to be built in the lower level model of selecting the next item.
Multi-stage testing	In multi-stage testing, a test consists of a network of smaller tests. After finishing the first small test, an examinee is directed to the next test on the basis of his/her ability level. In this way, the difficulty level of the small tests is adapted to the estimated ability level of the examinee. For assembling such a network of small tests, sets of small tests that only differ in difficulty level have to be first assembled. Then these tests have to be assigned to a stage in the network. Optimal multi-stage test assembly is very complicated because all routes through the network have to result in tests that meet the test characteristics.

depends on the way the item and test characteristics are described in the test blueprint. When a very strict formulation of characteristics is used in the test blueprint, the 0–1 LP model seems preferable, because it guarantees that the resulting test meets all constraints. The WDM model is more flexible. It also gives the test assembler more opportunities to prioritize some characteristics above others.

Several Test-Construction Models

In equations 1 to 5 and 6 to 9, general models are given for optimal construction of a single test. Many different models are available for different test-construction problems [7]. An overview of different models is given in Table 1. In this table, the special features of the models are described.

Current Developments

In the past twenty years, many optimal-test-construction models have been developed and algorithms for test construction have been fine-tuned. Although tests are assembled to be optimal, this does not imply that their quality is perfect. An upper bound to the quality of the test is defined by the quality of items in the pool. The next step in optimization, therefore, is to optimize composition and usage of the item pool.

Optimum item pool design [9] focuses on item pool development in management. An optimal blueprint is developed to guide the item writing process.

This blueprint is not only based on test characteristics, but also takes features of the item selection algorithm into account. The goal of these design methods is to develop an item pool with minimal costs and optimal item usage.

Besides, exposure-control methods have been developed that can be used to optimize the usage of item banks. These methods are applicable for testing programs that use an item pool over a period of time. In optimal test construction, the best items are selected for a test. As a consequence, a small portion of the items in the pool is selected for the majority of the tests. This problem became most obvious when computerized adaptive testing was introduced. To deal with problems of unequal usage of items in the pool, several exposure-control methods are available, both to deal with overexposure of the popular items (e.g., Sympson–Hetter method [6]) or to deal with underexposure (e.g., progressive method [4]).

References

[1] Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability, in *Statistical Theories of Mental Test Scores*, F.M. Lord & M.R. Novick, eds, Addison-Wesley, Reading, pp. 397–479.

[2] Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*, Kluwer Academic Publishers, Boston.

[3] Lord, F.M. (1980). *Application of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum Associates, Hillsdale.

[4] Revuelta, J. & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing, *Journal of Educational Measurement* **35**, 311–327.

[5] Stocking, M.L. & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing, *Applied Psychological Measurement* **17**, 277–292.

[6] Sympson, J.B. & Hetter, R.D. (1985). Controlling item-exposure rates in computerized adaptive testing, *Proceedings of the 27th Annual Meeting of the Military Testing Association*, Navy Personnel Research and Development Center, San Diego, pp. 973–977.

[7] van der Linden, W.J. (2005). *Optimal Test Assembly*, Springer-Verlag, New York.

[8] van der Linden, W.J. & Boekkooi-Timminga, E. (1989). A maximin model for test assembly with practical constraints, *Psychometrika* **54**, 237–247.

[9] van der Linden, W.J., Veldkamp, B.P. & Reese, L.M. (2000). An integer programming approach to item bank design, *Applied Psychological Measurement* **24**, 139–150.

[10] Veldkamp, B.P. (1999). Multiple objectives test assembly problems, *Journal of Educational Measurement* **36**, 253–266.

BERNARD P. VELDKAMP

Test Dimensionality: Assessment of

Marc E. Gessaroli and Andre F. de Champlain

Volume 4, pp. 2014–2021

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Test Dimensionality: Assessment of

Test scores are viewed as representations of a theoretical construct, or set of constructs, hypothesized to underlie examinee performances. A test score is typically interpreted as the manifestation of one or more latent traits (*see* **Latent Variable**) that the test is designed to measure. An examination of the *structural aspect* of construct validity [33] involves an assessment of test dimensionality. That is, the degree to which the dimensional structure of the response matrix is consistent with the domain(s) hypothesized to underlie performance must be examined. Assessing the dimensional structure of a response matrix therefore constitutes a central psychometric activity in that the primary concern of any organization involved in assessment is to ensure that scores and decisions reported to examinees are accurate and valid reflections of the proficiencies that were intended to be targeted by the examination. As well, understanding the dimensional structure of a test is important for other psychometric activities such as equating and calibration.

The term *test dimensionality* is, in some sense, misleading because it does not refer only to the particular set of items comprising a test. Rather, the dimensionality of a test is a function of the interaction between the set of items and the group of examinees responding to those items. An examination of the responses of examinee populations responding to the same set of items might quite possibly result in different conclusions regarding the dimensions underlying the trait. The differences among different examinee populations on variables that might be important in responding to the test items such as curriculum, prior experience, age, and so on, must be carefully considered before any generalizations are made.

What is meant by test dimensionality has been debated in the literature and is still unclear. Early work considered test dimensionality to be related to test homogeneity and reliability. Current definitions relate test dimensionality to some form of the principle of local independence (*LI*).

Definitions of Dimensionality

Dimensionality Based on Local Independence

The principle of local independence is achieved when for fixed levels of a vector of latent traits ($\boldsymbol{\theta}$) the responses for an examinee are statistically independent. More formally, if the item responses for p items are represented by random variables $\mathbf{Y} = Y_1, Y_2, \ldots, Y_p$, then the responses to the p are locally independent when, for m latent traits $\boldsymbol{\Theta}$ having fixed values $\boldsymbol{\theta}$,

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_p = y_p|\boldsymbol{\Theta} = \boldsymbol{\theta}) = \prod_{j=1}^{p} P(Y_j = y_j|\boldsymbol{\Theta} = \boldsymbol{\theta}). \quad (1)$$

The definition of local independence in (1) involves all 2^p higher-order interactions among the items and is known as *strong* local independence (*SLI*). A less stringent version of local independence considers only the second-order interactions among the items. Here, local independence holds if for all item responses $Y_j = y_j$ and $Y_k = y_k$ $(j \neq k; j, k = 1, \ldots p)$,

$$P(Y_j = y_j, Y_k = y_k|\boldsymbol{\Theta} = \boldsymbol{\theta}) = P(Y_j = y_j|\boldsymbol{\Theta} = \boldsymbol{\theta}) \times P(Y_k = y_k|\boldsymbol{\Theta} = \boldsymbol{\theta}). \quad (2)$$

In practice, (2) is usually assessed by

$$\sum_{j=1}^{p} \sum_{\substack{k=1 \\ j \neq k}}^{p} \operatorname{cov}(Y_j, Y_k|\boldsymbol{\Theta} = \boldsymbol{\theta}) = 0. \quad (3)$$

This definition of LI, known as *weak* local independence (*WLI*), only requires that the covariances between all pairs of items be zero for fixed values of the latent traits.

The existence of *SLI* implies *WLI*. As well, under multivariate normality, *SLI* holds if *WLI* is valid [29]. The very little research comparing analyses on the basis of the two forms of *LI* has found no practical differences between them [24].

McDonald [27] and McDonald and Mok [32] assert that the principle of local independence provides the definition of a latent trait. More formally, they state that $\Theta_1, \Theta_2, \ldots, \Theta_m$ are the m latent traits underlying the item responses if and only if, for $\boldsymbol{\Theta} = \theta$, *SLI* holds. Now, in practice, this definition has been relaxed to require *WLI* instead of *SLI.* Using this definition, the number of dimensions underlying test responses (d_{LI}) is equal to m.

Definitions of latent traits and test dimensionality using either *SLI* or *WLI* are based on precise theoretical requirements of statistical independence. As such, these two principles are reflective of the more general principle of *strict* local independence.

This is a mathematical definition of test dimensionality based on latent traits that, some have argued [11, 12], sometimes fails to capture all the dependencies among the item responses. Furthermore, it is possible to satisfy the mathematical definition yet not fully account for all of the psychological variables affecting the item responses. For example, suppose examinees, in a test of reading comprehension, were asked to read a passage relating to Greek Theatre and then answer a series of questions relating to the passage. It is quite possible that a single trait would account for the mathematical dependencies among the items. However, although local independence is satisfied by a single mathematical latent trait, there might be two psychological traits that influence the responses to the items. An examinee's response to an item might be influenced by some level of proficiency related to general reading comprehension as well as some specific knowledge of Greek Theatre. These two traits are confounded in the single latent trait that results in local independence [10].

Other researchers (e.g., [38, 42, 43]) state that the mathematical definition based on strict local independence is too stringent because it considers both major and minor dimensions. They argue that minor dimensions, although present mathematically, do not have an important influence on the item responses and probably are unimportant in a description of the dimensionality underlying the item responses [38]. This notion is the basis for a definition of *essential dimensionality* described next.

Dimensionality Based on Essential Independence

Using the same notation as above, Stout [43] states that a response vector, $\mathbf{y}$, of p items is said to be *essentially independent* (*EI*) with regard to the latent variables $\mathbf{\Theta}$, if, for every pair of responses $y_j, y_k (j, k = 1, \ldots, p)$,

$$\frac{2}{p(p-1)} \sum_{1 \le j < k \le p} |\mathrm{cov}(y_j, y_k)|\mathbf{\Theta} = \theta| \rightarrow 0 \; as \; p \rightarrow \infty. \tag{4}$$

EI is similar to *WLI* in that it only considers pairwise dependencies among the items. However, unlike *WLI, EI* does not require that conditional item covariances be equal to zero. Rather, *EI* requires that the mean $|\mathrm{cov}(y_j, y_k)|\mathbf{\Theta} = \theta|$ across item pairs is small (approaches 0) as test length p increases. As a result, *EI* only considers dominant dimensions while *WLI* in theory requires all dimensions, however minor, to satisfy (2) or (3). The *essential dimensionality* (d_{EI}) can subsequently be defined as the smallest number of dimensions required for *EI* to hold. From the above definitions, it is clear that $d_{\mathrm{EI}} \le d_{\mathrm{LI}}$. In the instance where $d_{\mathrm{EI}} = 1$, the item response matrix is said to be *essentially unidimensional.*

Methods to Assess Dimensionality

Owing to the increasing popularity of item response theory (IRT) (*see* **Item Response Theory (IRT) Models for Polytomous Response Data**; **Item Response Theory (IRT) Models for Rating Scale Data**), most early work in the assessment of test dimensionality focused specifically on the assessment of unidimensionality. Hattie [15, 16], in a comprehensive review of such techniques, identified indices that purported to assess whether a test was unidimensional. He found that most of the indices used (e.g., those based on reliability, homogeneity, principal components) were *ad hoc* in nature and were not based on any formal definition of dimensionality. There appeared to be confusion regarding the concepts of homogeneity, internal consistency, and dimensionality. For example, although the degree of reliability of a test was thought to be related to its dimensionality (i.e., higher reliability was indicative of a more unidimensional test), Green, Lissitz, and Mulaik [13] showed that it is possible for coefficient alpha to be large with a five-dimensional test.

Today's methods to assess dimensionality are more theoretically sound because they are based on either local independence or essential independence [8]. The methods, in some way, are related to the principles of local and essential independence because they provide indices measuring the degree to which the data are not conditionally independent. Some methods provide global indices while others assess dimensionality by assessing the amount of conditional dependence present between pairs of items.

Methods Based on Local Independence

Factor analysis is the regression of an observed variable on one or more unobserved variables. In dichotomous item scoring, Y is an observed binary variable having a value of 0 for an incorrect response and 1 for a correct response. The unobserved variable(s) are the latent traits needed to correctly answer these items.

Original factor analytic work described a linear relationship between the binary responses and the latent traits, and parameters were estimated by fitting a phi correlation matrix. There are two important weaknesses to this approach. First, predicted values of the dependent variable may be either greater than 1 or less than 0 where, in fact, the item scores are bounded by 0 and 1. This model misspecification leads, in some cases, to additional spurious factors described initially as 'difficulty factors.' McDonald and Ahlawat [31] clarified the issue by suggesting that these factors are attributable to the misfit of the model at the upper and lower extremes of the item response function where the relationship between the trait and item responses is nonlinear.

The fitting of a **tetrachoric correlation** matrix was suggested as a possible alternative to the fitting of the phi correlation matrix. While having a sound theoretical basis, there are practical issues associated with the calculation of a tetrachoric correlation matrix that do not allow for the general recommendation of this approach. First, non-Gramian matrices and Heywood cases have been reported. Also, the fitting of tetrachoric matrices has been found to yield poor results in the presence of guessing in the item responses. This is not surprising because the underlying latent distribution of each item assumed in calculating the tetrachoric correlations are the equivalent of a two-parameter normal ogive function, not the three-parameter function that would include a parameter for guessing [25].

Over the past few decades, a number of weighted least-squares estimation methods have been proposed within the context of **structural equation modeling** for use in linear **confirmatory factor analytic** models with dichotomous variables [3, 5, 21, 35]. The fit of a given m-dimensional factor model is typically assessed using a robust chi-square statistic. One limitation of using such methods for assessing dimensionality is that the recommended sample size is typically prohibitive in practical applications. Recently, diagonally weighted least-squares estimation methods [36] implemented in the software packages Mplus [37] and PRELIS/LISREL [22] have proven more useful with smaller sample sizes (*see* **Structural Equation Modeling: Software**). However, neither package currently offers an adjustment for guessing or smoothing when analyzing a tetrachoric correlation matrix. More research is needed to assess the utility of these approaches for assessing dimensionality.

McDonald [28] showed that the nonlinear relationship between the probability of correctly answering an item and the underlying trait(s) could be approximated by a third-order polynomial model. He classified this as a model that is linear in its coefficients but nonlinear in the traits [30]. McDonald's polynomial approximation is implemented in the computer program NOHARM [7]. Parameters are estimated by fitting the joint proportions among the items by unweighted least squares (*see* **Least Squares Estimation**). McDonald states that the magnitudes and patterns of the residual joint proportions (i.e., the difference between the observed and predicted values of the off-diagonal components of the matrix of joint proportions) provides evidence of the degree to which the fitted model achieves local independence. The sum or mean of the absolute residuals have been found to be generally related to the number of dimensions underlying the item response matrix [14, 16, 24]. As well as providing the residual joint proportions, NOHARM currently provides Tanaka's goodness of fit index [47] as a measure of model fit. This index has been used in structural equation modeling but little research has been carried out as to its utility in this context.

Another statistic using the residual joint proportions from NOHARM is the Approximate χ^2 statistic [9]. The Approximate χ^2 statistic was proposed as an *ad hoc* method to be used to aid practitioners when other, more theoretically sound procedures, are inappropriate. Although the authors acknowledge limitations of the statistic, the Approximate χ^2 statistic has performed quite well in identifying the correct dimensional structure with simulated data based on compensatory item response models [9, 48]. As with any significance test, the null hypothesis will always be rejected with large enough samples and caution should always be used when interpreting the results.

The common item responses models based on the normal ogive or logistic functions have been shown to be special cases of a more general nonlinear factor analytic model [28, 46]. McDonald [30] places these IRT models (*see* **Item Response Theory (IRT) Models for Polytomous Response Data**; **Item Response Theory (IRT) Models for Rating Scale Data**) into a third classification of factor analytic models – models that are nonlinear in both their coefficients and traits.

The full-information factor analysis (FIFA) methods proposed by Bock, Gibbons, and Muraki [1] and implemented in TESTFACT 4.0 [2] yield parameter estimates and fit statistics for a m-dimensional item response model. FIFA is theoretically sound because it uses information from the 2^p response patterns instead of only pairwise relationships to estimate the model parameters and thus uses the strong principle of local independence.

A measure of model misfit in TESTFACT is given by the likelihood ratio χ^2 test. Mislevy [34] indicates that this statistic might poorly approximate the theoretical chi-square distribution due to the incomplete cells in the 2^p response-pattern table. The authors of TESTFACT suggest the use of the difference between two likelihood ratio χ^2 statistics from nested models to test whether a higher dimensional model yields a significantly better fit to the data.

Little research has been carried out to investigate the performance of these statistics. However, Knol and Berger [24] found that the chi-square difference test was unable to correctly identify the number of dimensions in simulated data. De Champlain and Gessaroli [6], in a study investigating the effects of small samples and short test lengths, reported inflated rejections of the assumption of unidimensionality for unidimensionally simulated data. More research under a wider variety of conditions is warranted.

Other methods have been specifically developed to assess the amount of conditional dependence present in pairs of items assuming a single trait underlying the responses. The Q_3 statistic is a measure of conditional correlation between two items [49]. Chen and Thissen [4] assessed the performance of four **measures of association** – (a) Pearson χ^2, (b) Likelihood Ratio G^2, (c) Standardized ϕ Coefficient Difference, and (d) Standardized Log-Odds Ratio Difference (τ) – to test for conditional association between binary items in their two-way joint responses. These statistics are computed by comparing the cells in the 2×2 tables of observed and expected joint frequencies where the expected joint frequencies are obtained from an IRT model. Much more detail is available in Chen and Thissen [4]. Some results of investigations of the performance of these indices can be found in Chen and Thissen [4], Yen [49], and Zwick [51].

The utility of parametric models in assessing test dimensionality is dependent upon the item response model specified and the distributional assumptions of the latent trait(s). Incorrect assumptions and/or model specification might lead to inaccurate conclusions regarding the test dimensionality.

Holland and Rosenbaum [19] discuss an approach to assess the conditional independence among pairs of items without assuming an underlying item response model. Their procedure is based on the work of Holland [18] and Rosenbaum [39]. Conditional association for each pair of items is tested with the **Mantel–Haenszel statistic** [26]. Hattie [16] suggested that this approach was promising because of its link to local independence. However, relatively little research has been carried out studying the effectiveness of the procedure. See Ip [20] for a discussion of some issues.

Ip [20] states that little research has been given to controlling the Type I **error rate** associated with multiple tests inherent in testing the many item pairs for local independence. He proposes a method using a step-down Mantel–Haenzel approach to control for family-wise error rates. Ip suggests that this method might be used to provide more detailed information after a global test has concluded that a response matrix is not unidimensional.

Methods Based on Essential Independence

Over the past 15 to 20 years, Stout and his colleagues have carried out considerable work developing nonparametric methods based on the principle of essential independence described earlier. This work, while still being developed and refined, has resulted in three primary methods that aim to identify the nature and amount of multidimensionality in a test: (a) DIMTEST [42, 45] tests the null hypothesis that a response matrix is essentially unidimensional; (b) DETECT [23, 50] provides an index quantifying the amount of dimensionality present given a particular clustering or partitioning of the items in a test; and, (c) HCA/CCPROX [40], using agglomerative hierarchical cluster analysis, attempts to identify clusters of items that reflect the true (approximate) dimensionality underlying a set of item responses.

DIMTEST is the most popular and widely used procedure. Quite generally, DIMTEST tests the null hypothesis of essential unidimensionality by comparing the magnitudes of the conditional covariances of a subset of items, AT1, chosen to be (a) as unidimensional as possible, and (b) as dimensionally dissimilar to the other items on the test, with the conditional covariances of another subset of items on the test (PT). In the original version of DIMTEST, another subset of the test items, AT2, having the same number of items and similar item difficulties as AT1, is used to adjust the T-statistic for preasymptotic bias. Nandakumar [38] provides a more detailed explanation of DIMTEST.

DIMTEST generally has performed well with simulated data; it has high power in rejecting unidimensionality with multidimensional data and maintains a Type I error rate close to nominal values when the simulated data are unidimensional. However, Type I error rates are inflated in cases where AT1 has high-item discriminations relative to the other items on the test. In this case, AT2 often fails to adequately correct the T-statistic. As well, DIMTEST is not recommended for short test lengths (<20 items) because the necessary partitioning of items into AT1, AT2, and PT results in too few items in each subgroup for a proper analysis. Seraphine [41] concluded that DIMTEST had poorer performance in simulation studies where the data were based on noncompensatory or partially noncompensatory models. A comprehensive investigation of the performance of DIMTEST is provided by Hattie, Krakowski, Rogers, and Swaminathan [17].

Research into DIMTEST (and DETECT and HCA/CCPROX) is continuing. Recent research has suggested the use of nonparametric IRT parametric bootstrap method to correct for the preasymptotic bias in the T-statistic in DIMTEST. A method to use DIMTEST, HCA/CCPROX, and DETECT together to investigate multidimensionality is outlined by Stout, Habing, Douglas, Kim, Roussos, and Zhang [44].

Other Issues

The assessment of test dimensionality is complex. Disagreement exists in several areas such as whether dimensions should be defined and interpreted using psychological or statistical criteria, whether dimensions should be interpreted as only being examinee characteristics that influence the responses or whether item characteristics (such as item dependencies within a testlet) should be also be considered, and whether only dominant dimensions should be included in the definition and interpretation. Added to this are the different approaches to dimensionality assessment, including methods to investigate whether a test is unidimensional, others that attempt to find the number of dimensions, and yet others whose purpose is to determine the dimensional structure underlying the responses.

Although not discussed here, methods such as those based on factor analysis enable a more detailed analysis of the dimensional structure including the relative strengths of each dimension and the relative strengths of each dimension on individual items.

Essential versus Local Independence in Practice

Essential independence as defined by Stout [42, 43] requires that the conditional covariances among pairs of items are approximately zero. In this way, it uses the *weak* principle of local independence as its basis but, in theory, differs from *WLI* because it does not require that the conditional covariances be equal to zero. In practice, both methods evaluate whether the conditional covariances are sufficiently small to conclude that the proposed dimensional structure explains the item covariances. No distinction is made as to whether the conditional covariances are due to sampling error or the effect of additional or minor nuisance dimensions. In practice, the conclusions reached by the approaches associated with the two principles are usually quite similar. Further discussions are found in Gessaroli [8] and McDonald and Mok [32].

Dimensional Structure

The nature of the dimensionality underlying tests can be complex. *Simple structure* occurs when each trait influences independent clusters of items on the test. That is, each item is related to only one trait. In simulation studies, the amount of multidimensionality is often manipulated by either increasing or decreasing the magnitudes of the intertrait correlations. Higher intertrait correlations result in less multidimensionality, in the sense that the magnitudes of the conditional are smaller when fitting a unidimensional model. In the extreme case, perfectly correlated traits result in a unidimensional model. Items may also be related to more than one trait. This, combined with correlations among the traits, leads to complex multidimensional structures that are often difficult to identify with the dimensionality assessment methods described above. In general, the dimensionality assessment methods function best when the data have *simple structure*.

Confirmatory Analyses

The dimensionality assessment methods described earlier are exploratory. However, knowledge of the examinees, curriculum, test blueprint, and so on might lead a researcher to ask whether a particular model fits the data. In such cases, the researcher has a priori expectations about the traits underlying the performance as well as the relationship of the items to these traits. For example, an expectation of a model having *simple structure* might be tested. Factor analysis provides a logical basis for performing these confirmatory analyses. As of now, NOHARM provides the only analytical program for confirmatory factor analysis. However, TESTFACT 4.0 estimates parameters for a bifactor model. DIMTEST can be used in a confirmatory mode by allowing the researcher to choose items to be placed in the AT1 subtest. Methods to use DIMTEST, HCA/CCPROX, and DETECT in a confirmatory way are outlined in Stout, et al. [44].

Polytomous Item Responses

The discussion so far has been limited to the dimensionality assessment of binary-coded item responses. The relatively little work carried out for polytomous data has largely been extensions of existing methods for the binary case. Poly-DIMTEST is an extension of DIMTEST and can be used to test the assumption of essential unidimensionality. The structural equation modeling literature provides many extensions of binary to polytomous data [21]. Ip [20] discusses an extension of the assessment of conditional association between pairs of items to polytomous data.

References

[1] Bock, R.D., Gibbons, R. & Muraki, E. (1988). Full information item factor analysis, *Applied Psychological Measurement* **12**, 261–280.

[2] Bock, R.D., Gibbons, R., Schilling, S.G., Muraki, E., Wilson, D.T. & Wood, R. (2003). *TESTFACT 4.0 [Computer Software]*, Scientific Software International, Lincolnwood.

[3] Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures, *British Journal of Mathematical and Statistical Psychology* **37**, 62–83.

[4] Chen, W. & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory, *Journal of Educational and Behavioral Statistics* **22**, 265–289.

[5] Christofferson, A. (1975). Factor analysis of dichotomized variables, *Psychometrika* **40**, 5–32.

[6] De Champlain, A.F. & Gessaroli, M.E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths, *Applied Measurement in Education* **11**, 231–253.

[7] Fraser, C. & McDonald., R.P. (2003). NOHARM Version 3.0: A Windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory. [Computer software and manual] `http://www.niagarac.on.ca/~cfraser/download/noharmdl.html`.

[8] Gessaroli, M.E. (1994). The assessment of dimensionality via local and essential independence. A comparison in theory and practice, in *Modern Theories of Measurement: Problems and Issues*, D. Laveault, B.D. Zumbo, M.E. Gessaroli & M.W. Boss, eds, University of Ottawa, Ottawa, pp. 93–104.

[9] Gessaroli, M.E. & De Champlain, A.F. (1996). Using an approximate χ^2 statistic to test the number of dimensions underlying the responses to a set of items, *Journal of Educational Measurement* **33**, 157–179.

[10] Gessaroli, M.E. & Folske, J.C. (2002). Generalizing the reliability of tests comprised of testlets, *International Journal of Testing* **2**, 277–295.

[11] Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models, *British Journal of Mathematical and Statistical Psychology* **33**, 234–246.

[12] Goldstein, H. & Wood, R. (1989). Five decades of item response modelling, *British Journal of Mathematical and Statistical Psychology* **42**, 139–167.

[13] Green, S.B., Lissitz, R.W. & Mulaik, S.A. (1977). Limitations of coefficient alpha as an index of test unidimensionality, *Educational and Psychological Measurement* **37**, 827–838.

[14] Hambleton, R.K. & Rovinelli, R. (1986). Assessing the dimensionality of a set of test items, *Applied Psychological Measurement* **10**, 287–302.

[15] Hattie, J. (1984). An empirical study of various indices for determining unidimensionality, *Multivariate Behavioral Research* **19**, 49–78.

[16] Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items, *Applied Psychological Measurement* **9**, 139–164.

[17] Hattie, J., Krakowski, K., Rogers, H.J. & Swaminathan, H. (1996). An assessment of Stout's index of essential dimensionality, *Applied Psychological Measurement* **20**, 1–14.

[18] Holland, P.W. (1981). When are item response models consistent with observed data? *Psychometrika* **46**, 79–92.

[19] Holland, P.W. & Rosenbaum, P.R. (1986). Conditional association and unidimensionality in monotone latent variable models, *The Annals of Statistics* **14**, 1523–1543.

[20] Ip, E.H. (2001). Testing for local dependency in dichotomous and polytomous item response models, *Psychometrika* **66**, 109–132.

[21] Jöreskog, K.G. (1990). New developments in LISREL: analysis of ordinal variables using polychoric correlations and weighted least squares, *Quality and Quantity* **24**, 387–404.

[22] Jöreskog, K.G. & Sörbom, D. (2003). *LISREL 8.54 [Computer Software]*, Scientific Software International, Lincolnwood.

[23] Kim, H.R. (1994). New techniques for the dimensionality assessment of standardized test data, Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Department of Statistics.

[24] Knol, D.L. & Berger, M.P.F. (1991). Empirical comparison between factor analysis and multidimensional item response models, *Multivariate Behavioral Research* **26**, 456–477.

[25] Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading.

[26] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the retrospective study of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[27] McDonald, R.P. (1962). A note on the derivation of the latent class model, *Psychometrika* **27**, 203–206.

[28] McDonald, R.P. (1967). *Nonlinear Factor Analysis*, Psychometric Monographs No. 15, The Psychometric Society.

[29] McDonald, R.P. (1981). The dimensionality of tests and items, *British Journal of Mathematical and Statistical Psychology* **34**, 100–117.

[30] McDonald, R.P. (1982). Linear versus nonlinear models in item response theory, *Applied Psychological Measurement* **6**, 379–396.

[31] McDonald, R.P. & Ahlawat, K.S. (1974). Difficulty factors in binary data, *British Journal of Mathematical and Statistical Psychology* **27**, 82–99.

[32] McDonald, R.P. & Mok, M.M.-C. (1995). Goodness of fit in item response models, *Multivariate Behavioral Research* **30**, 23–40.

[33] Messick, S. (1995). Standards of validity and the validity of standards in performance assessment, *Educational Measurement: Issues and Practice* **14**, 5–8.

[34] Mislevy, R.J. (1986). Recent developments in the factor analysis of categorical variables, *Journal of Educational Statistics* **11**, 3–31.

[35] Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators, *Psychometrika* **49**, 115–132.

[36] Muthén, B. (1993). Goodness of fit with categorical and other non-normal variables, in *Testing Structural Equation Models*, K.A. Bollen & J.S. Long, eds, Sage, Newbury Park, pp. 205–243.

[37] Muthén, B. & Muthén, L. (2004). *Mplus Version 3 [Computer Software]*, Muthén & Muthén, Los Angeles.

[38] Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality, *Journal of Educational Measurement* **28**, 99–117.

[39] Rosenbaum, P. (1984). Testing the conditional independence and monotonicity assumptions of item response theory, *Psychometrika* **49**, 425–435.

[40] Roussos, L.A., Stout, W.F. & Marden, J.I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality, *Journal of Educational Measurement* **35**, 1–30.

[41] Seraphine, A.E. (2000). The performance of DIMTEST when latent trait and item difficulty distributions differ, *Applied Psychological Measurement* **42**, 82–94.

[42] Stout, W.F. (1987). A nonparametric approach for assessing latent trait unidimensionality, *Psychometrika* **52**, 589–617.

[43] Stout, W.F. (1990). A new item response theory modelling approach with applications to unidimensionality assessment and ability estimation, *Psychometrika* **55**, 293–325.

[44] Stout, W.F., Habing, B., Douglas, J., Kim, H.R., Roussos, L. & Zhang, J. (1996). Conditional covariance-based multidimensionality assessment, *Applied Psychological Measurement* **20**, 331–354.

[45] Stout, W.F., Nandakumar, R., Junker, B., Chang, H. & Steidinger, D. (1992). DIMTEST: a Fortran program for assessing dimensionality of binary items, *Applied Psychological Measurement* **16**, 236.

[46] Takane, Y. & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables, *Psychometrika* **52**, 393–408.

[47] Tanaka, J.S. (1993). Multifaceted conceptions of fit in structural equation models, in *Testing Structural Equation Models*, K.A. Bollen & J.S. Long, eds, Sage, Newbury Park, pp. 11–39.

[48] Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items, *Applied Psychological Measurement* **27**, 159–203.

[49] Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model, *Applied Psychological Measurement* **30**, 187–213.

[50] Zhang, J. & Stout, W.F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure, *Psychometrika* **64**, 213–249.

[51] Zwick, R. (1987). Assessing the dimensionality of NAEP reading data, *Journal of Educational Measurement* **24**, 293–308.

Marc E. Gessaroli and Andre
F. de Champlain

Test Translation

Scott L. Hershberger and Dennis G. Fisher

Volume 4, pp. 2021–2026

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Test Translation

'When we sometimes despair about the use of language as a tool for measuring or at least uncovering awareness, attitude, percepts and belief systems, it is mainly because we do not yet know *why* questions that look so similar actually produce such very different results, or how we can predict contextual effects on a question, or in what ways we can ensure that the respondents will all use the same frame of reference in answering an attitude question' [19, p. 49].

The Problem

Translating a test from the language in which it was originally written (the 'source' language) to a new language (the 'target' language) is not a simple process. Problems in test translation can be thought of as problems of test inequality – before we accept the value of a translated test, we should have evidence of its semantic and conceptual equivalence to the original scale. In the following, we define semantic and conceptual equivalence, and suggest strategies for maximizing the equivalence of a translated test and the source test, as well as ways of identifying inequality when it is present.

Semantic Equivalence

Definition

Two tests are semantically equivalent if the identification of words in the target language test has identical or similar meanings to those used in the source language scale. As an example of how semantic equivalence go awry, we cite a Spanish translation of the Readiness to Change Questionnaire (RCQ), which is used to assess stages of change among substance abusers [20]. This Spanish translation of the test has been criticized as having several items that do adequately capture the ideas or meanings expressed by the corresponding English items [7]. In item seven of the RCQ, the English version appears as 'Anyone can talk about wanting to do something about drinking, but I am actually doing something about it.' This was translated to 'Cualquiera puede manifestarar [*sic*] su intención de hacer algo en relacióccon la bebida, pero yo ya estoy haciéndolo.' The word 'manifestar' (which was misspelled as 'manifestarar') means 'to manifest', and is defined as 'to make evident or certain by showing or displaying' [17]. However, this definition did not relate to the original English version, which states the idea 'to talk about wanting'. Consequently, using the action verb 'manifestar' in the translated Spanish version did not convey, nor interpret, the original English action 'to talk about wanting'. Further, in the English version, a cognitive process is being described, whereas in the translated Spanish version, an action is being described. In the following section, we describe five methods for confirming the semantic equivalence of source and target language versions of a scale.

Confirming Semantic Equivalence

Direct Translation. In direct translation, the source language test is translated into the target language, presumably by someone who is fluent in both languages. This is the extent of the translation process. Obviously, as a method for developing semantically equivalent scales, direct translation leaves much to be desired. No external checks on the fidelity of the translation are made; semantic equivalence is taken on faith.

Translation/Back-translation. This is perhaps the most common method for developing semantically equivalent scales. Translation/back-translation is a cyclical process in which the source language test is translated into the target language scale. Then, a second translator attempts to translate the target language test back into the original source language scale. If this translation is not judged sufficiently close to the original source language, another target language translation is attempted that tries to eliminate discrepancies. This process continues until the target language test satisfactorily captures the wording and meaning of the original source language scale.

Ultimate Test. The ultimate test is a two-step process [5]. In the first step, a respondent is asked to perform a behavior using the instructions from the target language version of the scale. Presumably, if the respondent performs the correct behavior, we are sure that at least the instructions are semantically equivalent.

In the second step, bilingual respondents are assigned randomly to four groups. The first group is administered the test in the source language, the second group is administered the test in the target language, the third group is administered the first half of the test in the source language and the second half in the target language, and the fourth group is administered the first half of the test in the target language and the second half in the source language. Using the four-group design, semantic equivalence is indicated if the response distributions of the four groups do not statistically differ, and if the correlation between the two halves of the test in the third and fourth groups is statistically significant.

Parallel Blind Technique. In this technique, two target language versions of the test are independently created, after which the versions are compared, with any differences being reconciled for a third and final version [23].

Random Probe Technique. Here, the target language test is administered to target language speakers. In addition to responding to the items, the respondents provide explanations for their responses. These explanations should uncover any misconceptions about the meaning of the items [8].

Solving Problems of Semantic Equivalence

Although the five procedures discussed above will help identify problems in semantic equivalence, they all, by definition, occur after the initial test translation. Potentially less effort and time will be spent if potential problems in semantic equivalence are addressed during the initial construction of the target language version of the scale. We offer three suggestions for how to maximize the probability of semantic equivalence when the source language version of the test is first constructed.

Write with Translation in Mind. Behling and Law [3] strongly advise researchers to 'write with translation in mind'. That is, the source language version of the test should be written using words, phrases, sentence structures, and grammatical forms that will facilitate the later translation of the scale. They provide a number of useful suggestions: (a) write short sentences (b) where possible, write sentences in the active rather than passive voice, (c) repeat nouns rather than substitute with ambiguous pronouns, (d) do not use colloquialisms, metaphors, slang, out-dated expressions or unusual words, (e) avoid conditional verbs such as 'could', 'would', and 'should', (f) avoid subjective qualifiers such as 'usually', 'somewhat', 'a bit', and so on, (g) do not use interrogatively worded sentences, nor double negatives.

Decentering. Decentering follows the same iterative sequence as translation/back-translation with one critical difference: Both the target language version *and* the source language version of the test can be revised during this process [5]. Specifically, following the identification of discrepancies between the back-translated test and the source language scale, decisions are made as to whether the flaws in the target language or source language versions of the test are responsible. Thus, either version can be revised. Once revisions have been made, the translation/back-translation cycle begins anew, continuing until the versions are judged to be sufficiently equivalent semantically.

Multicultural Team Approach. In this approach, a bilingual team constructs the two-language versions of the test in tandem. Some applications of this approach result in two scales that, while (presumably) measuring the same construct, comprise different items in order to capture cultural differences in how the construct is expressed.

Conceptual Equivalence

Definition

Conceptually equivalent scales measure the same construct. Problems associated with conceptually inequivalent scales have to do with the operationalization of the construct in the source language version of the test – whether the test items represent an adequate sampling of behaviors reflecting the construct in both the source *and* target language versions of the scale. Conceptual inequality often occurs from differences in the true dimensionality of a construct between cultures. The dimensionality of a construct refers to how many factors or aspects there are to the construct; for example, the construct of intelligence is sometimes described by two factors or dimensions, fluid and crystallized intelligence. While the source language test may be sufficient to capture the hypothesized dimensionality of the construct, the target language version may not because the construct has a different dimensionality across cultures. For example, Smith et al. [21] evaluated whether the three factors used to describe responses in the United States to a circadian rhythm test – the construct here being a preference for morning or evening activities – was adequate to describe responses in Japan. The researchers found that one of the three factors was poorly represented in the data from a Japanese language version of the scale. In the following section, we describe four statistical approaches to evaluating the conceptual equality of different language versions of a scale.

Statistically Confirming Conceptual Equivalence

Correlations with Other Scales. Correlations between the target language version of the test and other scales should be similar to correlations between the source language version and the same scales. For example, in English-speaking samples, we would expect a relatively high correlation between the RCQ and the SOCRATES, because both measure a substance abuser's readiness to change. We should also expect a similarly high correlation between the two scales in Spanish-speaking samples. Finding a comparable correlation does not provide strong evidence for conceptual equivalence, but neither does finding an incomparable correlation, because differences in correlation could be due to flaws in translation (a lack of semantic equality).

Exploratory Factor Analysis. Exploratory factor analysis (*see* **Factor Analysis: Exploratory**) is a technique for identifying **latent variables** (factors) by using a variety of measured variables. The analysis is considered exploratory when the concern is with determining how many factors are necessary to explain the relationships among the measured variables. Similarity of the factor structures found in the source and target language samples is evidence of conceptual equivalence. Although confirmatory factor analysis (*see* **Factor Analysis: Confirmatory**) (see below) is generally more useful for evaluating conceptual equivalence, exploratory factor analysis can still be of assistance, especially if the source and target language versions of the scales are being constructed simultaneously, as in the multicultural team approach described earlier.

There are several methods available for comparing the similarity of factors found in the source and target language samples. The *congruence coefficient* [10] measures the correlation between two factors. Another measure, the *salient variable similarity index* [6], evaluates the similarity of factors based on how many of the same items load 'significantly' (saliently) on both. *Configurative matching* [13] determines factor similarity by determining the correlations between factors from two samples that are rotated simultaneously.

Confirmatory Factor Analysis. In contrast to the raw empiricism involved in 'discovering' factors in exploratory factor analysis, confirmatory factor analysis (CFA) is a 'hypotheticist procedure designed to test a hypothesis about the relationship of certain hypothetical common factor variables, whose number and interpretation are given in advance' [18, p. 265]. In general terms, CFA is not concerned with discovering a factor structure, but with confirming the existence of a specific factor structure. Therein lies its relevance to evaluating the conceptual equivalence of source language and target language scales. Assuming we have confirmed our hypothesized factor structure for the source language scale, we want to determine whether the target language test has the same factor structure. In order for the two factor structures to be considered conceptually equivalent, a confirmatory factor model is specified that simultaneously tests equality in five areas: (a) number of factors (i.e., the dimensionality of the factor structure), (b) magnitude of the item loadings on the factors, (c) correlations among the factors, (d) error variances of the items, and (e) factor means.

Most CFA software programs allow one to test whether the same model fits across multiple samples. Space prohibits a detailed description of the procedures involved in CFA. Kline [14] provides an excellent introduction, and Kojima et al. [15] is a clear example of an application of CFA to equivalence in test translation. Typically, we begin by specifying that all parameters of the model (i.e., factor loadings, factor correlations, error variances, and factor means) are identical for the source and target language test samples. This model of *factor invariance* (identical factor structures) is evaluated by examining one or more indices of the model's fit to the data. One of the most widely used indices for assessing the fit of a model is the χ^2(chi-square) goodness-of-fit statistic. When evaluating factor invariance and testing factorial invariance, researchers are interested in a nonsignificant χ^2 goodness-of-fit test. A nonsignificant χ^2 indicates that the two scales are factor-invariant, that there is conceptual equivalence between the source and the target language scales. On the other hand, a significant χ^2 indicates that the two scales are *not* factor-invariant, but differ in one or more of the five ways noted above. The idea that the two scales are well described by identical factor structures is rejected by the data.

If the χ^2 goodness-of-fit statistic indicates that factor structures differ between samples, we usually proceed to discover why they differ. The word 'discover' is important because now we are no longer interested in *confirming* the equality of the two scales, but in *exploring* why they are unequal. Although the procedures involved in this exploration differ from those used in traditional exploratory factor analysis, philosophically and scientifically, we have returned to the raw empiricism of exploratory factor analysis. Various models are fit to the two sample data, allowing the two samples to differ in one of the five ways noted above, and equating across the other four. For example, we might specify that the factor correlations differ between the two samples, but specify that the number of factors, the factor loadings, the item error variances, and the factor means be the same. Each of the revised models (M_1) is associated with its own χ^2 goodness-of-fit statistic, which is compared to the χ^2 statistic associated with the factor invariance model (M_0), the model that specified equality between the source and target language scales on all parameters. The χ^2 associated with M_1 must always be no smaller than the χ^2 associated with M_0. If the difference in the two model chi-squares, which is also chi-square distributed, is significant, M_1 fits the data better than M_0. Our interpretation at this point would be that the factor correlations differ between the two scales. The process of model modification could take any number of paths at this point; opinion varies as to the 'best' sequence of model modifications [4]. We could choose to specify, one at a time, test inequality on either the number of factors, factor loadings, item error variances, or factor means while specifying equality on the other parameters. Alternatively, we could specify a model that retains test inequality for the factor correlations, and add, one at a time, test inequality for the number of factors, item error variances, or factor means. Let us call either of these models M_2. We would then evaluate the significance of $\chi^2_{M_2} - \chi^2_{M_1}$, and use that result to determine subsequent model modifications, until, finally, we find a model that fits the data well and whose fit cannot be improved by any other subsequent model modifications. The researcher must always bear in mind, however, that the validity of the final model is quite tentative, and should be cross-validated on new samples. In any event, unless the model of factor invariance, M_0, fits the data well, the source and target language scales are not conceptually equivalent.

Item Response Theory. Item response theory (IRT) is another powerful method for testing conceptual equivalence (*see* **Item Response Theory (IRT) Models for Dichotomous Data**). Again, space limitations prohibit us from discussing IRT in detail; [9] provides an excellent introduction and [2] is a clear example of an application of IRT to equivalence in test translation.

In IRT, the probability of a correct response to an item is modeled using one or more parameters. A typical set of item parameters is item difficulty and item discrimination. For two tests to be conceptually equivalent, corresponding items on the tests must have identical item difficulties and item discriminations. In IRT, an item is said to have **differential item functioning** (DIF) when the item difficulty and item discrimination of the same item on two versions of the test differ significantly. DIF can be tested statistically using a number of methods. The *parameter equating method* [16] tests differences between two tests' item parameters using a chi-square statistics. The **Mantel–Haenszel method** [11] (*see* **Item Bias Detection: Classical Approaches**) also uses a chi-square statistic to test whether the observed frequencies of responses to one test version differs significantly from the frequencies expected if there were no differences in the item parameters. When all items within a scale are assumed to have the same discrimination, a t Test, the *item difficulty shift statistic* [12], is used to test for differences in an item's difficulty between two tests. Yet another measure of item DIF is the *mean square residual* [24], which involves analyzing the fit of each item in the source language test group and the target language test group. An item should either fit or fail to fit in a similar manner for both samples; thus it relates to the concept being measured in the same way for each group. Thissen et al. [22] describe a model comparison method for testing DIF that is analogous to the model comparison method used to test for factor invariance in CFA. Item parameter estimates are obtained from a model that equates the parameters between the two groups, and item parameter estimates are obtained from a second model that allows one or more of the item parameters to differ between the two groups. If the difference between the two models' chi-squares is significant, the item parameters differ between the groups.

Some methods for detecting DIF do not require IRT estimation of item parameters. For example, the *delta-plot method* [1] involves plotting pairs of item difficulties (deltas) for the same item from the two groups. If the two tests are conceptually equivalent, we would expect that the item difficulties would order themselves in the same way in the two groups, thus forming a 45° line from the origin when plotted.

Conclusion

Although the methods involved in confirming semantic and conceptual equivalence have been described separately, in practice, it is frequently difficult to determine whether apparent flaws in test translation are attributable to the one or the other or both. For example, in [21] the researchers attributed factor structure differences not to true cultural differences in the nature of the construct (conceptual inequivalence), but to flaws in the Japanese translation (semantic inequivalence). Although the researchers were certain that semantic inequivalence was the culprit, the evidence they present, at least to our understanding, does not argue strongly for the lack of either type of equivalence. No doubt, in this study, and in many other studies, test translation problems will arise from a bit of each.

References

[1] Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias, in *Handbook of Methods for Detecting Test Bias*, R.A. Berk, ed., The John Hopkins University Press, Baltimore, pp. 96–116.

[2] Beck, C.T. & Gable, R.K. (2003). Postpartum depression screening scale: Spanish version, *Nursing Research* **52**, 296–306.

[3] Behling, O. & Law, K.S. (2000). *Translating Questionnaires and Other Research Instruments: Problems and Solutions*, Sage Publications, Thousand Oaks.

[4] Bollen, K.A. (2000). Modeling strategies: in search of the holy grail, *Structural Equation Modeling* **7**, 74–81.

[5] Brislin, R.W. (1973). Questionnaire wording and translation, in *Cross-Cultural Research Methods*, R.W. Brislin, W.J. Lonner & R.M. Thorndike, eds, John Wiley & Sons, New York, pp. 32–58.

[6] Cattell, R.B. (1978). *The Scientific Use of Factor Analysis in Behavioral and Life Sciences*, Plenum Press, New York.

[7] Fisher, D.G., Rainof, A., Rodríguez, J., Archuleta, E. & Muñiz, J.F. (2002). Translation problems of the Spanish version of the readiness to change questionnaire, *Alcohol & Alcoholism* **37**, 100–101.

[8] Guthery, D. & Lowe, B.A. (1992). Translation problems in international marketing research, *Journal of Language for International Business* **4**, 1–14.

[9] Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*, Sage Publications, Newbury Park.

[10] Harman, H.H. (1976). *Modern Factor Analysis*, 3rd Edition, University of Chicago Press, Chicago.

[11] Holland, P.W. & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure, in *Testing Validity*, H. Wainer & H.I. Braun, eds, Lawrence Erlbaum, Hillsdale, pp. 129–146.

[12] Ironson, G.H. (1982). Use of chi-square and latent trait approaches for detecting item bias, in *Handbook of Methods for Detecting Test Bias*, R.A. Berk, ed., The John Hopkins University Press, Baltimore, pp. 117–160.

[13] Kaiser, H., Hunka, S. & Bianchini, J. (1971). Relating factors between studies based upon different individuals, *Multivariate Behavior Research* **6**, 409–422.

[14] Kline, R.B. (1998). *Principles and Practice of Structural Equation Modeling*, The Guilford Press, New York.

[15] Kojima, M., Furukawa, T.A., Takahashi, H., Kawaii, M., Nagaya, T. & Tokudome, S. (2002). Cross-cultural validation of the beck depression inventory-II in Japan, *Psychiatry Research* **110**, 291–299.

[16] Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum, Hillsdale.

[17] Mish, F.C., ed. (2000). *Merriam-Webster's Collegiate Dictionary*, Springfield Merriam-Webster.

[18] Mulaik, S.A. (1988). Confirmatory factor analysis, in *Handbook of Multivariate Experimental Psychology*, 2nd Edition, J.R. Nesselroade & R.B. Cattell, eds, Plenum Press, New York, pp. 259–288.

[19] Oppenheim, A.N. (1992). *Questionnaire Design, Interviewing and Attitude Measurement*, Continuum International, London.

[20] Rodríguez-Martos, A., Rubio, G., Auba, J., Santo-Domingo, J., Torralba, Ll. & Campillo, M. (2000). Readiness to change questionnaire: reliability study

of its Spanish version, *Alcohol & Alcoholism* **35**, 270–275.

[21] Smith, C.S., Tisak, J., Bauman, T. & Green, E. (1991). Psychometric equivalence of a translated circadian rhythm questionnaire: implications for between- and within-population assessments, *Journal of Applied Psychology* **76**, 628–636.

[22] Thissen, D., Steinberg, L. & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models, in *Differential Item Functioning: Theory and Practice*, P.W. Holland & H. Wainer, eds, Lawrence Erlbaum, Hillsdale, pp. 147–169.

[23] Werner, O. & Campbell, T.D. (1970). Translating, working through interpreters, and problems of decentering, in *A Handbook of Method in Cultural Anthropology*, R. Noll & R. Cohen, eds, Columbia University Press, New York, pp. 398–420.

[24] Wright, B.D. & Masters, G.N. (1982). *Rating Scale Analysis*, MESA Press, Chicago.

SCOTT L. HERSHBERGER AND DENNIS G. FISHER

Tetrachoric Correlation

Scott L. Hershberger

Volume 4, pp. 2027–2028

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Tetrachoric Correlation

The *tetrachoric correlation* is used to correlate two artificially dichotomized variables, X and Y, which have a *bivariate-normal distribution* (*see* **Catalogue of Probability Density Functions**) [8]. In a bivariate-normal distribution, the distribution of Y is normal at fixed values of X, and the distribution of X is normal at fixed values of Y. Two variables that are bivariate-normally distributed must each be normally distributed. The calculation of the tetrachoric correlation involves corrections that approximate what the **Pearson product-moment correlation** would have been if the data had been continuous. If instead of the tetrachoric correlation, the Pearson product-moment correlation formula is directly applied to the data (we are not willing to assume that X and Y are truly bivariate-normally distributed), the resulting correlation is referred to as a *phi correlation.*

Tetrachoric correlations are often useful in behavioral genetic research (*see* **Correlation Issues in Genetics Research**). For example, we might have twin concordance data for a psychiatric illness such as schizophrenia; either both twins are diagnosed as schizophrenic (concordant twins) or only one of the co-twins is diagnosed as schizophrenic (discordant twins) [4]. Instead of assuming that is there is true phenotypic discontinuity in schizophrenia due to a single gene of large effect, we assume that the discontinuity is an arbitrary result of classifying people by kind rather than by degree. In the latter case, the phenotype is truly continuous, the result of many independent genetic factors, leading to a continuous distribution of genotypes for schizophrenia. Models that assume that a continuous distribution of genotypes underlie an artificially dichotomized phenotype are referred to as *continuous liability models* [7].

The genetic analysis of continuous liability models (*see* **Liability Threshold Models**) assumes that the liability for the phenotype (e.g., schizophrenia) in pairs of twins is bivariate-normal with zero mean vector and a correlation ρ between the liabilities of twin pairs. This is the tetrachoric correlation. If both twins are above a genetic *threshold t*, then both twins will be diagnosed as schizophrenic. If both are below the genetic threshold, then neither twin will be diagnosed as schizophrenic. If one twin is above the threshold and the other below, then the first will be diagnosed as schizophrenic and the other will not.

The probability that both twins will be diagnosed as schizophrenic is thus

$$p_{11} = \int_t^\infty \int_t^\infty \theta(x, y, \rho)\mathrm{dydx}, \qquad (1)$$

where $\theta(x, y, \rho)$ is the bivariate-normal probability function. Similarly, the probability that the first twin will be diagnosed as schizophrenic and the second will not is

$$p_{10} = \int_t^\infty \int_{-\infty}^t \theta(x, y, \rho)\mathrm{dydx}. \qquad (2)$$

Similar expressions follow for the other categories of twin diagnosis, p_{01} and p_{00}.

In order to obtain values for t, the genetic threshold, and ρ, the tetrachoric correlation between twins, the bivariate-normal probability integrals are evaluated numerically for selected values of t and ρ values that maximize the likelihood of the data [3]. The observed data are the number of twin pairs in each of the four cells of the twin **contingency table** for schizophrenia. Thus, the log-likelihood to be maximized for a given contingency table is

$$L = C + \sum_i \sum_j N_{ij} \ln p_{ij}, \qquad (3)$$

where C is some constant. Estimates of t and ρ that produce the largest value of L for a contingency table are **maximum likelihood estimates**.

As an example of computing the tetrachoric correlation for a dichotomous phenotype, consider the monozygotic twin concordance data ($N = 56$ pairs) for a male homosexual orientation from [2]:

		Twin1	
		Yes	No
Twin 2	Yes	29	14
	No	13	0

Assuming a model in which a homosexual orientation is due to additive genetic and random environmental effects, the tetrachoric correlation between monozygotic twins is 0.50. The phi correlation between the twins' observed dichotomized phenotypes is 0.32.

The tetrachoric correlation is also often used in the factor analysis of dichotomous item data for the following reason. In the context of cognitive testing, where an item is scored 'right' or 'wrong', a measure of the difficulty of an item is the proportion of the sample that passes the item [1]. Factor analyzing binary item data on the basis of phi correlations may lead to the extraction of spurious factors known as *difficulty factors*, because there may be a tendency to yield factors defined solely by items of similar difficulty. Why difficulty factors are produced by factoring phi correlations has been attributed to the restricted range of phi correlations; $-.8 \leq 0 \leq -.8$ [5], and to the erroneous application of the linear common factor model to the inherently nonlinear relationship that may exist between item responses and latent factors [6]. In either case, use of the tetrachoric correlation instead of the phi correlation may prevent the occurrence of spurious difficulty factors.

References

[1] Allen, M.J. & Yen, W.M. (1979). *Introduction to Measurement Theory*, Wadsworth, Monterey.

[2] Bailey, J.M. & Pillard, R.C. (1991). A genetic study of male sexual orientation, *Archives of General Psychiatry* **48**, 1089–1096.

[3] Eaves, L.J., Last, K.A., Young, P.A. & Martin, P.A. (1978). Model-fitting approaches to the analysis of human behavior, *Heredity* **41**, 249–320.

[4] Gottesman, I.I. & Shields, J. (1982). *Schizophrenia: The Epigenetic Perspective*, Cambridge University Press, New York.

[5] Harman, H.H. (1976). *Modern Factor Analysis*, 3rd Edition, University of Chicago Press, Chicago.

[6] McDonald, R.P. & Ahlawat, K.S. (1974). Difficulty factors in binary data, *British Journal of Mathematical and Statistical Psychology* **27**, 82–99.

[7] Neale, M.C. & Cardon, L.R. (1992). *Methodology for Genetic Studies of Twins and Families*, Kluwer Academic Press, Boston.

[8] Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory*, 3rd Edition, McGraw-Hill, New York.

Scott L. Hershberger

Theil Slope Estimate

CLIFFORD E. LUNNEBORG

Volume 4, pp. 2028–2030

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Theil Slope Estimate

The Linear Regression Model

In simple linear regression, the expected or mean value of a response variable, Y, is modeled as a linear function of the value of an explanatory variable, X:

$$E(Y|X = x_i) = \alpha + \beta x_i, \tag{1}$$

(*see* **Multiple Linear Regression**); that is, for each value of X of interest to the researcher, the values of Y are distributed about this mean value. A randomly chosen observation, y_i, from the distribution of responses for $X = x_i$ is modeled as

$$y_i = \alpha + \beta x_i + e_i, \tag{2}$$

where e_i is a random draw from a distribution of deviations. The task in regression is to estimate the unknown regression constants, α and β, and, often, to test hypotheses about their values.

Estimation and hypothesis testing depend upon a sample of n paired observations, $(x_i, y_i), i = 1, 2, \ldots, n$. The n values of the explanatory variable are treated as fixed constants, values specified by the researcher.

The Least Squares Estimates of the Regression Parameters

The most common estimates of the regression parameters are those based on minimizing the average squared discrepancy between the sampled values of Y and the estimated means of the distributions from which they were sampled (*see* **Least Squares Estimation**); that is, we choose as estimates of the linear model intercept and slope those values, $\widehat{\alpha}$ and $\widehat{\beta}$, that minimize the mean squared deviation:

$$MSD = \tfrac{1}{n}\sum_{i=1}^{n}\left[y_i - \left(\widehat{\alpha} - \widehat{\beta}x_i\right)\right]^2. \tag{3}$$

The resulting estimates can be expressed as functions of the sample means, standard deviations, and correlation: $\widehat{\beta} = r_{xy}[SD(y)/SD(x)]$ and $\widehat{\alpha} = \overline{y} - \widehat{\beta\overline{x}}$.

The Normal Linear Regression Model

The normal regression model assumes that response observations are sampled independently and that the deviations, the e_is, are distributed as the normal random variable with a mean of zero and a variance, σ^2, that does not depend upon the value of x_i. Under this model, the least squares slope and intercept estimates are unbiased and with sampling distributions that are normal with variances that are a function of σ^2 and the mean and variance of the x_is. As σ^2 is estimated by $(n/n-2)MSD$, this leads to hypothesis tests and confidence-bound estimates based on one of Student's t distributions.

Nonparametric Estimation and Hypothesis Testing

The least squares slope estimate can be influenced strongly by one or a few outlying observations thus providing a misleading summary of the general strength of relation of the response observations to the level of the explanatory variable. A more robust estimator has been proposed by Theil [4].

Order the paired observations, (x_i, y_i), in terms of the sizes of the x_is, letting $(x_{[1]}, y_1)$ designate the pair with the smallest value of X and $(x_{[n]}, y_n)$ the pair with the largest value of X.

For every pair of explanatory variable scores for which $x_{[j]} < x_{[k]}$, we can compute the two-point slope

$$S_{jk} = \frac{(y_k - y_j)}{(x_{[k]} - x_{[j]})}. \tag{4}$$

If there are no ties in the values of X, there will be $(n-1)n/2$ such slopes; if there are ties, the number will be smaller.

The Theil slope estimator is the median of these two-point slopes,

$$\widehat{\beta}_T = Mdn(S_{jk}). \tag{5}$$

Outlying observations will be accompanied by two-point slopes that are either unusually small or unusually large. These are discounted in the Theil estimator.

Conover [1] describes how the ordered set of two-point slopes, the S_{jk}s, can be used as well to find a **confidence interval** for β.

An intercept estimator related to the Theil slope estimate has been proposed by Hettmansperger, McKean & Sheather [2]. If we use the Theil slope estimate to compute a set of differences of the form, $a_i = y_i - \widehat{\beta}_T x_i$, then the regression intercept can be estimated robustly by the median of these differences,

$$\widehat{\alpha}_H = Mdn(a_i). \tag{6}$$

To carry out Theil's nonparametric test of the hypothesis that $\beta = \beta_0$, we first compute the n differences, $D_i = y_i - \beta_0 x_i$. If β has been correctly described by the null hypothesis, any linear dependence of the y_is on the x_is will have been accounted for. As a result, the D_is will be uncorrelated with the x_is. Hollander and Wolf [3] propose carrying out Theil's test by computing **Kendall's rank correlation** between the D_is and the x_is, $\tau(D, x)$, and testing whether τ differs significantly from zero. Conover [1] suggests using the **Spearman rank correlation**, $\rho(D, x)$, for the same test. Hollander and Wolf [3] provide, as well, an alternative derivation of the Theil test.

Example

As an example of the influence of an outlier on the estimation of regression parameters, [5] gives the artificial example in Table 1.

The y value of 12.74 clearly is out of line with respect to the other observations. The least squares estimates of the regression parameters are $\widehat{\alpha} = 3.002$ and $\widehat{\beta} = 0.500$. As there are no ties among the x values, there are 55 two-point slopes. Their median provides the Theil estimate of the slope parameter, $\widehat{\beta}_T = 0.346$. The corresponding Hettmansperger intercept estimate is $\widehat{\alpha}_H = 4.004$. Both differ considerably from the least squares estimates.

Table 1 An Anscombe influence example

x	y
4	5.39
5	5.73
6	6.08
7	6.42
8	6.77
9	7.11
10	7.46
11	7.81
12	8.15
13	12.74
14	8.84

Following the procedure outlined in [1], the 95% confidence interval for β is given by the 17th and 39th of the ordered two-point slopes: [0.345, 0.347]. When one ignores the aberrant two-point slopes associated with the point (13, 12.74), the other two-point slopes are in remarkable agreement for this example.

References

[1] Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Edition, Wiley, New York.

[2] Hettmansperger, T.P., McKean, J.W. & Sheather, S.J. (1997). Rank-based analyses of linear models, in *Handbook of Statistics*, Vol. 15, Elsevier-Science, Amsterdam.

[3] Hollander, M. & Wolfe, D.A. (1999). *Nonparametric Statistical Methods*, 2nd Edition, Wiley, New York.

[4] Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis, *Indagationes Mathematicae* **12**, 85–91.

[5] Weisberg, S. (1985). *Applied Linear Regression*, 2nd Edition, Wiley, New York.

Clifford E. Lunneborg

Thomson, Godfrey Hilton

Pat Lovie

Volume 4, pp. 2030–2031

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Thomson, Godfrey Hilton

Born: March 27, 1881, in Carlisle, England.
Died: February 9, 1955, in Edinburgh, Scotland.

Godfrey Thomson's life did not start propitiously. His parents separated when he was an infant and his mother took him to live with her own mother and sisters in her native village of Felling, on industrial south Tyneside. There he attended local schools until the age of 13, narrowly avoiding being placed into work as an engineering patternmaker by winning a scholarship to Rutherford College in Newcastle-upon-Tyne. Three years later, he was taken on as a pupil-teacher (essentially, an apprenticeship) in his old elementary school in Felling, but had to attend Rutherford College's science classes in the evenings and weekends. His performance in the London University Intermediate B.Sc. examinations earned him a Queen's Scholarship in 1900, which allowed him to embark on a full-time science course and also train as a teacher at the Durham College of Science (later renamed Armstrong College) in Newcastle, at that time part of the University of Durham. He graduated in 1903, with a distinction in Mathematics and Physics, the year after obtaining his teaching certificate. With the aid of a Pemberton Fellowship from Durham, he set off to study at the University of Strasburg in 1903, gaining a Ph.D. for research on Hertzian waves three years later.

The Queen's Scholarship, however, had strings attached that required the beneficiary to teach for a certain period in 'in an elementary school, the army, the navy, or the workhouse'. Apparently, an assistant lectureship at Armstrong College qualified under these headings! As teaching educational psychology was one of Thomson's duties, he felt obliged to learn something of psychology generally, and was mildly surprised to find that he enjoyed the experience. However, it was during a summer vacation visit to C. S. Myers's laboratory in Cambridge in 1911 that his interest was caught by **William Brown's** book *The Essentials of Mental Measurement* [1]. Although Thomson's initial foray was on the psychophysical side (and led to publications that earned him a D.Sc. in 1913), he was also intrigued by Brown's criticisms of **Charles Spearman's** two-factor theory of human ability. According to Thomson [7], sitting at his fireside and armed with only a dice, a house slipper, and a notepad, he was able to generate sets of artificial scores with a correlational structure consistent with Spearman's theory, but without needing its cornerstone, the single underlying general factor g. The publication of this finding in 1916 [3] marked both the start of Thomson's many significant contributions to the debates on intelligence and the newly emerging method of **factor analysis**, and also of a long running, and often bitter, quarrel with Spearman.

Thomson's own notion of intelligence evolved into his 'sampling hypothesis' in which the mind was assumed to consist of numerous connections or 'bonds' and that, inevitably, tests of different mental abilities would call upon overlapping samples of these bonds. In Thomson's view, therefore, the correlational structure that resulted suggested a statistical rather than a mental phenomenon. *The Factorial Analysis of Human Ability* [4], published in five editions between 1939 and 1951, was Thomson's major work on factor analysis. He also coauthored with Brown several further editions in 1921, 1925, and 1940, of the book, *The Essentials of Mental Measurement,* that had so fired him in 1911.

However, it is for his work on devising mental tests that Thomson is best remembered. This began in 1920 for the newly promoted Professor Thomson with a commission from Northumberland County Council for tests that could be used in less-privileged schools to select pupils who merited the opportunity of a secondary education – as Thomson himself had benefited many years earlier. By 1925, after a year spent in the United States with E. L. Thorndike, he had accepted a chair of education at Edinburgh University, with the associated post of Director of Moray House teacher training college, and had begun to formulate what became known as the Moray House tests; these tests would be widely used in schools throughout the United Kingdom and many other countries. Thomson and his many collaborators were also involved in a large-scale study of how the test results from schoolchildren related to various social factors, including family size and father's occupation, and to geographical region.

Thomson received many honors from learned societies and academies abroad. He was knighted in 1949. Even after retiring in 1951, he was still writing and working diligently on data from a longitudinal Scottish study. Thomson's final book, *The Geometry of Mental Measurement* [6], was published in 1954.

'God Thom', as he was nicknamed (though not wholly affectionately), by his students in Edinburgh, died from cancer in 1955 at the age of 73.

Further material on Thomson's life and work can be found in [2, 5, 7].

References

[1] Brown, W. (1911). *The Essentials of Mental Measurement*, Cambridge University Press, London.

[2] Sharp, S. (1997). "Much more at home with 3.999 pupils than with four": the contributions to psychometrics of Sir Godfrey Thomson, *British Journal of Mathematical and Statistical Psychology* **50**, 163–174.

[3] Thomson, G.H. (1916). A hierarchy without a general factor, *British Journal of Psychology* **8**, 271–281.

[4] Thomson, G. (1939). *The Factorial Analysis of Human Ability*, London University Press, London.

[5] Thomson, G. (1952). Godfrey Thomson, in *A History of Psychology in Autobiography*, Vol. 4, E.G. Boring, H.S. Langfeld, H. Werner & R.M. Yerkes, eds, Clark University Press, Worcester, pp. 279–294.

[6] Thomson, G.H. (1954). *The Geometry of Mental Measurement*, London University Press, London.

[7] Thomson, G. (1969). *The Education of an Englishman*, Moray House College of Education, Edinburgh.

PAT LOVIE

Three Dimensional (3D) Scatterplots

Daniel B. Wright and Siân E. Williams

Volume 4, pp. 2031–2032

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Three Dimensional (3D) Scatterplots

A standard **scatterplot** is appropriate to display the relationship between two continuous variables. But what happens if there are three variables? If the third variable is categorical, it is customary to print different symbols for the different points. If the variable is metric, then many packages allow the user to set the size of the data points to represent the value on the third variable and the result is the **bubble plot**. This is what is normally recommended when graphing three continuous variables and the sample size is small. Another possibility is to make a series of two-variable scatterplots for each bivariate comparison, sometimes called a **scatterplot matrix**. But if it is the three-way relationships that is of interest, these bivariate scatterplots are not appropriate.

An alternative procedure available in many graphics packages is to plot the data points for three variables in a three dimensional space. Like the standard scatterplot, the data points are placed at their appropriate location within a coordinate space, but, this time, the space is three dimensional. Because paper and computer screens are two dimensional, it is important to use some of the available features, such as rotation of axes, so that the all the dimensions are clear.

Figure 1(a) shows data on the baseline scores for working memory span using three tests: digit span, visual spatial span, and Corsi block span [1]. These were expected to be moderately correlated, and they are. While this plot can help in understanding the patterns in the data, it can still be difficult to make sense of the data. The Corsi task is more complex than the other two tasks, and the researchers were interested in how well the digit and visual spatial tasks could predict the Corsi scores. The resulting regression plane has been added to Figure 1(b). This helps to show the general pattern of the cloud of data points. Other planes could be used instead (for example, from more robust methods, polynomials, etc.).

There is a sense in which the three-dimensional scatterplots attempt to make the two-dimensional page into a three-dimensional object, and this can never be wholly satisfactory. Using size, contours, and colors to show values on other dimensions within a two-dimensional space are often easier for the reader to interpret.

Reference

[1] Wright, D.B. & Osborne, J.E. (in press). Dissociation, cognitive failures, and working memory, *American Journal of Psychology*.

DANIEL B. WRIGHT AND SIÂN E. WILLIAMS

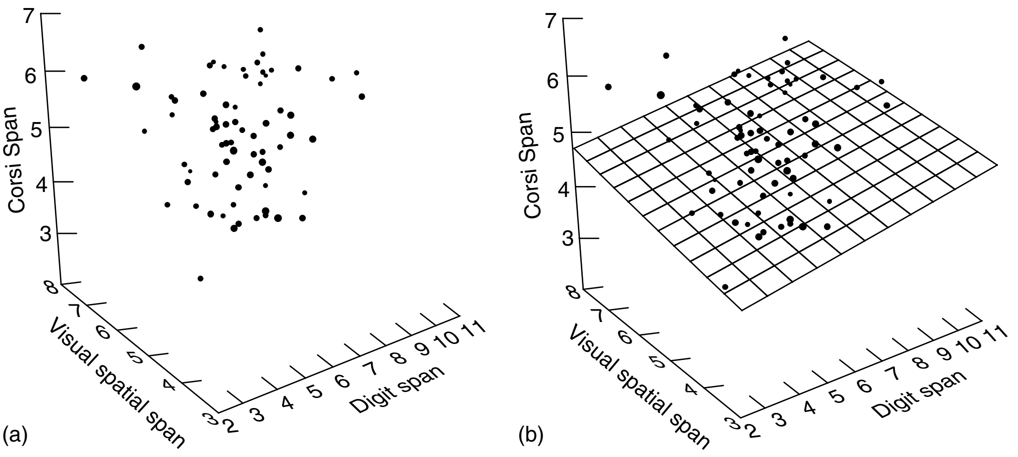


Figure 1 Showing three dimensions in a scatterplot. Figure 1a shows only the data points. Figure 1b includes the linear regression plane

Three-mode Component and Scaling Methods

PIETER M. KROONENBERG

Volume 4, pp. 2032–2044

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Three-mode Component and Scaling Methods

Introduction

What is Three-mode Analysis?

Most statistical methods are used to analyze the scores of objects (subjects, groups, etc.) on a number of variables, and the data can be arranged in a two-way *matrix*, that is, a rectangular arrangement of rows (objects) and columns (variables) (*see* **Multivariate Analysis: Overview**). However, data are often far more complex than this, and one such complexity is that data have been collected under several conditions or at several time points; such data are referred to as *three-way data* (see Figure 1). Thus, for each condition there is a matrix, and the set of matrices for all conditions can be arranged next to each other to form a broad matrix of subjects by variables times conditions. Alternatively, one may create a tall matrix of subjects times conditions by variables. The third possibility is to arrange the set of matrices in a three-dimensional block or three-way *array*, so that metaphorically the data now fit into a box. The collection of techniques that attempt to analyze such data boxes are referred to as *three-mode methods*, and making sense of such data is the art of *three-mode analysis*. Thus three-mode analysis is the analysis of data that fit into boxes.

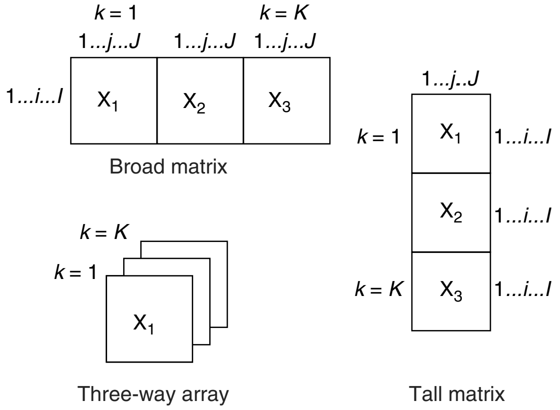


Figure 1 Two-way matrices and a three-way array

Usually a distinction is made between *three-way* and *three-mode*. The word *way* is more general and points to the three-dimensional arrangement irrespective of the content of the data, while the word *mode* is more specific and refers to the content of each of the ways. Thus objects, variables, and conditions can be the three modes of a data array. When the same entities occur twice, as is the case in a correlation matrix, and we have correlation matrices for the same variables measured in several samples, one often speaks of a two-mode three-way data array, where the variables and the samples are the two modes. However, to avoid confusion and wordiness, we generally refer to three-way data and three-way arrays, and to three-mode methods and three-mode analysis, with the exception of the well-established name of three-way analysis of variance. The word *two-mode analysis* is then reserved for the analysis of two-way matrices.

Why Three-mode Analysis?

Given that there are so many statistical and data-analytic techniques for two-way data, why are these not sufficient for three-way data? The simplest answer is that two-mode methods do not respect the three-way design of the data. Such disrespect is not unusual as, for instance, time series data are often analyzed as if the time mode was an unordered mode and the time sequence is only used in interpretation.

Three-way data are supposedly collected because all three modes were necessary to answer the pertinent research questions. Such research questions can be facetiously summarized as: ‘Who does what to whom and when?’, or more specifically: ‘Which groups of subjects behave differently on which variables under which conditions?’ or in an agricultural setting ‘Which plant varieties behave in a specific way in which locations on which attributes?’. Such questions cannot be answered with two-mode methods, because there are no separate parameters for all three modes. When analyzing three-way data with two-mode methods, one has to rearrange the data as in Figure 1, and this means that either the subjects and conditions are combined to a single mode (‘tall matrix’) or the variables and conditions are so combined (‘broad matrix’). Thus, two of the modes are always confounded and no independent parameters for these modes are present in the model itself.

In general, a three-mode model is much more parsimonious for three-way data then an appropriate two-mode model. To what extent this is true depends very much on the specific model used. In some three-mode component models, low-dimensional representations are defined for all three modes, which can lead to enormous reductions in parameters. Unfortunately, it cannot be said that this means that automatically the results of a three-mode analysis are always easier to interpret. Again this depends on the questions asked and the data and models used.

An important aspect of three-mode models, especially in the social and behavioral sciences, is that they allow the analysis of individual differences. The subjects from whom the data have been collected do not disappear in sufficient statistics for distributions, such as means, (co)variances or correlations, and possibly higher-order moments such as the **kurtosis** and **skewness**, but they are examined in their own right. This implies that often the data at hand are taken as is, and not necessarily as a random sample from a larger population in which the subjects are in principle exchangeable. Naturally, this affects the generalizability but that is considered inevitable. At the same time, however, the subjects are recognized as the 'data generators' and are awarded a special status, for instance, when statistical stability is determined via **bootstrap** or **jackknife** procedures. Furthermore, it is nearly always the contention of the researcher that similar samples are or may become available, so that at least part of the results are valid outside the context of the specific sample.

Three-way Data

As mentioned above, three-way data fit into three-dimensional boxes, which in the social and behavioral sciences often take the form of subjects by variables by conditions.

The first way (subjects) has index i running along the vertical axis, the second way or variables index j runs along the horizontal axis, and the third way or conditions index k runs along the 'depth' axis of the box. The *number of levels* in each way is I, J, and K. The $I \times J \times K$ three-way data matrix $\underline{\mathbf{X}}$ is thus defined as the collection of elements, x_{ijk} with the indices $i = 1, \ldots, I$; $j = 1, \ldots, J$; $k = 1, \ldots, K$.

Athree-way array can also be seen as a collection 'normal' (=two-way) matrices or *slices*. There are three different arrangements for this, as is shown in Figure 2. Furthermore, one can break up a three-way matrix into one-way submatrices (or vectors), called *fibers* (see Figure 3). The slices are referred to as *frontal* slices, *horizontal* slices, and *lateral* slices. The fibers are referred to *rows*, *columns*, and *tubes*. The prime reference paper for three-mode terminology is [35].

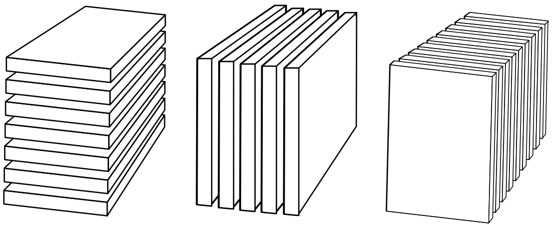


Figure 2 Slices of a three-mode data array; horizontal, lateral and frontal, respectively

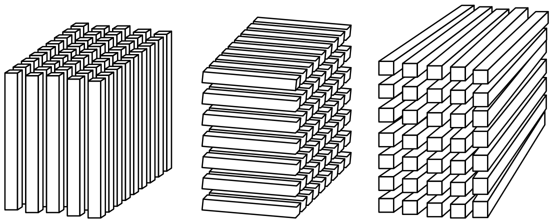


Figure 3 Fibers of a three-mode data array; columns, rows, and tubes, respectively

A Brief Example: Abstract Paintings

Consider the situation in which a number of persons (Mode 1) have rated twenty abstract paintings (Mode 2) using some 10 different rating scales which measure the feelings these paintings elicit (Mode 3). Suppose a researcher wants to know if there is a common structure underlying the usage of the rating scales with respect to the paintings, how the various subjects perceive this common structure, and/or whether subjects can be seen as types or combination of types in their use of the rating scales to describe their emotions. Although all subjects might agree on the kinds (or dimensions) of feelings elicited by the paintings, for some subjects some of these dimensions might be more important and/or more correlated than for other subjects, and one could imagine that different types of subjects evaluate the paintings in different ways. One can gain insight into such problems by constructing graphs or clusters not only for the paintings, and for the rating scales, but also for the subjects, and find ways to combine the information from all three modes into a coherent story about the ways people look at and feel about abstract paintings.

Models and Methods for Three-way Data

Three-mode Component Models

Three-mode analysis as an approach towards analyzing three-way data started with Ledyard Tucker's publications [53, 54, 55]). By early 2004, his work on three-mode analysis was cited about 500 to 600 times in the journal literature. He called his main model *three-mode factor analysis*, but it is now generally referred to as *three-mode component analysis*, or more specifically the *Tucker3 model*. Before his series papers, various authors had investigated ways to deal with sets of matrices, especially from a purely linear algebra point of view, but studying three-way *data* really started with Tucker's seminal work. In the earlier papers, Tucker formulated two models (the principal component model and a common factor model) and several computational procedures. He also wrote and collaborated on about 10 applications, not all of them published. Levin, one of his Ph.D. students, wrote an expository paper on applying the technique in psychology [44]. After that, the number of applications and theoretical papers gradually, but slowly, increased. A second major step was taken when Kroonenberg and De Leeuw [38] presented an improved, least-squares solution for the original component model as well as a computer program to carry out the analysis. Kroonenberg [37] also presented an overview of the then state of the art with respect to Tucker's component model, as well as an annotated bibliography [36].

Table 1 gives an overview of the major three-mode models which are being used in practice. Without going into detail here, a perusal of the table will make clear how these models are related to one another by imposing restrictions or adding extensions.

Table 1 Major component and scaling models for three-way data

Model	Sum notation	Matrix and vector notation
SVD	$x_{ij} \approx \sum_{s=1}^{S} w_{ss}(a_{is}b_{js})$	$\mathbf{X} = \mathbf{AWB}' = \sum_{s=1}^{S} w_{ss}(\mathbf{a}_s \otimes \mathbf{b}_s)$
Tucker2	$x_{ijk} \approx \sum_{p=1}^{P}\sum_{q=1}^{Q} h_{pqk}(a_{ip}b_{jq})$	$\mathbf{X}_k \approx \mathbf{AH}_k\mathbf{B}'$
Tucker3	$x_{ijk} \approx \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}(a_{ip}b_{jq}c_{kr})$	$\underline{\mathbf{X}} \approx \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{pqr}(\mathbf{a}_p \otimes \mathbf{b}_q \otimes \mathbf{c}_r)$
Parafac	$x_{ijk} \approx \sum_{s=1}^{S} \tilde{w}_{kss}(\tilde{a}_{is}\tilde{b}_{js})\ [\tilde{w}_{kss} = g_{sss}\tilde{c}_{ks}]$	$\mathbf{X}_k \approx \tilde{\mathbf{A}}\tilde{\mathbf{W}}_k\tilde{\mathbf{B}}'$
	$x_{ijk} \approx \sum_{s=1}^{S} g_{sss}(\tilde{a}_{is}\tilde{b}_{js}\tilde{c}_{ks})$	$\underline{\mathbf{X}} \approx \sum_{s=1}^{S} g_{sss}(\tilde{\mathbf{a}}_s \otimes \tilde{\mathbf{b}}_s \otimes \tilde{\mathbf{c}}_s)$
IDIOSCAL	$x_{ii'k} \approx \sum_{p=1}^{P}\sum_{p'=1}^{P} h_{pp'k}(a_{ip}a_{i'p'})$	$\mathbf{X}_k \approx \mathbf{AH}_k\mathbf{A}'$
INDSCAL	$x_{ii'k} \approx \sum_{s=1}^{S} \tilde{w}_{kss}(\tilde{a}_{is}\tilde{a}_{i's})\ [\tilde{w}_{kss} = g_{sss}\tilde{c}_{ks}]$	$\mathbf{X}_k \approx \tilde{\mathbf{A}}\tilde{\mathbf{W}}_k\tilde{\mathbf{A}}'$
	$x_{ijk} \approx \sum_{s=1}^{S} g_{sss}(\tilde{a}_{is}\tilde{a}_{i's}\tilde{c}_{ks})$	$\underline{\mathbf{X}} \approx \sum_{s=1}^{S} g_{sss}(\tilde{\mathbf{a}}_s \otimes \tilde{\mathbf{a}}_s \otimes \tilde{\mathbf{c}}_s)$

Notes: $\underline{\mathbf{G}} = (g_{pqr})$ is a full $P \times Q \times R$ core array, $\mathbf{H}_k$ is a full $P \times Q$ slice of the extended core array; $\mathbf{W}$ and $\mathbf{W}_k$ are diagonal $S \times S$ matrices. Unless adorned with a tilde ($\sim$), $\mathbf{A} = (a_{ip})$, $\mathbf{B} = (b_{jq})$, and $\mathbf{C} = (c_{kr})$ are orthonormal. The scaling models are presented in their inner-product form.

These relationships are explained in some detail in [3, 33, 34], and [37].

Three-mode Factor Models

A stochastic version of Tucker's common three-mode factor model was first proposed by Bloxom [8], and this was further developed by Bentler and coworkers ([5, 6, 7], and [43]). Bloxom [9] discussed Tucker's factor models in term of higher-order composition rules. Much later the model was treated in extenso with many additional features by Oort ([46, 47]), while Kroonenberg and Oort [39] discuss the link between stochastic three-mode factor models and three-mode component models for what they call multimode covariance matrices. Multimode PLS models were developed by Lohmöller [45].

Parallel Factor Models

Parallel to the development of three-mode component analysis, Harshman ([25, 26]) conceived a component model which he called the *parallel factors model* – (PARAFAC). He conceived this model as an extension of regular component analysis and, using the parallel proportional profiles principle proposed by Cattell [18], he showed that the model solved the rotational indeterminacy of ordinary two-mode **principal component analysis**.

At the same time, Carroll and Chang [15] proposed the same model, calling it canonical decomposition (CANDECOMP). However, their development was primarily related to individual differences scaling, and their main contribution was to algorithmic aspects of the model without further developing the full potential for the analysis of 'standard' three-way arrays. This is the main reason why in this article the model is consistently referred to as the Parafac model.

A full-blown exposé of the model and some extensions is contained in [27] and [28], a more applied survey can be found in [29], and a tutorial with a chemical slant in [11]. The Parafac model has seen a large upsurge in both theoretical development and applications, when it was realized in (analytical) chemistry that physical models of the same form were encountered frequently and the parameter estimation of these models could be solved by the Parafac/Candecomp algorithm; for details see the book by Smilde, Geladi, and Bro [52].

Three-mode Models for Categorical Data

Van Mechelen and coworkers have developed a completely new paradigm for tackling binary three-mode data using Boolean algebra to construct models and express relations between parameters, see especially [19]. Another important approach to handling categorical data was presented by Sands and Young [50] who used optimal scaling of the categorical variables in conjunction with the Parafac model. Large three-way **contingency tables** have been tackled with three-mode **correspondence analysis** [13] and association models [2].

Hierarchies of Three-mode Component Models

Hierarchies of three-way models from both the French and Anglo-Saxon literature, which include the Tucker2 and Tucker3 models respectively, have been presented by Kiers ([33, 34]).

Individual Differences Scaling Models

The work of Carroll and Chang [15] on individual differences multidimensional scaling, or the INDSCAL model, formed a milestone in three-way analysis, and by early 2004 their paper was cited around 1000 times in the journal literature. They extended existing procedures for two-way data, mostly symmetric summed similarity matrices, to three-way data, building upon less far-reaching earlier work of Horan [30]. Over the years, various relatives of this model have been developed, and important from this article's point of view are the IDIOSCAL model [16] a less restricted variant of INDSCAL, the Parafac model which can be interpreted as an asymmetric INDSCAL model and the Tucker2 model which also belongs to the class of individual differences models of which it is the most general representative. The INDSCAL model has also been called a '*generalized subjective metrics model*' [50]. Other similar models have been developed within the context of **multidimensional scaling**, and general discussions of individual differences models and their interrelationships can, for instance, be found in [3] and [56] and their references.

Three-mode Cluster Models

Carroll and Arabie [14] developed a clustering version of the INDSCAL model with a set of common clusters with each sample or subject in the third mode having individual weights associated with these clusters. The procedure was called *individual differences clustering*–INDCLUS and is applied to sets of similarity matrices. Within that tradition, several further models were suggested, including some (ultrametric) **tree models** (see, for example, [17] and [22]).

Sato and Sato [51] presented a **fuzzy clustering method** for three-mode data by treating the problem as a multicriteria optimization problem and searching for a Pareto efficient solution. Coppi [21] is another contribution to this area.

On the basis of multivariate modeling using **maximum likelihood estimation** Basford [4] developed a mixture method approach to clustering three-mode continuous data and this approach has seen considerable application in agriculture. Extensions to categorical data can be found in Hunt and Basford [31], while further contributions in this vein have been made by Rocci and Vichi [48].

Other Three-way and Three-mode Models and Techniques

In several other fields, three-mode and three-way developments have taken place such as **unfolding**, block models, **longitudinal data**, clustering trajectories, conjoint analysis, PLS modeling, and so on, but an overview will not be given here. In France, several techniques for three-mode analysis have been developed, especially STATIS [41] and AFM [23] which are being used in francophone and Mediterranean countries, but not much elsewhere. An extensive bibliography on the website of the Three-Mode Company contains references to most of the papers dealing with three-mode and three-way issues (http:\\three-mode.leidenuniv.nl\).

Multiway and Multimode Models

Several extensions now exist generalizing three-mode techniques to multiway data. The earliest references are probably [15] for multiway CANDECOMP, [40] for generalizing Tucker's nonleast squares solution to the Tucker4 model, and [32] for generalizing the least-squares solution to analyze the Tucker*n* model. The book [52] contains the references to the many developments that have taken place especially in chemometrics with respect to multiway modeling.

Detailed Examples

In this section, we will present two examples of the analysis of three-way data: one application of the Tucker3 model and one application of individual differences scaling.

A Tucker3 Component Analysis: Stress and Coping at School

Coping Data: Description. The small example data set to demonstrate three-mode data consists of the ratings of 14 selected Dutch primary school children (mean age 9.8 years). On the basis of the results of a preliminary three-mode analysis, these 14 children were selected from a larger group of 390 children so that they were maximally different from each other and had relatively clear structures in the analysis (for a full account, see [49]).

The children were presented with a questionnaire describing six situations: Restricted in class by the teacher (*TeacherNo*), restricted at home by the mother (*MotherNo*), too much work in class (*WorkLoad*), class work was too difficult (*TooDifficult*), being bullied at school (*Bullied*) and not being allowed to participate in play with the other children (*NotParticipate*). For each of these situations, the children had to indicate what they felt (*Emotions*: Sad, Angry, Annoyed, Afraid) and how they generally dealt with the situation (*Strategies*: Avoidance Coping, Approach Coping, Seeking Social Support, Aggression). The data set has the form of 6 situations by 8 emotions and strategies by 13 children.

The specific interest of the present study is whether children have a different organization of the interrelations between situations, and emotions & strategies. In particular, we assume that there is a single configuration of situations and a single configuration of emotions & strategies, but not all children use the same strategies and do not have the same emotions when confronted with the various situations. In terms of the analysis model, we assume that the children may rotate and stretch or shrink the two configurations before they are combined. In other words, the combined configuration of the situations and the emotions & strategies may look different from one child to another.

The Tucker3 model and fit to the data. The Tucker3 model (see Table 1) has component matrices for each of the three modes **A**, **B**, and **C**, and a core array $\underline{\mathbf{G}}$ which contains information about the strength of the relationships of the components from the three modes. In particular, in the standard form of the model, g^2_{pqr} indicates the explained variability of the pth components of the children, the qth component of the emotions & strategies and the rth component of the situations (see Figure 6, Panel 1).

The model chosen for this example was the 3 × 3 × 2-model with 3 children components, 3 emotion & strategy components, and 2 situation components with a relative fit of 0.45. The components of the modes account for 0.27 0.11, and 0.07 (children), 0.26, 0.12, and 0.07 (emotions & strategies), and 0.29 and 0.16 (situations); all situations fit reasonably (i.e., about average = 0.45) except for restricted by the mother (*MotherNo*, relative fit = 0.18). The standardized residuals of the emotions and strategies were about equal, but the variability in Afraid, Social Support, and Aggression was not well fitted compared to the average of 0.45 (0.17, 0.10, and 0.19, respectively). The relative fit of most children was around the average, except that children 6 and 14 fitted rather badly (0.14 and 0.08, respectively).

Interpretation. *Individual differences between children.* The 14 children were children were selected to show individual differences and these differences are evident from the graphs of the subject space: Component 1 versus 2 (Figure 4) and Components 1 versus 3 (Figure 5). To get a handle on the kind of individual differences the scores of the children on the components were correlated with background variables available. The highest correlations for the first component were with 'Happy in school', 'Quality of relationship with the teacher', 'Having a good time at school' (average about 0.65). On the whole for most children except for 1 and 6 who have negative values on the first component, their scores on the first component go together with positive scores on general satisfaction with school. The second component correlated with 'Not ill versus ill' (however, only 10 and 1 were ill at the time) and Emotional support by the teacher (about 0.55 with 3,6,7 low scores and 1,8,13 high ones). Finally, the third component correlated 0.70 with Internalizing problem behavior (a CBCL scale–see [1]), but only 8 of the 14 have valid scores on this variable, so that this cannot be taken too seriously. The height of the correlations makes that we can use these variables for interpretation of the biplots as shown below, but correlations based on 14 scores with missing values on some of them do not make for very strong statements.

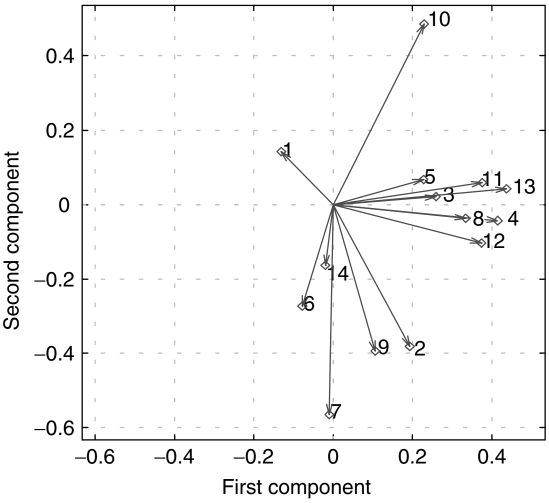


Figure 4 Coping data: three-dimensional children space: 1st versus 2nd Component

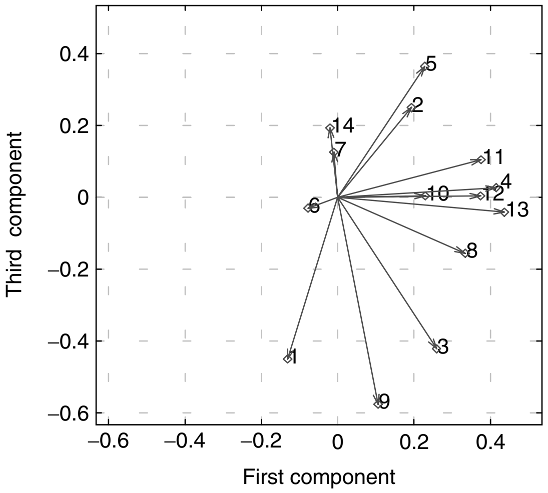


Figure 5 Coping data: three-dimensional children space: 1st versus 3rd Component

Even though the correlations are all significant, one should not read too much into them with respect to generalizing to the sample (and population) from which they were drawn, as the children were a highly selected set from the total group, but it serves to illustrate that appropriate background information can be used to enhance the interpretability of the subject space.

How children react differently in different situations. The whole purpose of the analysis is to see whether and how the children use different coping strategies and have different emotions in the situation presented to them. To evaluate this, we may look at joint biplots (see [37], Chap. 6). Figure 6 gives a brief summary of their construction. For, say, the rth component of the third mode C, the corresponding core slice $\mathbf{G}_r$ is divided between the other two modes A and B to construct the coordinates for the joint **biplot**. In particular, $\mathbf{A}^*$ and $\mathbf{B}^*$ are computed using a singular value decomposition of the core slice $\mathbf{G}_r = \mathbf{U}_r\mathbf{\Lambda}_r\mathbf{V}_r$.

Evaluating several possibilities, it was decided to use the Situation mode as reference mode (i.e., mode C in Figure 6), so that the Children and Emotions & Strategies appear in the joint biplot. For the first situation component the three joint biplot dimensions accounted for 21.8%, 7.2%, and 0.4% respectively so that only the first two biplot dimensions needed to be portrayed, and the same was true for the second situation component where the three joint biplot dimensions accounted for 10.3%, 5.2%, and 0.0001%, respectively.

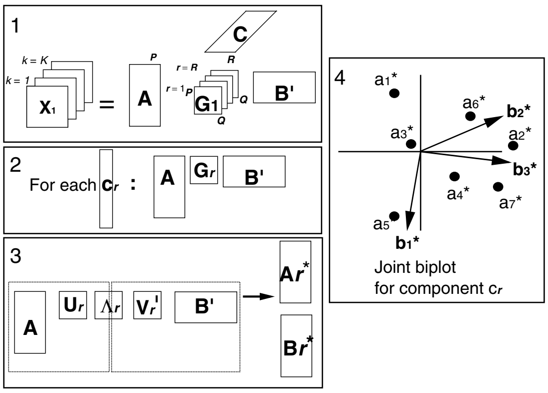


Figure 6 Joint biplot construction for the rth component of the third mode C

To facilitate interpretation, two joint biplots are presented for the *first situation* component. One plot (Figure 7) for the situations loading positively on the component (Being bullied and Not being allowed to participate), and one plot (Figure 8) for the situations loading negatively on the component (Class work too difficult and Too much work in class). This can be achieved by mirroring one of the modes around the origin. Here, this was done for the emotions & strategies. The origin in this graph is the estimated mean scores for all emotions & strategies and a zero

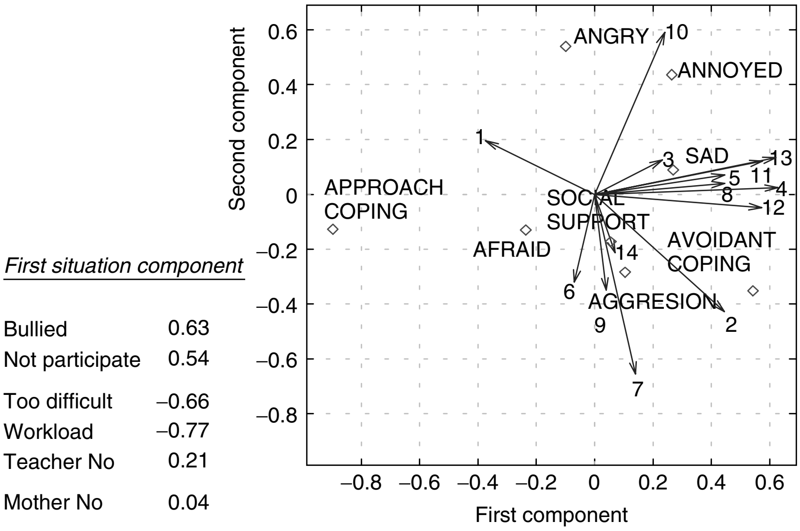
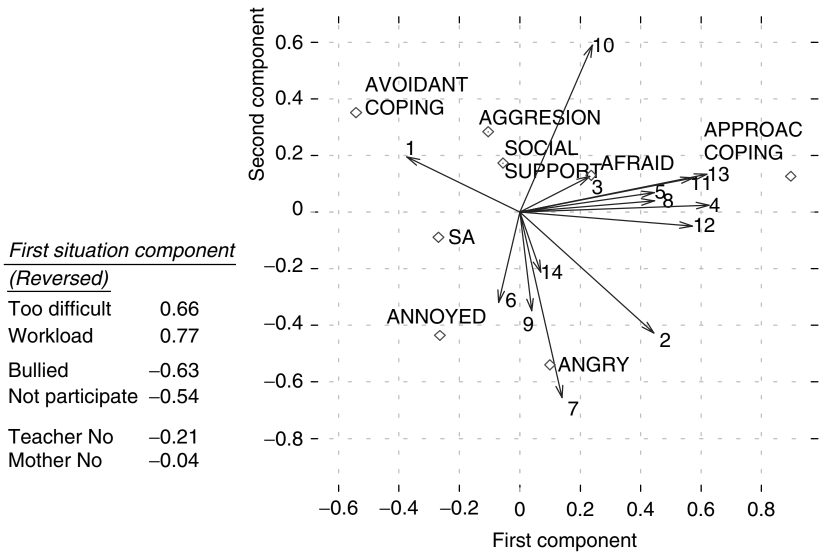


Figure 7 Coping Data: Joint biplot plot for bullying and not being allowed to participate (situations with positive scores on first situation component)



First situation component (Reversed)	
Too difficult	0.66
Workload	0.77
Bullied	−0.63
Not participate	−0.54
Teacher No	−0.21
Mother No	−0.04

Figure 8 Coping data: joint biplot plot for class work too difficult and too much work in class (Situations with negative scores on first situation component)

value on the component of the situations represent the estimated mean of that situation. To interpret the plot, the projections of children on the emotions & strategies are used, with high positive values indicating that the child has comparatively high scores for such an emotion (strategy). The situations weight the values of these projections, so that it can be established whether the child uses a particular emotion in a particular situation relative to other situations.

Bullying and not being allowed to participate (1st situation component). Most children, especially 5, 8, 9, 11, 12, and 13, use an avoidant coping strategy comparatively more often than an approach coping strategy when bullied or being left out (Figure 7). From the external variables we know that these are typically the children who are happy at school and who do well. Only child 1 (which is not so happy at school and does not do so well) seems to do the reverse, that is, using an approach coping rather than an avoidant coping strategy. Child 10 who according to the external variables is more ill than the others, is particularly angry and annoyed in such situations, while 2 and 7 resort towards aggressive behavior. Large differences with respect to social support are not evident.

Class work too difficult and too much work in class (1st situation component). When faced with too difficult or too much class work, the well-adjusted children, especially 5, 8, 9, 11, 12, and 13, use an approach coping strategy comparatively more often than an avoidant coping strategy, and only 1 seems to do the reverse, that is, using an avoidant rather than an approach coping strategy (Figure 8). Child 10 who is more ill than the others, is somewhat aggressive in such situations, while the more robust children 2 and 7 are particularly angry and rather annoyed. Large differences with respect to social support are not evident.

Restricted by the teacher and the mother (2nd situation component). These two situations have high loadings on the second situation components and not on the first, so that the joint plot associated with this component should be inspected. For the children who are happy at school and do well (i.e., 3, 4, 8, 12, 13), being restricted by the teacher and to a lesser extent by the mother is particularly met with avoidance coping, 10 and 11 are comparatively angry as well, 5 seeks some social support and is relatively angry, afraid, and sad (Figure 9). Child 14 is fairly unique in that it uses more approach coping, but it is comparatively sad and annoyed as well. Children 2, 6, and 7 are primarily more

Second situation component

Teacher no	0.67
Mother no	0.39
Workload	–0.10
Not participate	–0.27
Too difficult	–0.27
Bullied	–0.42

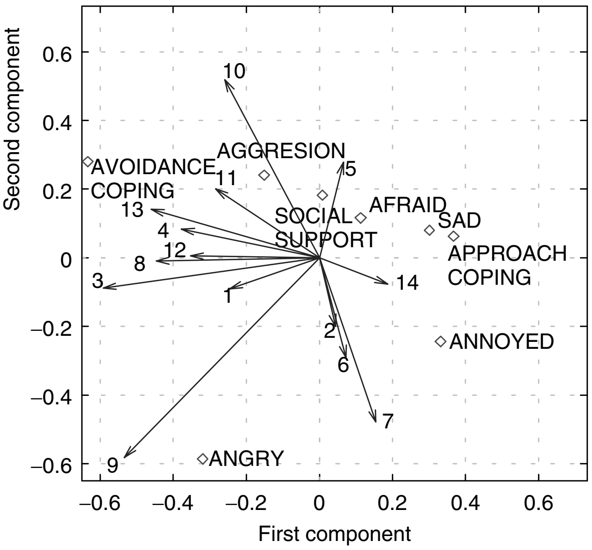


Figure 9 Coping data: Joint biplot plot for class work too difficult and Too much work in class. (Situations with positive scores on second situation component)

angry and annoyed but do not favor one particular strategy over an another, and finally child 9's reaction is primarily one of anger over any other emotions.

Note that being bullied also has sizable (negative) loading on the second situation dimension, indicating that being bullied is more complex than is shown in Figure 7 and the reverse pattern from the one discussed for being restricted by the teacher and the mother is true for bullying. To get a complete picture for bullying the information of the two joint plots should be combined, which is not easy to do. In a paper especially devoted to the substantive interpretation of the data, one would probably search for a rotation of the situation space such that bullying loads only on one component and interpret the associated joint plot especially for bullying; however, this will not be pursued here.

INDSCAL and IDIOSCAL: Typology of Pain

Pain data: Description. In this example, subjects were requested to indicate the similarities between certain pain sensations. The question is whether the subjects perceived pain in a similar manner, and in which way and to what extent pain sensations were considered as being similar.

The similarities were converted to dissimilarities to make them comparable to distances, but we will treat the dissimilarities as *squared* distances. As is shown in the MDS-literature, double-centering squared distances gives scalar products which can be analyzed by scalar-product models, such as INDSCAL and IDIOSCAL (see [15] and also Table 1). The INDSCAL model assumes that there exists a common stimulus configuration, which is shared by all judges (subjects), and that this configuration has the same structure for all judges, except that they may attach different importance (salience) to each of the (fixed) axes of the configuration. This results in some judges having configurations which are stretched out more along one of the axes. The IDIOSCAL model is similar except that each judge may rotate the axes of the common configuration over an angle before stretching.

Apart from the investigation into pain perception, we were also interested in examining for this data set Arabie, Carroll, and De Sarbo's [3] claim that IDIOSCAL '[..] has empirically yielded disappointing results in general' (p. 45) by comparing the IDIOSCAL and INDSCAL models.

Results. On the basis of a preliminary analysis, 16 of the 41 subjects were chosen for this example on the basis of the fit of the IDIOSCAL model to their data. The analysis reported here is a two-component

Table 2 Pain Data: IDIOSCAL subject weights and cosines and INDSCAL subject weights (sorted with respect to INDSCAL weights)

Type of	IDIOSCAL			IDIOSCAL	INDSCAL	
	subject weights			cosines	subject weights	
subject	(1,1)	(2,2)	(1,2)	(1,2)	(1,1)	(2,2)
Control	0.71	0.15	−0.27	−0.82	**0.45**	0.03
Chronic Pain	0.59	0.20	−0.22	−0.64	**0.38**	0.05
Chronic Pain	0.66	0.25	−0.18	−0.46	**0.37**	0.13
Control	0.53	0.22	−0.15	−0.44	**0.31**	0.09
Control	0.66	0.09	−0.03	−0.13	0.29	0.19
Chronic Pain	0.41	0.23	−0.24	−0.79	0.28	0.02
Control	0.50	0.14	−0.06	−0.22	0.24	0.14
RSI Pain	0.42	0.32	0.27	0.75	0.07	**0.42**
RSI Pain	0.53	0.32	0.19	0.45	0.16	**0.38**
Chronic Pain	0.34	0.45	0.17	0.44	0.09	**0.36**
RSI Pain	0.52	0.24	0.15	0.44	0.16	**0.33**
RSI Pain	0.45	0.29	0.12	0.34	0.15	**0.31**
Chronic Pain	0.35	0.47	0.08	0.19	0.14	**0.30**
Control	0.51	0.30	0.08	0.22	0.19	**0.30**
Control	0.25	0.32	0.08	0.32	0.08	0.23
RSI Pain	0.52	0.09	0.04	0.16	0.22	0.19

IDIOSCAL solution with a fit of 39.2%, indicating that the data are very noisy. In Figure 10 violent pains, such as shooting, burning, cramping, intense pain are in the same region of the space, as are the less dramatic ones, such are mild, moderate, and annoying, and also tiring, miserable, and distressing form a group.

The subject weights are shown in the left-hand panel of Table 2. The table provides the weights allocated to the first dimension and to the second dimension as well as the 'individual orientation' expressed as a cosine between the two dimensions. Clearly the subjects fall in two groups, those that put the dimensions under an acute angle and those that put them at an obtuse angle. Proof enough for an individual differences of orientation scaling, it seems. However, one problem is that for identifiability of the model, it has to be assumed that the stimulus space is orthogonal. To check whether this was problematic we also performed an INDSCAL analysis. This analysis provided a fit of 38.3%, hardly worse than the previous analysis, and given that its interpretation is more straightforward it is clearly to be preferred. The additional complexity of the IDIOSCAL model was only apparent in this case and the results support the conclusion in [3].

In Figure 10, we have drawn the orientation of the two INDSCAL axes, which have an inner product of −0.31 and thus make an angle of 108 degrees. In the right hand panel of Table 2, we see the INDSCAL subject weights, which also show the two groups found earlier. Staying with the basic INDSCAL interpretation, we see that one group of subjects (1,2,3,10,11,12,13,14) tends to emphasize the axis of

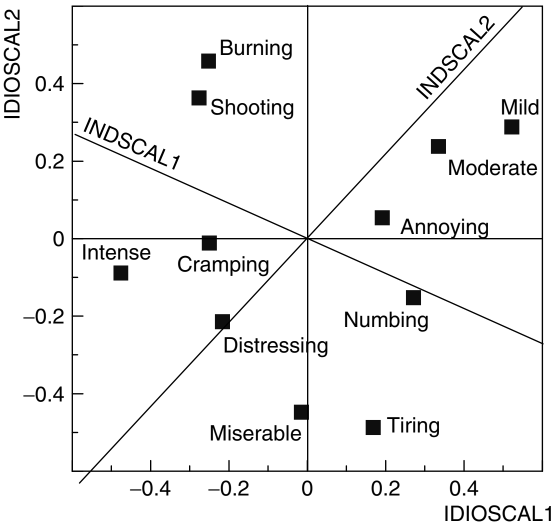


Figure 10 Grigg pain data: two-dimensional IDIOSCAL stimulus space with the best fitting INDSCAL axes

burning, shooting, intense, cramping pain in contrast with mild, numbing, and tiring. The other group of subjects (5,6,7,8,9,15,16) contrast mild and moderate pain with intense, tiring, distressing, and miserable, and place burning and shooting somewhere in the middle.

In the original design, the subjects consisted of three groups: chronic pain sufferers, repetitive-strain-injury sufferers, and a control group. If the information available is correct, then the empirical division into two groups runs right through two of the design groups, but all of the RSI sufferers are in the second group. In drawing conclusions, we have to take into consideration that the subjects in this example are a special selection from the real sample.

Further Reading

The earliest book length treatment of three-mode component models is [37], followed by [12], which emphasizes chemistry application, as does the recent book [52]. Multiway scaling models were extensively discussed in [3], and a recent comprehensive treatment of multidimensional scaling, which contains chapters on three-way scaling methods, is [10], while also [24] is an attractive recent book on the analysis of proximity data which also pays attention to three-way scaling.

Two collections of research papers on three-mode analysis have been published, which contain contributions of many people who were working in the field at the time. The first collection, [42], contains full-length overviews including the most extensive treatment of the Parafac model [27] and [28]. The second collection, [20], consists of papers presented at the 1988 conference on Multiway analysis in Rome. Finally, several special issues on three-mode analysis have appeared over the years: *Computational Analysis & Data Analysis, 18(1)* in 1994, *Journal of Chemometrics, 14(3)* in 2000, and *Journal of Chemometrics, 18(1)* in 2004.

Acknowledgments

This contribution was written while Fellow-in-Residence at the Netherlands Institute for Advanced Study in the Humanities and Social Sciences (NIAS). My thanks go to Irma Röder (Coping data) and Lynn Grigg (University of Queensland, Brisbane, Australia; Pain data) for their permission to use their data sets as examples. The last example is a reworked version of an earlier paper and is reproduced by permission of Springer Verlag, Berlin.

References

[1] Achenbach, T.M. & Edelbrock, C. (1983). *Manual for the Child Behavior Check List and revised child behavior profile*, Department of Psychiatry, University of Vermont, Burlington.

[2] Anderson, C.J. (1996). The analysis of three-way contingency tables by three-mode association models, *Psychometrika* **61**, 465–483.

[3] Arabie, P., Carroll, J.D. & DeSarbo, W.S. (1987). *Three-way scaling and clustering*, Sage Publications, Beverly Hills.

[4] Basford, K.E. & McLachlan, G.J. (1985). The mixture method of clustering applied to three-way data, *Journal of Classification* **2**, 109–125.

[5] Bentler, P.M. & Lee, S.-Y. (1978). Statistical aspects of a three-mode factor analysis model, *Psychometrika* **43**, 343–352.

[6] Bentler, P.M. & Lee, S.-Y. (1979). A statistical development of three-mode factor analysis, *British Journal of Mathematical and Statistical Psychology* **32**, 87–104.

[7] Bentler, P.M., Poon, W.-Y. & Lee, S.-Y. (1988). Generalized multimode latent variable models: Implementation by standard programs, *Computational Statistics and Data Analysis* **6**, 107–118.

[8] Bloxom, B. (1968). A note on invariance in three-mode factor analysis, *Psychometrika* **33**, 347–350.

[9] Bloxom, B. (1984). Tucker's three-mode factor analysis model, in *Research Methods for Multimode Data Analysis*, H.G. Law, C.W. Snyder Jr, J.A. Hattie, & R.P. McDonald, eds, Praeger, New York, pp. 104–121.

[10] Borg, I. & Groenen, P. (1997). *Modern Multidimensional Scaling: Theory and Applications*, Springer, New York. (Chapters 20 and 21).

[11] Bro, R. (1997). PARAFAC. Tutorial and applications, *Chemometrics and Intelligent Laboratory Systems* **38**, 149–171.

[12] Bro, R. (1998). Multi-way analysis in the food industry. Models, algorithms, and applications, PhD thesis, University of Amsterdam, Amsterdam.

[13] Carlier, A. & Kroonenberg, P.M. (1996). Decompositions and biplots in three-way correspondence analysis, *Psychometrika* **61**, 355–373.

[14] Carroll, J.D. & Arabie, P. (1983). INDCLUS: an individual differences generalization of the ADCLUS model and the MAPCLUS algorithm, *Psychometrika* **48**, 157–169.

[15] Carroll, J.D. & Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition, *Psychometrika* **35**, 283–319.

[16] Carroll, J.D. & Chang, J.J. (1972). IDIOSCAL: A generalization of INDSCAL allowing IDIOsyncratic reference system as well as an analytic approximation to INDSCAL, in *Paper Presented at the Spring Meeting of the Classification Society of North America*, Princeton, NJ.

[17] Carroll, J.D., Clark, L.A. & DeSarbo, W.S. (1984). The representation of three-way proximity data by single and multiple tree structure models, *Journal of Classification* **1**, 25–74.

[18] Cattell, R.B. & Cattell, A.K.S. (1955). Factor rotation for proportional profiles: analytical solution and an example, *The British Journal of Statistical Psychology* **8**, 83–92.

[19] Ceulemans, E., Van Mechelen, I. & Leenen, I. (2003). Tucker3 hierarchical classes analysis, *Psychometrika* **68**, 413–433.

[20] Coppi, R. & Bolasco, S., eds (1989). *Multiway Data Analysis*, Elsevier Biomedical, Amsterdam.

[21] Coppi, R. & D'Urso, P. (2003). Three-way fuzzy clustering models for LR fuzzy time trajectories, *Computational Statistics & Data Analysis* **43**, 149–177.

[22] De Soete, G. & Carroll, J.D. (1989). Ultrametric tree representations of three-way three-mode data, in *Multiway Data Analysis*, R., Coppi & S., Bolasco, eds, Elsevier Biomedical, Amsterdam, pp. 415–426.

[23] Escofier, B. & Pagès, J. (1994). Multiple factor analysis (AFMULT package), *Computational Statistics and Data Analysis* **18**, 121–140.

[24] Everitt, B.S. & Rabe-Hesketh, S. (1997). *The Analysis of Proximity Data*, Arnold, Londen, (Chapters 5 and 6).

[25] Harshman, R.A. (1970). Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multi-modal factor analysis, *UCLA Working Papers in Phonetics* **16**, 1–84.

[26] Harshman, R.A. (1972). Determination and proof of minimum uniqueness conditions for PARAFAC1, *UCLA Working Papers in Phonetics* **22**, 111–117.

[27] Harshman, R.A. & Lundy, M.E. (1984a). The PARAFAC model for three-way factor analysis and multidimensional scaling, in *Research Methods for Multimode Data Analysis*, H.G. Law, C.W. Snyder Jr, J.A. Hattie & R.P. McDonald, eds, Praeger, New York, pp. 122–215.

[28] Harshman, R.A. & Lundy, M.E. (1984b). Data preprocessing and the extended PARAFAC model, in *Research Methods for Multimode Data Analysis*, H.G. Law, C.W. Snyder Jr, J.A. Hattie & R.P. McDonald, eds, Praeger, New York, pp. 216–284.

[29] Harshman, R.A. & Lundy, M.E. (1994). PARAFAC: Parallel factor analysis, *Computational Statistics and Data Analysis* **18**, 39–72.

[30] Horan, C.B. (1969). Multidimensional scaling: combining observations when individuals have different perceptual structures, *Psychometrika* **34**, 139–165.

[31] Hunt, L.A. & Basford, K.E. (2001). Fitting a mixture model to three-mode three-way data with categorical and continuous variables, *Journal of Classification* **16**, 283–296.

[32] Kapteyn, A., Neudecker, H. & Wansbeek, T. (1986). An approach to n-mode components analysis, *Psychometrika* **51**, 269–275.

[33] Kiers, H.A.L. (1988). Comparison of "Anglo-Saxon" and "French" three-mode methods, *Statistique et Analyse Des Données* **13**, 14–32.

[34] Kiers, H.A.L. (1991). Hierarchical relations among three-way methods, *Psychometrika* **56**, 449–470.

[35] Kiers, H.A.L. (2000). Towards a standardized notation and terminology in multiway analysis, *Journal of Chemometrics* **14**, 105–122.

[36] Kroonenberg, P.M. (1983). Annotated bibliography of three-mode factor analysis, *British Journal of Mathematical and Statistical Psychology* **36**, 81–113.

[37] Kroonenberg, P.M. (1983). *Three-Mode Principal Component Analysis: Theory and Applications*, (*Errata, 1989; available from the author*). DSWO Press, Leiden.

[38] Kroonenberg, P.M. & De Leeuw, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms, *Psychometrika* **45**, 69–97.

[39] Kroonenberg, P.M. & Oort, F.J. (2003). Three-mode analysis of multi-mode covariance matrices, *British Journal of Mathematical and Statistical Psychology* **56**, 305–336.

[40] Lastovicka, J.L. (1981). The extension of component analysis to four-mode matrices, *Psychometrika* **46**, 47–57.

[41] Lavit, C., Escoufier, Y., Sabatier, R. & Traissac, P. (1994). The ACT (STATIS method), *Computational Statistics and Data Analysis* **18**, 97–119.

[42] Law, H.G. Snyder Jr, C.W., Hattie, J.A. & McDonald, R.P., eds (1984). *Research Methods for Multimode Data Analysis*, Praeger, New York.

[43] Lee, S.-Y. & Fong, W.-K. (1983). A scale invariant model for three-mode factor analysis, *British Journal of Mathematical and Statistical Psychology* **36**, 217–223.

[44] Levin, J. (1965). Three-mode factor analysis, *Psychological Bulletin* **64**, 442–452.

[45] Lohmöller, J.-B. (1989). *Latent Variable Path Modeling with Partial Least Squares*, Physica, Heidelberg.

[46] Oort, F.J. (1999). Stochastic three-mode models for mean and covariance structures, *British Journal of Mathematical and Statistical Psychology* **52**, 243–272.

[47] Oort, F.J. (2001). Three-mode models for multivariate longitudinal data, *British Journal of Mathematical and Statistical Psychology* **54**, 49–78.

[48] Rocci, R. & Vichi, M. (2003). Simultaneous component and cluster analysis: the between and within approaches, in *Paper presented at the International Meeting of the Psychometric Society*, Chia Laguna (Cagliary), Italy, 7–10 July.

[49] Röder, I. (2000). Stress in children with asthma. Coping and social support in the school context. PhD thesis, Department of Education, Leiden University, Leiden.

[50] Sands, R. & Young, F.W. (1980). Component models for three-way data: ALSCOMP3, an alternating least

squares algorithm with optimal scaling features, *Psychometrika* **45**, 39–67.

[51] Sato, M. & Sato, Y. (1994). On a multicriteria fuzzy clustering method for 3-way data, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **2**, 127–142.

[52] Smilde, A.K., Geladi, P. & Bro, R. (2004). *Multiway Analysis in Chemistry*, Wiley, Chicester.

[53] Tucker, L.R. (1963). Implications of factor analysis of three-way matrices for measurement of change, in *Problems in Measuring Change*, C.W., Harris, ed., University of Wisconsin Press, Madison, pp. 122–137.

[54] Tucker, L.R. (1964). The extension of factor analysis to three-dimensional matrices, in *Contributions to Mathematical Psychology*, H. Gulliksen & N. Frederiksen, eds, Holt, Rinehart and Winston, New York, pp. 110–127.

[55] Tucker, L.R. (1966). Some mathematical notes on three-mode factor analysis, *Psychometrika* **31**, 279–311.

[56] Young, F.W. & Hamer, R.M. (1987). *Multidimensional Scaling: History, Theory, and Applications*, Lawrence Erlbaum, Hillsdale. (reprinted 1994).

PIETER M. KROONENBERG

Thurstone, Louis Leon

Randall D. Wight and Philip A. Gable

Volume 4, pp. 2045–2046

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Thurstone, Louis Leon

Born: May 29, 1887, in Chicago.
Died: September 19, 1955, near Rapid City, Michigan.

Born in 1887 in Chicago to Swedish immigrants, Louis Leon Thunström began school in Berwyn, IL, moved with his parents to Centerville, MS, to Stockholm, Sweden, and back to Jamestown, NY – all before turning 14. Refusing to return to the United States without his books, Leon personally carried his three favorites onboard ship, among them Euclid. At 18, his name appeared in print for the first time in a *Scientific American* letter, suggesting how to manage the tension between Niagara Fall's tourists and its energy output. Shortly thereafter, to ease assimilation, his parents changed the spelling of the family name to Thurstone. After high school, Thurstone entered Cornell and studied engineering. As one of his undergraduate projects, Thurstone built – and later patented – a motion picture camera and projector that eliminated flicker. Those designs attracted the attention of Thomas Edison, who invited Thurstone to spend the summer following his 1912 master of engineering degree as an assistant in Edison's lab.

Thurstone became an engineering instructor at the University of Minnesota in the fall of 1912. While teaching, Thurstone pursued an interest in learning and enrolled in undergraduate experimental psychology courses. His interest and the inspired instruction he received prompted him to seek graduate study in psychology. In 1914 at age 27, he entered the University of Chicago. In 1915 and 1916, Thurstone accepted a graduate assistantship in applied psychology from Walter Bingham at the Carnegie Institute of Technology. In 1917, Thurstone received a Ph.D. from the University of Chicago, apparently without being in residence for at least two years. His dissertation, published in 1919, examined the learning curve equation. Thurstone joined the Carnegie faculty and advanced from assistant to full professor and to department chair. Between 1919 and 1923, Thurstone created psychometric instruments assessing aptitude, clerical skill, ingenuity, and intelligence.

Carnegie closed its applied psychology program in 1923, and Thurstone moved to Washington, D.C., to work for the Institute for Government Research, a foundation trying to improve civil service examinations. The American Council on Education (ACE) was located in the same Dupont Circle building as the foundation's office. Thurstone engaged the ACE's staff in conversation centering on creating college admission examinations, and in 1924, the ACE began to financially support Thurstone in that endeavor. The tests Thurstone developed in this initiative during the following years included linguistic and quantitative subscores and evolved into the Scholastic Aptitude Test, or SAT. The year 1924 also saw Thurstone's marriage to Thelma Gwinn and his accepting an offer to join the University of Chicago faculty.

Measurement theory drew Thurstone's attention during the early years at Chicago, 1924 through 1928. In contrast to psychophysical scales that related stimulation to experience, Thurstone developed theory and techniques for scaling psychological dimensions without physical referents, deriving accounts using dispersion as a unit of psychological measure [5]. Beginning in the late 1920s, this work grew into procedures that explored the structure of **latent variables** underlying the response patterns he found using his scaled instruments. Thurstone's influential concept of 'simple structure' (*see* **History of Factor Analysis: A Statistical Perspective**) guided the use of factor analysis in describing psychologically meaningful constructs. The development of multiple **factor analysis** is among his most widely known achievements. Notably, he reinterpreted 'g' in **Spearman's** theory of general intelligence as a special case of a multidimensional factor structure [3, 4].

In the coming years, Leon and Thelma employed factor analytic techniques to improve college entrance exams and to measure primary mental abilities. Their work bore much fruit, including providing the University of Chicago with the country's first credit by examination. Colleagues also recognized Thurstone's accomplishments. He was elected president of the American Psychological Association in 1932, elected charter president of the Psychometric Society in 1936, and elected to the National Academy of Sciences (NAS) in 1938. Shortly after the NAS election, Ernest Hilgard reports being a dinner guest in Thurstone's home and remembers Thurstone express surprise that as the son of immigrants, he (Thurstone) could successfully follow such a circuitous route to academic success.

During much of the twentieth century, Thurstone was a leading figure in psychometric and psychophysical theory and practice and in the investigation of attitudes, intelligence, skills, and values. His commitment to the scientific process is perhaps best seen in the weekly Wednesday evening seminars he conducted in his home – chalkboard and all – over the course of 30 years, often hosting 30 people at a time, conversing primarily over work in progress. After retiring from Chicago in 1952, he accepted a position at the University of North Carolina-Chapel Hill, where he established the L. L. Thurstone Psychometrics Laboratory. The home he and Thelma built near the campus included a seminar room with a built-in chalkboard where the seminar tradition continued unabated. On leave from Chapel Hill in the spring of 1954, Thurstone returned to Sweden as a visiting professor at the University of Stockholm, lecturing there and at other universities in northern Europe. This would be his last trip to the continent. In September 1955, Thurstone died at his summer home on Elk Lake in Michigan's upper peninsula (see [1], [2], [6] for more details of his life and work).

References

[1] Adkins, D. (1964). Louis Leon Thurston: creative thinker, dedicated teacher, eminent psychologist, in *Contributions to Mathematical Psychology*, N. Frederiksen & H. Gulliksen, eds, Holt Rinehart Winston, New York, pp. 1–39.

[2] Jones, L.V. (1998). L. L. Thurstone's vision of psychology as a quantitative rational science, in *Portraits of Pioneers in Psychology*, Vol. 3, G.A. Kimble & M. Werthheimer, eds, American Psychological Association, Washington.

[3] Thurstone, L.L. (1938). *Primary Mental Abilities*, University of Chicago Press, Chicago.

[4] Thurstone, L.L. (1947). *Multiple-Factor Analysis: A Development and Expansion of The Vectors of Mind*, University of Chicago Press, Chicago.

[5] Thurstone, L.L. & Chave, E.J. (1929). *The Measurement of Attitude*, University of Chicago Press, Chicago.

[6] Thurstone, L.L. (1952). In *A History of Psychology in Autobiography*, Vol. 4, E.G. Boring, H.S. Langfeld, H. Werner & R.M. Yerkes, eds, Clark University Press, Worcester, pp. 295–321.

Randall D. Wight and Philip A. Gable

Time Series Analysis

P.J. Brockwell

Volume 4, pp. 2046–2055

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Time Series Analysis

Time Series Data

A time series is a set of observations x_t, each one associated with a particular time t, and usually displayed in a *time series plot* of x_t as a function of t. The set of times T at which observations are recorded may be a discrete set, as is the case when the observations are recorded at uniformly spaced times (e.g., daily rainfall, hourly temperature, annual income, etc.), or it may be a continuous interval, as when the data is recorded continuously (e.g., by a seismograph or electrocardiograph). A very large number of practical problems involve observations that are made at uniformly spaced times. For this reason the, present article focuses on this case, indicating briefly how missing values and irregularly-spaced data can be handled.

Example 1 Figure 1 shows the number of accidental deaths recorded monthly in the US. for the years 1973 through 1978. The graph strongly suggests (as is usually the case for monthly data) the presence of a periodic component with period 12 corresponding to the seasonal cycle of 12 months, as well as a smooth trend accounting for the relatively slow change in level of the series, and a random component accounting for irregular deviations from a deterministic model involving trend and seasonal components only.

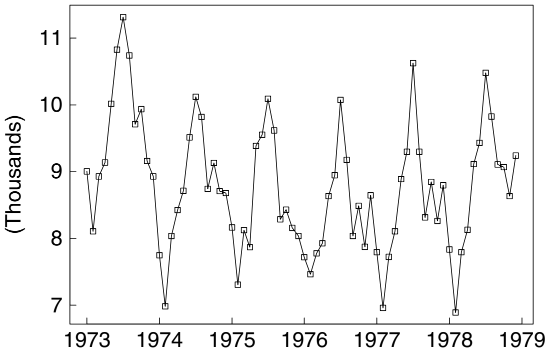


Figure 1 Monthly accidental deaths in the US from 1973 through 1978

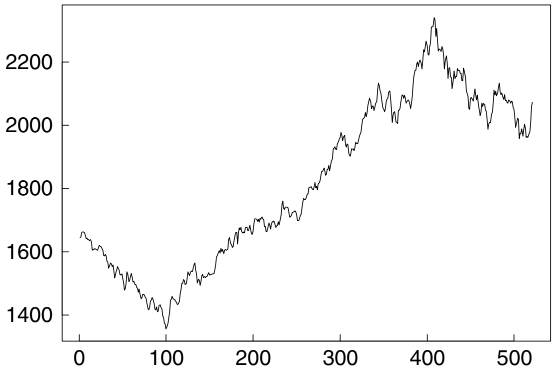


Figure 2 The Australian All-Ordinaries Index of Stock Prices on 521 successive trading days up to July 18, 1994

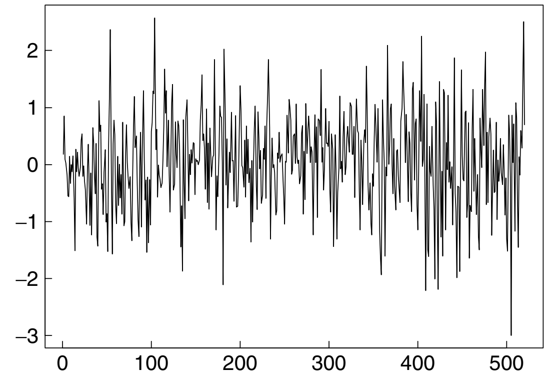


Figure 3 The daily percentage changes in the all-ordinaries index over the same period

Example 2 Figure 2 shows the closing value in Australian dollars of the Australian All-Ordinaries index (an average of 100 stocks sold on the Australian Stock Exchange) on 521 successive trading days, ending on July 18, 1994. It displays irregular variation around a rather strong trend. Figure 3 shows the daily percentage changes in closing value of the index for each of the 520 days ending on July 18, 1994. The trend apparent in Figure 2 has virtually disappeared, and the series appears to be varying randomly around a mean value close to zero.

Objectives

The objectives of time series analysis are many and varied, depending on the particular field of application. From the observations $x_1, \ldots, x_n$, we may wish to make inferences about the way in which the data are generated, to predict future values of the series, to detect a 'signal' hidden in noisy data, or simply to find a compact description of the available observations.

In order to achieve these goals, it is necessary to postulate a mathematical model (or family of models), according to which we suppose that the data is generated. Once an appropriate family has been selected, we then select a *specific* model by estimating model parameters and checking the resulting model for goodness of fit to the data. Once we are satisfied that the selected model provides a good representation of the data, we use it to address questions of interest. It is rarely the case that there is a 'true' mathematical model underlying empirical data, however, systematic procedures have been developed for selecting the best model, according to clearly specified criteria, within a broad class of candidates.

Time Series Models

As indicated above, the graph of the accidental deaths series in Figure 1 suggests representing x_t as the sum of a slowly varying *trend* component, a period-12 *seasonal* component, and a *random* component that accounts for the irregular deviations from the sum of the other two components. In order to take account of the randomness, we suppose that for each t, the observation x_t is just one of many possible values of a random variable X_t that we *might* have observed. This leads to the following *classical decomposition model* for the accidental deaths data,

$$X_t = m_t + s_t + Y_t, \quad t = 1, 2, 3, \ldots, \tag{1}$$

where the sequence $\{m_t\}$ is the *trend* component describing the long-term movement in the level of the series, $\{s_t\}$ is a *seasonal* component with known period (in this case, 12), and $\{Y_t\}$ is a sequence of random variables with mean zero, referred to as the *random* component. If we can characterize m_t, s_t, and Y_t in simple terms and in such a way that the model (1) provides a good representation of the data, then we can proceed to use the model to make forecasts or to address other questions related to the series. Completing the specification of the model by estimating the trend and seasonal components and characterizing the random component is a major part of time series analysis. The model (1) is sometimes referred to as an *additive* decomposition model (*see* **Additive Models**). Provided the observations are all positive, the *multiplicative* model,

$$X_t = m_t s_t Y_t, \tag{2}$$

can be reduced to an additive model by taking logarithms of each side to get an additive model for the logarithms of the data.

The general form of the additive model (1) supposes that the seasonal component s_t has known period d (12 for monthly data, 4 for quarterly data, etc.), and satisfies the conditions

$$s_{t+d} = s_t \text{ and } \sum_{t=1}^{d} s_t = 0, \tag{3}$$

while $\{Y_t\}$ is a *weakly stationary sequence of random variables*, that is, a sequence of random variables satisfying the conditions,

$$E(Y_t) = \mu, \ E(Y_t^2) < \infty \text{ and}$$
$$\mathrm{Cov}(Y_{t+h}, Y_t) = \gamma(h) \text{ for all } t, \tag{4}$$

with $\mu = 0$. The function γ is called the *autocovariance function* of the sequence $\{Y_t\}$, and the value $\gamma(h)$ is the *autocovariance at lag h*. In the special case when the random variables Y_t are independent and identically distributed, the model (1) is a classical regression model and $\gamma(h) = 0$ for all $h \neq 0$. However, in time series analysis, it is the dependence between Y_{t+h} and Y_t that is of special interest, and which allows the possibility of using past observations to obtain forecasts of future values that are better in some average sense than just using the expected value of the series. A measure of this dependence is provided by the autocovariance function. A more convenient measure (since it is independent of the origin and scale of measurement of Y_t) is the *autocorrelation function*,

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}. \tag{5}$$

From observed values $y_1, \ldots, y_n$ of a weakly stationary sequence of random variables $\{Y_t\}$, good estimators of the mean $\mu = E(Y_t)$ and the autocovariance function $\gamma(h)$ are the *sample mean* and *sample autocovariance function*,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{6}$$

and

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{i=1}^{n-|h|}(y_{i+|h|} - \hat{\mu})(y_i - \hat{\mu}), \quad -n < h < n, \tag{7}$$

respectively. The autocorrelation function of $\{Y_t\}$ is estimated by the *sample autocorrelation function*,

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}. \tag{8}$$

Elementary techniques for estimating m_t and s_t can be found in many texts on time series analysis (e.g., [6]). More sophisticated techniques are employed in the packages X-11 and the updated version X-12 described in [12], and used by the US Census Bureau. Once estimators $\hat{m}_t$ of m_t and $\hat{s}_t$ of s_t have been obtained, they can be subtracted from the observations to yield the **residuals**,

$$y_t = x_t - \hat{m}_t - \hat{s}_t. \tag{9}$$

A stationary time series model can then be fitted to the residual series to complete the specification of the model. The model is usually chosen from the class of *autoregressive moving average (or ARMA) processes*, defined below in ARMA Processes.

Instead of estimating and subtracting off the trend and seasonal components to generate a sequence of residuals, an alternative approach, developed by Box and Jenkins [4], is to apply *difference operators* to the original series to remove trend and seasonality. The *backward shift operator* B is an operator that, when applied to X_t, gives X_{t-1}. Thus,

$$BX_t = X_{t-1}, \quad B^j X_t = X_{t-j}, j = 2, 3, \ldots.$$

The *lag-1 difference operator* is the operator $\nabla = (1 - B)$. Thus,

$$\nabla X_t = (1 - B)X_t = X_t - X_{t-1}. \tag{10}$$

When applied to a polynomial trend of degree p, the operator ∇ reduces it to a polynomial of degree $p - 1$. The operator ∇^p, denoting p successive applications of ∇, therefore, reduces any polynomial trend of degree p to a constant. Usually, a small number of applications of ∇ is sufficient to eliminate trends encountered in practice. Application of the *lag-d difference operator*, $\nabla_d = (1 - B^d)$ (not to be confused with ∇^d) to X_t gives

$$\nabla_d X_t = (1 - B^d) = X_t - X_{t-d}, \tag{11}$$

eliminating any seasonal component with period d. In the Box–Jenkins approach to time series modeling, the operators ∇ and ∇_d are applied as many times as is necessary to eliminate trend and seasonality, and the sample mean of the differenced data subtracted to generate a sequence of residuals y_t, which are then modeled with a suitably chosen ARMA process in the same way as the residuals (9).

Figure 4 shows the effect of applying the operator ∇_{12} to the accidental deaths series of Figure 1. The seasonal component is no longer apparent, but there is still an approximately linear trend. Further application of the operator ∇ yields the series shown in Figure 5 with relatively constant level. This new series is a good candidate for representation by a stationary time series model.

The daily percentage returns on the Australian All-Ordinaries Index shown in Figure 2 already show no sign of trend or Seasonality, and can be modeled as a stationary sequence without preliminary detrending or deseasonalizing.

In cases where the variability of the observed data appears to change with the level of the data, a preliminary transformation, prior to detrending and deseasonalizing, may be required to stabilize

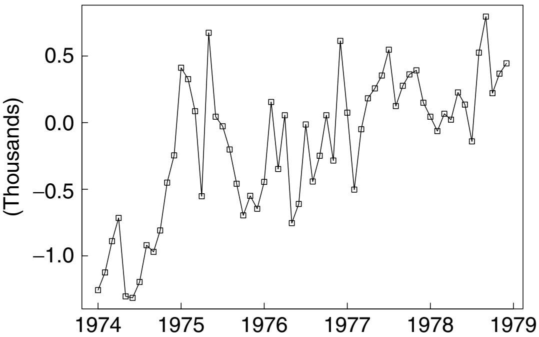


Figure 4 The differenced series $\{\nabla_{12}x_t, t = 13, \ldots, 72\}$ derived from the monthly accidental deaths $\{x_1, \ldots, x_{72}\}$ shown in Figure 1

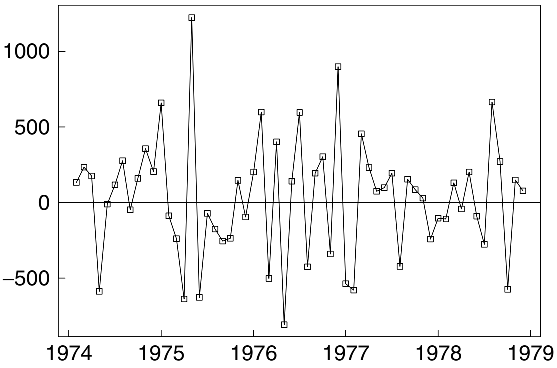


Figure 5 The differenced series $\{\nabla\nabla_{12}x_t, t = 14, \ldots, 72\}$ derived from the monthly accidental deaths $\{x_1, \ldots, x_{72}\}$ shown in Figure 1

the variability. For this purpose, a member of the family of *Box–Cox transformations* (see e.g., [6]) is frequently used.

ARMA Processes

For modeling the residuals $\{y_t\}$ (found as described above), a very useful parametric family of zero-mean stationary sequences is furnished by the *autoregressive moving average (or ARMA) processes*. The ARMA(p, q) process $\{Y_t\}$ with autoregressive coefficients, $\phi_1, \ldots, \phi_p$, moving average coefficients, $\theta_1, \ldots, \theta_q$, and white noise variance σ^2, is defined as a weakly stationary solution of the difference equations,

$$(1 - \phi_1 B - \cdots - \phi_p B^p)Y_t$$
$$= (1 + \theta_1 + \cdots + \theta_q B^q)Z_t, \quad t = 0, \pm 1, \pm 2, \ldots, \tag{12}$$

where B is the backward shift operator, the polynomials $\phi(z) = 1 - \phi_1 z - \cdots \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$ have no common factors, and $\{Z_t\}$ is a sequence of uncorrelated random variables with mean zero and variance σ^2. Such a sequence $\{Z_t\}$ is said to be *white noise with mean 0 and variance* σ^2, indicated more concisely by writing $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$.

The equations (12) have a unique stationary solution if, and only if, the equation $\phi(z) = 0$ has no root with $|z| = 1$, however, the possible values of $\phi_1, \ldots, \phi_p$ are usually assumed to satisfy the stronger restriction,

$$\phi(z) \neq 0 \text{ for all complex } z \text{ such that } |z| \leq 1. \tag{13}$$

The unique weakly stationary solution of equation (12) is then

$$Y_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \tag{14}$$

where ψ_j is the coefficient of z^j in the power-series expansion,

$$\frac{\theta(z)}{\phi(z)} = \sum_{j=0}^{\infty} \psi_j z^j, \quad |z| \leq 1.$$

Since Y_t in (14) is a function only of Z_s, $s \leq t$, the series $\{Y_t\}$ is said to be a *causal function* of $\{Z_t\}$, and the condition (13) is called the *causality condition* for the process (12). (Condition (13) is also frequently referred to as a *stability* condition.)

In the causal case, simple recursions are available for the numerical calculation of the sequence $\{\psi_j\}$ from the autoregressive and moving average coefficients $\phi_1, \ldots, \phi_p$, and $\theta_1, \ldots, \theta_q$ (see e.g., [6]). To every noncausal ARMA process, there is a causal ARMA process with the same autocovariance function, and, *under the assumption that all of the joint distributions of the process are multivariate normal*, with the same joint distributions. This is one reason for restricting attention to causal models. Another practical reason is that if $\{Y_t\}$ is to be simulated sequentially from the sequence $\{Z_t\}$, the causal representation (14) of Y_t does not involve *future* values Z_{t+h}, $h > 0$.

A key feature of ARMA processes for modeling dependence in a sequence of observations is the extraordinarily large range of autocorrelation functions exhibited by ARMA processes of different orders (p, q) as the coefficients $\phi_1, \ldots, \phi_p$, and $\theta_1, \ldots, \theta_q$, are varied. For example, if we take any set of observations, $x_1, \ldots, x_n$ and compute the sample autocorrelations, $\hat{\rho}(0)(= 1), \hat{\rho}(1), \ldots, \hat{\rho}(n - 1)$, then for any $k < n$, it is always possible to find a causal AR(k) process with autocorrelations $\rho(j)$ satisfying $\rho(j) = \hat{\rho}(j)$ for every $j \leq k$.

The mean of the ARMA process defined by (12) is $E(Y_t) = 0$, and the autocovariance function can be found from the equations obtained by multiplying each side of (12) by Y_{t-j}, $j = 0, 1, 2, \ldots$, taking expectations and solving for $\gamma(j) = E(Y_t Y_{t-j})$. Details can be found in [6].

Example 3 The causal ARMA(1,0) or AR(1) process is a stationary solution of the equations

$$Y_t = \phi Y_{t-1} + Z_t, \quad |\phi| < 1, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2).$$

The autocovariance function of $\{Y_t\}$ is $\rho(h) = \sigma^2 \phi^{|h|}/(1 - \phi^2)$.

Example 4 The ARMA$(0, q)$ or MA(q) process is the stationary series defined by

$$Y_t = \sum_{j=0}^{q} \theta_j Z_{t-j}, \quad \theta_0 = 1, \quad \{Z_t\} \sim \mathrm{WN}(0, \sigma^2).$$

The autocovariance function of $\{Y_t\}$ is $\gamma(h) = \sigma^2 \sum_{j=0}^{|h|} \theta_j \theta_{j+|h|}$ if $h \leq q$ and $\gamma(h) = 0$ otherwise.

A process $\{X_t\}$ is said to be an *ARMA process with mean* μ if $\{Y_t = X_t - \mu\}$ is an ARMA process as defined by equations of the form (12)).

In the following section, we consider the problem of fitting an ARMA model of the form (12), that is, $\phi(B)Y_t = \theta(B)Z_t$, to the residual series $y_1, y_2, \ldots$, generated as described in Time Series Models. If the residuals were obtained by applying the differencing operator $(1 - B)^d$ to the original observations $x_1, x_2, \ldots$, then we are effectively fitting the model

$$\phi(B)(1 - B)^d X_t = \theta(B)Z_t, \quad \{Z_t\} \sim \ \mathrm{WN}(0, \sigma^2) \tag{15}$$

to the original data. If the order of the polynomials $\phi(B)$ and $\theta(B)$ are p and q respectively, then the model (15) is called an *ARIMA(*p, d, q*) model* for $\{X_t\}$. If the residuals y_t had been generated by differencing also at some lag greater than 1 to eliminate seasonality, say, for example, $y_t = (1 - B^{12})(1 - B)^d x_t$, and the ARMA model (12) were then fitted to y_t, then the model for the original data would be a more general ARIMA model of the form,

$$\phi(B)(1 - B^{12})(1 - B)^d X_t = \theta(B)Z_t,$$
$$\{Z_t\} \sim \mathrm{WN}(0, \sigma^2). \tag{16}$$

Selecting and Fitting a Model to Data

In Time Series Models, we discussed two methods for eliminating trend and seasonality with the aim of transforming the original series to a series of residuals suitable for modeling as a zero-mean stationary series. In ARMA Processes, we introduced the class of ARMA models with a wide range of autocorrelation functions. By suitable choice of ARMA parameters, it is possible to find an ARMA process whose autocorrelations match the sample autocorrelations of the residuals $y_1, y_2, \ldots$, up to any specified lag. This is an intuitively appealing and natural approach to the problem of fitting a stationary time series model. However, except when the residuals are truly generated by a purely autoregressive model (i.e., an ARMA$(p, 0)$ model), this method turns out to give greater large-sample mean-squared errors than the method of *maximum Gaussian likelihood* described in the following paragraph.

Suppose for the moment that we know the orders p and q of the ARMA process (12) that is to be fitted to the residuals $y_1, \ldots, y_n$, and suppose that $\boldsymbol{\beta} = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)$, is the vector of parameters to be estimated. The Gaussian likelihood $L(\boldsymbol{\beta}; y_1, \ldots, y_n)$, is the likelihood computed under the assumption that the joint distribution from which $y_1, \ldots, y_n$ are drawn is multivariate normal (*see* **Maximum Likelihood Estimation**). Thus,

$$L(\boldsymbol{\beta}; y_1, \ldots, y_n) = (2\pi)^{-n/2} (\det \Gamma_n)^{-1/2} \times \exp\left(-\frac{1}{2} \mathbf{y}_n \Gamma_n^{-1} \mathbf{y}_n\right), \tag{17}$$

where $\mathbf{y}_n = (y_1, \ldots, y_n)'$, Γ_n is the matrix of autocovariances $\left[\gamma(i - j)\right]_{i,j=1}^{n}$, and $\gamma(h)$ is the autocovariance function of the model defined by (12). Although direct calculation of L is a daunting task, L can be reexpressed in the *innovations form* of Schweppe [22], which is readily calculated from the minimum mean-squared error one-step linear predictors of the observations and their mean-squared errors. These in turn can be readily calculated from the *innovations algorithm* (see [5]).

At first glance, maximization of Gaussian likelihood when the observations appear to be non-Gaussian may seem strange. However, if the noise sequence $\{Z_t\}$ in the model (12) is any independent identically distributed sequence (with finite variance), the large-sample joint distribution of the estimators (assuming the true orders are p and q) is the same as in the Gaussian case (see [15], [5]). This large-sample distribution has a relatively simple Gaussian form that can be used to specify large-sample confidence intervals for the parameters (under the assumption that the observations are generated by the fitted model).

Maximization of L with respect to the parameter vector $\boldsymbol{\beta}$ is a nonlinear optimization problem, requiring the use of an efficient numerical maximization algorithm (*see* **Optimization Methods**). For this reason, a variety of simpler estimation methods have been developed. These generally lead to less efficient estimators, which can be used as starting points for the nonlinear optimization. Notable among these are the Hannan–Rissanen algorithm for general ARMA processes, and the Yule–Walker and Burg algorithms for purely autoregressive processes.

The previous discussion assumes that the orders p and q are *known*. However, this is rarely, if ever, the case, and they must be chosen on the basis of the observations. The choice of p and q is referred to as the problem of *order selection*. The shape of the sample autocorrelation function gives some clue as to the order of the ARMA(p, q) model that best represents the data. For example, a sample autocorrelation function that appears to be roughly of the form $\phi^{|h|}$ for some ϕ such that $|\phi| < 1$ suggests (see Example 3) that an AR(1) model might be appropriate, while a sample autocorrelation function that is small in absolute value for lags $h > q$ suggests (see Example 4) that an MA(r) model with $r \le q$ might be appropriate.

A systematic approach to the problem was suggested by Akaike [1] when he introduced the *information criterion* known as AIC (*see* **Akaike's Criterion**). He proposed that p and q be chosen by minimizing AIC($\hat{\boldsymbol{\beta}}(p, q)$), where $\hat{\boldsymbol{\beta}}(p, q)$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ for fixed p and q and

$$\mathrm{AIC}(\boldsymbol{\beta}) = -2\ln(L(\boldsymbol{\beta})) + 2(p + q + 1). \qquad (18)$$

The term $2(p + q + 1)$ can be regarded as a *penalty factor* that prevents the selection of excessive values for p and q and the accumulation of additional parameter estimation errors. If the data are truly generated by an ARMA(p, q) model, it has been shown that the AIC criterion tends to overestimate p and q and is not consistent as the sample size approaches infinity. Consistent estimation of p and q can be obtained by using information criteria with heavier penalty factors such as the (Bayesian information criterion) BIC. Since, however, there is rarely a 'true' model generating the data, consistency is not necessarily an essential property of order selection methods. It has been shown in [23] that although the AIC criterion does not give consistent estimation of p and q, it is optimal in a certain sense with respect to prediction of future values of the series. A refined small-sample version of AIC, known as AICC, has been developed in [18], and a comprehensive account of model selection can be found in [7].

Having arrived at a potential ARMA model for the data, the model should be checked for goodness of fit to the data. On the basis of the fitted model, the minimum mean-squared error linear predictors $\hat{Y}_t$ of each Y_t in terms of Y_s, $s < t$, and the corresponding mean-squared errors s_t can be computed (see Prediction below). In fact, they are computed in the course of evaluating the Gaussian likelihood in its innovations form. If the fitted model is valid, the properties of the rescaled one-step prediction errors $(Y_t - \hat{Y}_t)/\sqrt{(s_t)}$ should be similar to those of the sequence Z_t/σ in the model (12), and can, therefore, be used to check the assumed white noise properties of $\{Z_t\}$ and whether or not the assumption of independence and/or normality is justified. A number of such tests are available (see e.g., [6], chap. 5.)

Prediction

If $\{Y_t\}$ is a weakly stationary process with mean, $E(Y_t) = \mu$, and autocovariance function, $\mathrm{Cov}(Y_{t+h}, Y_t) = \gamma(h)$, a fundamental property of conditional expectation tells us that the 'best' (minimum mean squared error) predictor of Y_{n+h}, $h > 0$, in terms of $Y_1, \ldots, Y_n$, is the conditional expectation $E(Y_{n+h}| Y_1, \ldots, Y_n)$. However, this depends in a complicated way on the joint distributions of the random variables Y_t that are virtually impossible to estimate on the basis of a single series of observations $y_1, \ldots, y_n$. However, if the sequence $\{Y_t\}$ is Gaussian, the best predictor of Y_{n+h} in terms of $Y_1, \ldots, Y_n$ is a linear function, and can be calculated as described below.

The best *linear* predictor of Y_{n+h} in terms of $\{1, Y_1, \ldots, Y_n\}$, that is, the linear combination $P_n Y_{n+h} = a_0 + a_1 Y_1 + \cdots + a_n Y_n$, which minimizes the mean squared error, $E[(Y_{n+h}-a_0-\cdots-a_nY_n)^2]$, is given by

$$P_nY_{n+h} = \mu + \sum_{i=1}^{n} a_i(Y_{n+1-i} - \mu), \qquad (19)$$

where the vector of coefficients $\mathbf{a} = (a_1, \ldots, a_n)'$ satisfies the linear equation,

$$\Gamma_n\mathbf{a} = \boldsymbol{\gamma}_n(h), \qquad (20)$$

with $\boldsymbol{\gamma}_n(h) = (\gamma(h), \gamma(h+1), \ldots, \gamma(h+n-1)$ and $\Gamma_n = [\gamma(i-j)]_{i,j=1}^{n}$. The mean-squared error of the best linear predictor is

$$E(Y_{n+h} - P_nY_{n+h})^2 = \gamma(0) - \mathbf{a}'\boldsymbol{\gamma}_n(h). \qquad (21)$$

Once a satisfactory model has been fitted to the sequence $y_1, \ldots, y_n$, it is, therefore, a straightforward but possibly tedious matter to compute best linear predictors of future observations by using the mean and autocovariance function of the fitted model and solving the linear equations for the coefficients $a_0, \ldots, a_n$. If n is large, then the set of linear equations for $a_0, \ldots, a_n$ is large, and, so, recursive methods using the Levinson–Durbin algorithm or the Innovations algorithm have been devised to express the solution for $n = k + 1$ in terms of the solution for $n = k$, and, hence, to avoid the difficulty of inverting large matrices. For details see, for example, [6], Chapter 2.

If an ARMA model is fitted to the data, the special linear structure of the ARMA process arising from the defining equations can be used to greatly simplify the calculation of the best linear h-step predictor P_nY_{n+h} and its mean-squared error, $\sigma_n^2(h)$. If the fitted model is Gaussian, we can also compute 95% prediction bounds, $P_nY_{n+h} \pm 1.96\sigma_n(h)$. For details see, for example, [6], Chapter 5.

Example 5 In order to predict future values of the causal AR(1) process defined in Example 3, we can make use of the fact that linear prediction is a linear operation, and that $P_nZ_t = 0$ for $t > n$ to deduce that

$$\begin{aligned} P_nY_{n+h} &= \phi P_nY_{n+h-1} = \phi^2 P_nY_{n+h-2} = \cdots \\ &= \phi^h Y_n, \quad h \geq 1. \end{aligned}$$

In order to obtain forecasts and prediction bounds for the *original* series that were transformed to generate the residuals, we simply apply the inverse transformations to the forecasts and prediction bounds for the residuals.

The Frequency Viewpoint

The methods described so far are referred to as ‘time-domain methods’ since they focus on the evolution in time of the sequence of random variables $X_1, X_2, \ldots$ representing the observed data. If, however, we regard the sequence as a random function defined on the integers, then an alternative approach is to consider the decomposition of that function into sinusoidal components, analogous to the Fourier decomposition of a deterministic function. This approach leads to the *spectral representation* of the sequence $\{X_t\}$, according to which every weakly stationary sequence has a representation

$$X_t = \int_{-\pi}^{\pi} \mathrm{e}^{i\omega t}\, \mathrm{d}Z(t), \qquad (22)$$

where $\{Z(t),\ -\pi \leq t \leq \pi\}$ is a process with uncorrelated increments. A detailed discussion of this approach can be found, for example, in [2], [5], and [21], but is outside the scope of this article. Intuitively, however, the expression (22) can be regarded as representing the random function $\{X_t,\ t = 0, \pm 1, \pm 2 \ldots\}$ as the limit of a linear combination of sinusoidal functions with uncorrelated random coefficients. The analysis of weakly stationary processes by means of their spectral representation is referred to as ‘frequency domain analysis’ or spectral analysis. It is equivalent to time-domain analysis, but provides an alternative way of viewing the process, which, for some applications, may be more illuminating. For example, in the design of a structure subject to a randomly fluctuating load, it is important to be aware of the presence in the loading force of a large sinusoidal component with a particular frequency in order to ensure that this is not a resonant frequency of the structure.

Multivariate Time Series

Many time series arising in practice are best analyzed as components of some vector-valued (multivariate) time series $\{\mathbf{X}_t\} = (X_{t1}, \ldots, X_{tm})'$ in which each of the component series $\{X_{ti},\ t = 1, 2, 3, \ldots\}$ is a univariate time series of the type already discussed. In multivariate time series modeling, the goal is to account for, and take advantage of, the dependence not only between observations of a single component at different times, but also between the different component series. For example, if X_{t1} is the daily percentage change in the closing value of the Dow-Jones Industrial Average in New York on trading day t, and if X_{t2} is the analogue for the Australian All-Ordinaries Index (see Example 2), then $\{\mathbf{X}_t\} = (X_{t1}, X_{t2})'$ is a bivariate time series in which there is very little evidence of autocorrelation in either of the two component series. However, there is strong evidence of correlation between X_{t1} and $X_{(t+1)2}$, indicating that the Dow-Jones percentage change on day t is of value in predicting the All-Ordinaries percentage change on day $t + 1$. Such dependencies in the multivariate case can be measured by the covariance matrices,

$$\Gamma(t + h, t) = \left[\mathrm{cov}(X_{(t+h),i}, X_{t,j})\right]_{i,j=1}^{m}, \qquad (23)$$

where the number of components, m, is two in this particular example. Weak stationarity of the multivariate series $\{\mathbf{X}_t\}$ is defined as in the univariate case to mean that all components have finite second moments and that the mean vectors $E(\mathbf{X}_t)$ and covariance matrices $\Gamma(t + h, t)$ are independent of t.

Much of the analysis of multivariate time series is analogous to that of univariate series, multivariate ARMA (or VARMA) processes being defined again by linear equations of the form (12) but with vector-valued arguments, and matrix coefficients. There are, however, some new important considerations arising in the multivariate case. One is the question of VARMA nonidentifiability. In the univariate case, an AR(p) process cannot be reexpressed as a finite order moving average process. However, this is not the case for VAR(p) processes. There are simple examples of VAR(1) processes that are also VMA(1) processes. This lack of identifiability, together with the large number of parameters in a VARMA model and the complicated shape of the likelihood surface, introduces substantial additional difficulty into maximum Gaussian likelihood estimation for VARMA processes. Restricting attention to VAR processes eliminates the identifiability problem. Moreover, the fitting of a VAR(p) process by equating covariance matrices up to lag p is asymptotically efficient and simple to implement using a multivariate version of the Levinson-Durbin algorithm (see [27], and [28]).

For nonstationary univariate time series, we discussed in the section 'Objectives' the use of differencing to transform the data to residuals suitable for modeling as zero-mean stationary series. In the multivariate case, the concept of *cointegration*, due to Granger [13], plays an important role in this connection. The m-dimensional vector time series $\{\mathbf{X}_t\}$ is said to be integrated of order d (or $I(d)$) if, when the difference operator ∇ is applied to each component $d - 1$ times, the resulting process is nonstationary, while if it is applied d times, the resulting series is stationary. The $I(d)$ process $\{\mathbf{X}_t\}$ is said to be cointegrated with cointegration vector $\boldsymbol{\alpha}$, if $\boldsymbol{\alpha}$ is an $m \times 1$ vector such that $\{\boldsymbol{\alpha}'\mathbf{X}_t\}$ is of order less than d. Cointegrated processes arise naturally in economics. [11] gives as an illustrative example the vector of tomato prices in Northern and Southern California (say X_{t1} and X_{t2}, respectively). These are linked by the fact that if one were to increase sufficiently relative to the other, the profitability of buying in one market and selling in the other would tend to drive the prices together, suggesting that although the two series separately may be nonstationary, the difference varies in a stationary manner. This corresponds to having a cointegration vector $\boldsymbol{\alpha} = (1, -1)'$. Statistical inference for multivariate models with cointegration is discussed in [10].

State-space Models

State-space models and the associated Kalman recursions have had a profound impact on time series analysis. A linear state-space model for a (possibly multivariate) time series $\{\mathbf{Y}_t, t = 1, 2, \ldots\}$ consists of two equations. The first, known as the *observation equation*, expresses the w-dimensional observation vector $\mathbf{Y}_t$ as a linear function of a v-dimensional state variable $\mathbf{X}_t$ plus noise. Thus,

$$\mathbf{Y}_t = G_t\mathbf{X}_t + \mathbf{W}_t, \quad t = 1, 2, \ldots, \qquad (24)$$

where $\{\mathbf{W}_t\}$ is a sequence of uncorrelated random vectors with $E(\mathbf{W}_t) = \mathbf{0}$, $\mathrm{cov}(\mathbf{W}_t) = R_t$, and $\{G_t\}$ is a sequence of $w \times v$ matrices. The second equation, called the **state equation**, determines the state $\mathbf{X}_{t+1}$ at time $t + 1$ in terms of the previous state $\mathbf{X}_t$ and a noise term. The state equation is

$$\mathbf{X}_{t+1} = F_t\mathbf{X}_t + \mathbf{V}_t, \quad t = 1, 2, \ldots, \qquad (25)$$

where $\{F_t\}$ is a sequence of $v \times v$ matrices, $\{\mathbf{V}_t\}$ is a sequence of uncorrelated random vectors with $E(\mathbf{V}_t) = \mathbf{0}$, $\mathrm{cov}(\mathbf{V}_t) = Q_t$, and $\{\mathbf{V}_t\}$ is uncorrelated with $\{\mathbf{W}_t\}$ (i.e., $E(\mathbf{W}_t\mathbf{V}_s') = 0$ for all s and t). To complete the specification, it is assumed that the initial state $\mathbf{X}_1$ is uncorrelated with all of the noise terms $\{\mathbf{V}_t\}$ and $\{\mathbf{W}_t\}$.

An extremely rich class of models for time series, including and going well beyond the ARIMA models described earlier, can be formulated within this framework (see [16]). In econometrics, the structural time series models developed in [17], in which trend and seasonal components are allowed to evolve randomly, also fall into this framework. The power of the state-space formulation of time series models depends heavily on the Kalman recursions, which allow best linear predictors and best linear estimates of various model-related variables to be computed in a routine way. Time series with missing values are also readily handled in the state-space framework.

More general state-space models, in which the linear relationships (24) and (25) are replaced by the specification of conditional distributions, are also widely used to generate an even broader class of models, including, for example, models for time series of counts such as the numbers of reported new cases of a particular disease (see e.g., [8]).

Additional Topics

Although linear models for time series data have found broad applications in many areas of the physical, biological, and behavioral sciences, there are also many areas where they have been found inadequate. Consequently, a great deal of attention has been devoted to the development of nonlinear models for such applications. These include threshold models, [25], bilinear models, [24], random-coefficient autoregressive models, [20]), Markov switching models, [14], and many others. For financial time Series, the ARCH (autoregressive conditionally heteroscedastic) model of Engle [9], and its generalized version, the GARCH model of Bollerslev [3], have been particularly successful in modeling financial returns data of the type illustrated in Figure 4.

Although the sample autocorrelation function of the series shown in Figure 4 is compatible with the hypothesis that the series is an independent and identically distributed white noise sequence, the sample autocorrelation functions of the absolute values and the squares of the series are significantly different from zero, contradicting the hypothesis of independence. ARCH and GARCH processes are white noise sequences that are nevertheless dependent. The dependence is introduced in such a way that these models exhibit many of the distinctive features (heavy-tailed marginal distributions and persistence of volatility) that are observed in financial time series.

The recent explosion of interest in the modeling of financial data, in particular, with a view to solving option-pricing and asset allocation problems, has led to a great deal of interest, not only in ARCH and GARCH models, but also to stochastic volatility models and to models evolving continuously in time. For a recent account of time series models specifically related to financial applications, see [26].

Apart from their importance in finance, continuous-time models provide a very useful framework for the modeling and analysis of discrete-time series with irregularly spaced data or missing observations. For such applications, see [19].

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory*, B.N. Petrov & F. Saki, eds, Akademia Kiado, Budapest, pp. 267–281.

[2] Bloomfield, P. (1976). *Fourier Analysis of Time Series: An Introduction*, John Wiley & Sons, New York.

[3] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **31**, 307–327.

[4] Box, G.E.P., Jenkins, G.M. & Reinsel, G.C. (1994). *Time Series Analysis; Forecasting and Control*, 3rd Edition, Prentice-Hall, Englewood Cliffs.

[5] Brockwell, P.J. & Davis, R.A. (1991). *Time Series: Theory and Methods*, 2nd Edition, Springer-Verlag, New York.

[6] Brockwell, P.J. & Davis, R.A. (2002). *Introduction to Time Series and Forecasting*, 2nd Edition, Springer-Verlag, New York.

[7] Burnham, K.P. & Anderson, D. (2002). *Model Selection and Multi-Model Inference*, 2nd Edition, Springer-Verlag, New York.

[8] Chan, K.S. & Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving counts, *Journal American Statistical Association* **90**, 242–252.

[9] Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation, *Econometrica* **50**, 987–1087.

[10] Engle, R.F. & Granger, C.W.J. (1987). Co-integration and error correction: representation, estimation and testing, *Econometrica* **55**, 251–276.

[11] Engle, R.F. & Granger, C.W.J. (1991). *Long-Run Economic Relationships*, Advanced Texts in Econometrics, Oxford University Press, Oxford.

[12] Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C. & Chen, B.-C. (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program (with discussion and reply), *The Journal of Business Economic Statistics* **16**, 127–177.

[13] Granger, (1981). Some properties of time series data and their use in econometric model specification, *Journal of Econometrics* **16**, 121–130.

[14] Hamilton, J.D. (1994). *Time Series Analysis*, Princeton University Press, Princeton.

[15] Hannan, E.J. (1973). The asymptotic theory of linear time series models, *Journal of Applied Probability* **10**, 130–145.

[16] Hannan, E.J. & Deistler, M. (1988). *The Statistical Theory of Linear Systems*, John Wiley & Sons, New York.

[17] Harvey, A.C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

[18] Hurvich, C.M. & Tsai, C.L. (1989). Regression and time series model selection in small samples, *Biometrika* **76**, 297–307.

[19] Jones, R.H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations, *Technometrics* **22**, 389–395.

[20] Nicholls, D.F. & Quinn, B.G. (1982). *Random Coefficient Autoregressive Models: An Introduction*, Springer Lecture Notes in Statistics, p. 11.

[21] Priestley, M.B. (1981). *Spectral Analysis and Time Series*, Vols. I and II, Academic Press, New York.

[22] Schweppe, F.C. (1965). Evaluation of likelihood functions for Gaussian signals, *IEEE Transactions on Information Theory* **IT-11**, 61–70.

[23] Shibata, R. (1980). Asymptotically efficient selection of the order for estimating parameters of a linear process, *Annals of Statistics* **8**, 147–164.

[24] Subba-Rao, T. & Gabr, M.M. (1984). *An Introduction to Bispectral Analysis and Bilinear Time Series Models*, Springer Lecture Notes in Statistics, p. 24.

[25] Tong, H. (1990). *Non-linear Time Series: A Dynamical Systems Approach*, Oxford University Press, Oxford.

[26] Tsay, (2001). *Analysis of Financial Time Series*, John Wiley & Sons, New York.

[27] Whittle, P. (1963). On the fitting of multivariate autoregressions and the approximate canonical factorization of a spectral density matrix, *Biometrika* **40**, 129–134.

[28] Wiggins, R.A. & Robinson, E.A. (1965). Recursive solution to the multichannel filtering problem, *Journal of Geophysics Research* **70**, 1885–1891.

(*See also* **Longitudinal Data Analysis**; **Repeated Measures Analysis of Variance**)

P.J. BROCKWELL

Tolerance and Variance Inflation Factor

Jeremy Miles

Volume 4, pp. 2055–2056

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Tolerance and Variance Inflation Factor

In regression analysis (*see* **Regression Models**), one outcome is regressed onto one or more predictors in order to explain the variation in the outcome variable. Where there is more than one predictor variable, the relationships between the predictors can affect both the regression estimate and the standard error of the regression estimate (*see* **Multiple Linear Regression**).

In its original usage, multicollinearity referred to a perfect linear relationship between the independent variables, or a weighted sum of the independent variables; however, it is now used to refer to large (multiple) correlations amongst the predictor variables, known as **collinearity**.

The effect of collinearity is to increase the standard error of the regression coefficients (and hence to increase the confidence intervals and decrease the P values).

The standard error of a regression estimate of the variable j $(\hat{\beta}_j)$ is given by

$$se(\hat{\beta}_j) = \sqrt{\frac{\sigma^2}{\Sigma x_j^2} \times \frac{1}{1 - R_j^2}} \qquad (1)$$

where R_j^2 is the R^2 found when regressing all other predictors onto the predictor j. (Note that when there is only one variable in the regression equation, or when the correlation between the predictors is equal to zero, the value for the part of the equation $1/(1 - R_j^2)$ is equal to 1.) The term $1/(1 - R_j^2)$ is known as the *variance inflation factor* (VIF). When the correlation changes from 0 (or when additional variables are added), the value of the VIF increases, and the value of the standard error of the regression parameter increases with the square root of the VIF.

The reciprocal of the VIF is called the *tolerance*. It is equal to $1 - R_j^2$, where each predictor is regressed on all of the other predictors in the analysis.

A rule of thumb that is sometimes given for the tolerance and the VIF is that the tolerance should not be less than 0.1, and that therefore the VIF should not be greater than 10, although this is dependent on other factors, not least the sample size.

Further information on these measures can be found in [1].

Reference

[1] Fox, J. (1991). *Regression Diagnostics*, Sage Publications, Newbury Park.

JEREMY MILES

Transformation

Neal Schmitt

Volume 4, pp. 2056–2058

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Transformation

A transformation is any systematic alteration in a set of scores such that some characteristics of the scores are changed while other characteristics remain unchanged. Transformations of a set of scores are done for several reasons. Statistically, transformations are done to provide a data set that possesses characteristics that make it acceptable for a variety of parametric statistical procedures. In **analysis of variance** applications, a data set that is normally distributed and that possesses homogeneity of variance across treatment conditions is necessary to draw appropriate inferences about mean differences between conditions. Normality of the distribution of a set of scores and equality in the scale units on a measure is a necessary precondition for the application of the **multivariate analyses** (e.g., **factor analyses**, **structural equation modeling**) routinely applied in the social and behavioral sciences. When data are not normally distributed, transformations that produce distributions that more closely approximate normality can be used, provided it is reasonable to assume that the underlying construct being measured is normally distributed. Transformations are also done frequently to provide data that are more easily communicated to the consuming public (e.g., percentiles, IQ scores).

The data we use in the behavioral sciences involve the assignment of numbers to objects that have some meaning in terms of the objects' physical properties. In considering the use of statistical analyses to summarize or analyze data, it is important that the transformations involved in these analyses and summaries do not alter the meaning of the basic properties to which they refer. **Stevens** [1] is credited with providing the widely accepted classification of scales into nominal, ordinal, interval, and ratio types (*see* **Scales of Measurement**). A nominal scale is a measurement that we use to categorize objects into discrete groups (e.g., gender, race, political party). In the case of nominal data, any one-to-one transformation that retains the categorization of individual cases into discrete groups is permissible. Typical summary statistics in this instance are the number of groups, the number of cases, and the modal category. An ordinal scale is one in which we can rank order cases such that they are greater or less than other cases on the attribute measured (e.g., class rank, pleasantness of odors). In this case, we report the median, percentiles, and the interquartile range as summary measures and can perform any transformation that preserves the rank order of the cases being measured. An interval scale (temperature) has rank order properties but, in addition, the intervals between cases are seen as equal to each other. Appropriate summary statistics include the mean and standard deviation. Associations between variables measured on interval scales can be expressed as Pearson correlations, which are the basic unit of many multivariate analysis techniques. Any linear transformation is appropriate; the mean and standard deviation may be changed, but the rank order and relative distance between cases must be preserved. Finally, ratio scales (e.g., height, weight) include a meaningful zero point and allow the expression of the equality of ratios. Only multiplicative transformations will preserve the unique properties of a ratio scale, an absolute zero point, and the capacity to form meaningful ratios between numbers on the scale.

Examples of some commonly used transformations are provided below (see Table 1) where X might be the original scores on some scale and T_1 to T_6 represent different transformations.

In this table, the first four transformations have been accomplished by adding, subtracting, multiplying, and dividing the original numbers by 2. Each of these four transformations is a linear transformation in that the mean and standard deviation of the column of numbers is changed, but the rank order and the relative size of the intervals between units on the scale remains unchanged. Further, if one computed Pearson correlations between these four transformations and the original numbers, all would correlate 1.00. The shape of the distribution of all five sets of numbers would be identical. These transformations are said to preserve the interval nature of these numbers and are routinely part of the computation of various statistics. Note that T_1 and T_2 would be inappropriate

Table 1 Common transformations of a set of raw scores (X)

X	T_1	T_2	T_3	T_4	T_5	T_6
0	2	−2	0	0	0	0.00
1	3	−1	2	0.5	1	1.00
2	4	0	4	1.0	4	1.41
3	5	1	6	1.5	9	1.73
4	6	2	8	2.0	16	2.00
5	7	3	10	2.5	25	2.24

with ratio data since the zero point is not preserved. The ratios between corresponding points on the transformed scale and the original scale do not have the same meaning.

The fifth transformation above was produced by taking the square of the original numbers and the sixth one was produced by taking the square root. These transformations might be used if the original distribution of a set of numbers was not normal or otherwise appropriate for statistical analyses such as analyses of variance. For example, reaction times to a stimulus might include a majority of short times and some very long ones. In this case, a square root transformation would make the data more appropriate for the use of parametric statistical computations. If a square, square root, or other nonlinear transformation is used, it is important to recognize that the units of measurement have been changed when we interpret or speak of the data. These transformations are said to be nonlinear; not only are the means and standard deviations of the transformed data different than X, but the size of the intervals between points on the scale are no longer uniform, the distribution of scores are altered, and the Pearson correlations between X and these two transformed scores would be less than 1.00. Other nonlinear transformations that are sometimes used include logarithmic and reciprocal transformations. One guide for the selection of an appropriate nonlinear transformation is to consider the ratio of the largest transformed score to the smallest transformed score and select the transformation that produces the smallest ratio. Using this 'rule of thumb', one is reducing the influence of extreme scores or outliers in an analysis, making it more likely that key assumptions of normality of the score distribution and homogeneity of variance are met.

There have been several popularly used transformations whose primary purpose was to increase the ability to communicate the meaning of data to the general public or to make data comparable across different sets of scores. One of the most widely used linear transformations is the z-score or standardized score. In this case, the mean of a set of scores is subtracted from the original or raw scores for each individual and this difference is divided by the standard deviation. Since the raw score is transformed by subtracting and dividing by a constant (i.e., the mean and standard deviation), this is a linear transformation (see examples above). Raw scores and z-scores are correlated 1.00 and have the same distribution but different means and standard deviations. Since the means and standard deviations of all z-scores are 0 and 1 respectively, z-scores are often used to compare individuals' scores on two or more measures. It is important to recognize that the z transformation does not result in a normal distribution of scores; if the original distribution was skewed, the transformed one will also be skewed.

The standard or z-score includes decimal numbers, and half the scores (if the distribution is normal or nearly so) are negative. For this reason, it is common practice to perform a second linear transformation on the z-scores so that they have a different mean and standard deviation. Standard scores on the Graduate Record Examination, for example, are transformed by multiplying the standard score by 100 and adding 500. This provides a score whose mean and standard deviation are 500 and 100 respectively. Many other test scores are reported as 'T' scores. These are transformed z-scores that have been multiplied by 10 and to which 50 has been added. So they have means and standard deviations of 50 and 10 respectively.

One early and well-known nonlinear transformation of scores in the behavioral sciences was the computation of an intelligence quotient (IQ) by Terman [2]. The IQ was computed by taking the person's mental age as measured by the test, dividing it by the person's chronological age, and multiplying by 100. This was a nonlinear transformation of scores since persons' chronological ages were not constant. This index helped to popularize the IQ test because it produced numbers that appeared to be readily interpretable by the general public. The fact that the IQ was unusable for measured attributes that had no relationship to chronological age and the fact that mental growth tends to asymptote around age 20 doomed the original IQ transformation. Today's IQs are usually 'deviation IQs' in which a standard score for people of particular age groups is computed and then transformed as above to create a distribution of scores with the desired mean and standard deviation.

If a normal distribution of scores is desired (i.e., we hold the assumption that a variable is normally distributed and we believe that the measurement scale we are using is faulty), we can transform a set of scores to a normal distribution. This mathematical transformation is available in many statistical texts (e.g., [3]). A relatively old normalizing transformation of scores was the stanine distribution. This transformation was done by computing the percentages of cases falling below a certain raw score and changing that raw score to its normal distribution equivalent using what is known about the normal density function and tables published at the end of most statistics texts.

Another common nonlinear transformation is the percentile scale. A percentile is the percentage of cases falling below a given raw score. Since the number of cases falling at each percentile is a constant (e.g., all percentile scores for a data set containing 1000 cases represent 10 cases), the distribution of percentiles is rectangular in shape, rather than normal or the shape of the original data. This distributional property means that it is inappropriate to use statistics that assume normally distributed data. The primary reason for computing percentiles is for public consumption since this index appears to be more easily understood and communicated to a statistically unsophisticated audience. Percentiles are the means of communicating many achievement test scores.

Transformations are useful tools both in providing a distribution of scores that is amenable to a variety of statistical analyses and in helping statisticians communicate the meaning of scores to the general public. It is, however, important to remember that we make certain assumptions about the phenomenon of interest when we transform scores and that we must return to the basic unit of measurement when we consider the practical implications of the data we observe or the manipulations of some variable(s).

References

[1] Stevens, S.S. (1946). On the theory of scales of measurement, *Science* **103**, 677–680.

[2] Terman, L.M. (1916). *The Measurement of Intelligence*, Houghton Mifflin, Boston.

[3] Winkler, R.L. & Hays, W.L. (1975). *Statistics: Probability, Inference, and Decision*, Holt, Rinehart, & Winston, New York.

NEAL SCHMITT

Tree Models

Richard Daniels and Dailun Shi

Volume 4, pp. 2059–2060

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Tree Models

Tree models, also known as *multinomial process tree models*, are data-analysis tools widely used in behavioral sciences to measure the contribution of different cognitive processes underlying observed data. They are developed exclusively for categorical data, with each observation belonging to exactly one of a finite set of categories. For categorical data, the most general statistical distribution is the multinomial distribution, in which observations are independent and identically distributed over categories, and each category has associated with it a parameter representing the probability that a random observation falls within that category. These probability parameters are generally expressed as functions of the statistical model's parameters, that is, they redefine the parameters of the multinomial distribution. Linear (e.g., **analysis of variance**) and nonlinear (e.g., **log-linear** and **logit**) models are routinely used for categorical data in a number of fields in the social, behavioral, and biological sciences. All that is required in these models is a suitable factorial experimental design, upon which a model can be selected without regard to the substantive nature of the paradigm being modeled.

In contrast, tree models are tailored explicitly to particular paradigms. In tree models, parameters that characterize the underlying process are often unobservable, and only the frequencies in which observed data fall into each category are known. A tree model is thus a special structure for redefining the multinomial category probabilities in terms of parameters that are designed to represent the underlying cognitive process that leads to the observed data. Tree models are formulated to permit statistical inference on the process parameters using observed data.

Tree models reflect a particular type of cognitive architecture that can be represented as a tree, that is, a graph having no cycles. In a tree that depicts the underlying cognitive process, each branch represents a different sequence of processing stages, resulting in a specific response category. From one stage to the next immediate stage in a processing sequence, one parameter is assigned to determine the link probability. The probability associated with a branch is the product of the link probabilities along that branch. Each branch must correspond to a category for which the number of observations is known; however, there can be more than one branch for a given category. The observed response patterns can thus be considered as the final product of a number of different cognitive processes, each of which occurs with a particular probability.

A key characteristic of tree models is that category probabilities are usually nonlinear polynomial functions of the underlying process parameters (in contrast to the classical models for categorical data mentioned above, which all have linearity built in at some level). On the other hand, tree models are much less detailed than more sophisticated cognitive models like neural networks. Thus, while tree models capture some, but not all, of the important variables in a paradigm, they are necessarily approximate and incomplete, and hence are confined to particular paradigms. Despite this disadvantage, the statistical tractability of a tree model makes it an attractive alternative to standard, multipurpose statistical models.

A comprehensive review of the theory and applications of tree models is given in Batchelder and Riefer (1999) [1]. For readers interested in learning more about tree models and statistical inference, Xiangen Hu has developed an informative website at `http://irvin.psyc.memphis.edu/gpt/`.

An Example: 'Who Said What' Task

To illustrate the structure of a tree model, consider the 'Who Said What' task. Perceivers first observe a discussion that involves members of two categories (e.g., men and women). In a subsequent recognition test, subjects are shown a set of discussion statements and asked to assign each statement to its speaker. Apart from statements that occurred in the discussion (called *old statements*), new statements are also included in the assignment phase. For each statement, participants must assign Source A (male), Source B (female), or N (new statement). Figure 1 depicts a tree model for the three types of statements. Note that there are a total of 7 process parameters $\{D_1, D_2, d_1, d_2, a, b,$ and $g\}$, 15 branches, and 9 response categories (A, B, and N for each tree).

The model assumes that a participant first detects whether a statement is old or new with probability D_1, D_2, or b for source A, B, or new statements, respectively. If an old statement is correctly detected as old, then d_1 and d_2 capture the capacity to correctly assign the old statement to source A and B, respectively. If the participant cannot directly

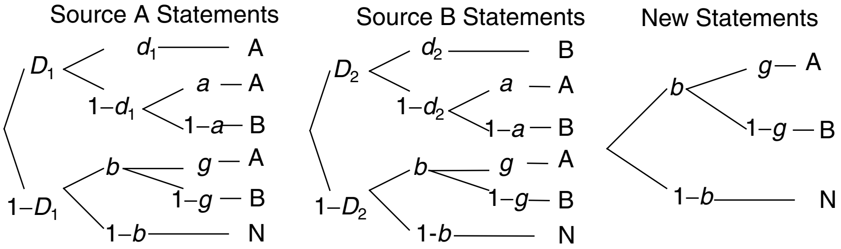


Figure 1 Tree models representing "who said what" task

attribute a statement to a source (with probability $1-d_i, i = 1, 2$), a guessing process determines the statement's source – the effectiveness of this process is measured by parameter a. If a statement is new, then another guessing process (the effectiveness of which is measured by parameter g) is used to determine the statement's source. Finally, if an old statement is not detected as old (with probability $1-D_i, i = 1, 2$), it is treated as a new statement; as such, the branches emanating from $1-D_i, i = 1, 2$, reproduce the new statement tree.

Several observations emerge from this example. First, the sequential nature of the process is based on both cognitive theory and assumptions about how statements are assigned to sources. Second, some parameters (e.g., a, g, and b) appear in more than one tree, implying, for example, that the probability of assigning a statement that is incorrectly detected as new to Source A is equal to the probability of assigning an incorrectly identified new statement to Source A. Since most of the parameters can be interpreted as conditional probabilities (i.e., conditional on the success or failure of other processes), it would perhaps be more appropriate to use different parameters to represent the same cognitive process in different trees. However, if S denotes the number of process parameters and J the number of resulting data categories, S must be no larger than $J-1$ for the model to be statistically well defined. As a result, model realism may be traded off to gain model tractability and statistical validity.

Finally, note that the category probabilities are the sums of the products of the underlying processing parameters. For example, the probability of correctly identifying a statement from Source A is$P(A|A) = D_1d_1 + D_1(1 - d_1)a + (1 - D_1)bg$. Similarly, the probability that a random observation falls into each of the other eight categories can be expressed as a function of the seven process parameters $(D_1, D_2, d_1, d_2, a, b, g)$. As such, the objective of tree modeling is to draw statistical inference on the process parameters using the sample frequencies of observations that fall into each data category, thus providing insight into the unknown cognitive processes.

Reference

[1] Batchelder, W.H. & Riefer, D.M. (1999). Theoretical and empirical review of multinomial process tree modeling, *Psychonomic Bulletin & Review* **6**(1), 57–86.

Richard Daniels and Dailun Shi

Trellis Graphics

Brian S. Everitt

Volume 4, pp. 2060–2063

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Trellis Graphics

Suppose in an investigation of crime in the USA we are interested in the relationship between crime rates in different states and the proportion of young males in the state, and whether this relationship differs between southern states and the others. To inspect the relationship graphically we might plot two graphs; the first a **scatterplot** of crime rate against proportion of young males for the southern states and the second the corresponding scatterplot for the rest. Such a diagram is shown in Figure 1 based on data given in [4].

Figure 1 is a simple example of a general scheme for examining high-dimensional structure in data by means of conditional one-, two- and three-dimensional graphs, introduced in [2]. The essential feature of such trellis displays (or casement displays as they sometimes called *see* **Scatterplot Matrices**) is the multiple conditioning that allows some type of graphic for two or more variables to be plotted for different values of a given variable (or variables). In Figure 1, for example, a simple scatterplot for two variables is shown conditional on the values of a third, in this case categorical, variable. The aim of trellis graphics is to help in understanding both the structure of the data and how well proposed models for the data actually fit.

Some Examples of Trellis Graphics

Blood Glucose Levels

Crowder and Hand [3] report an experiment in which blood glucose levels are recorded for six volunteers before and after they had eaten a test meal. Recordings were made at times −15,0,30,60,90,120,180,240, 300 and 360 min after feeding time. The whole process was repeated six times, with the meal taken at various times of the day and night. A trellis display of the relationship between glucose level and time for each subject, conditioned on the time a meal was taken is shown in Figure 2. There are clear differences in the way glucose level changes over time between the different meal times.

Married Couples

In [4] a set of data that give the heights and ages of both couples in a sample of married couples

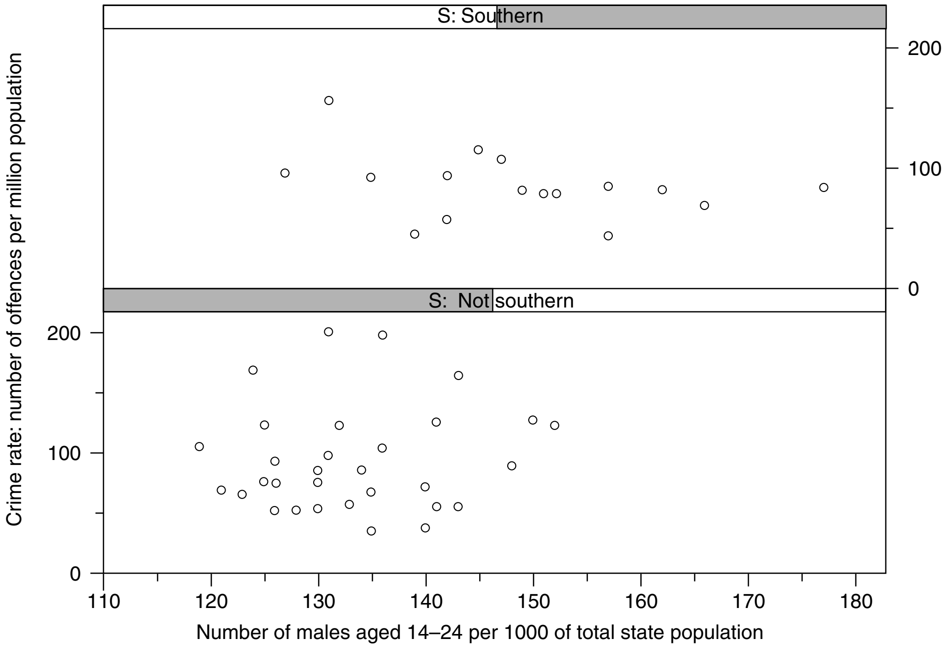


Figure 1 Scatterplots of crime rate against number of young males for southern states and other states

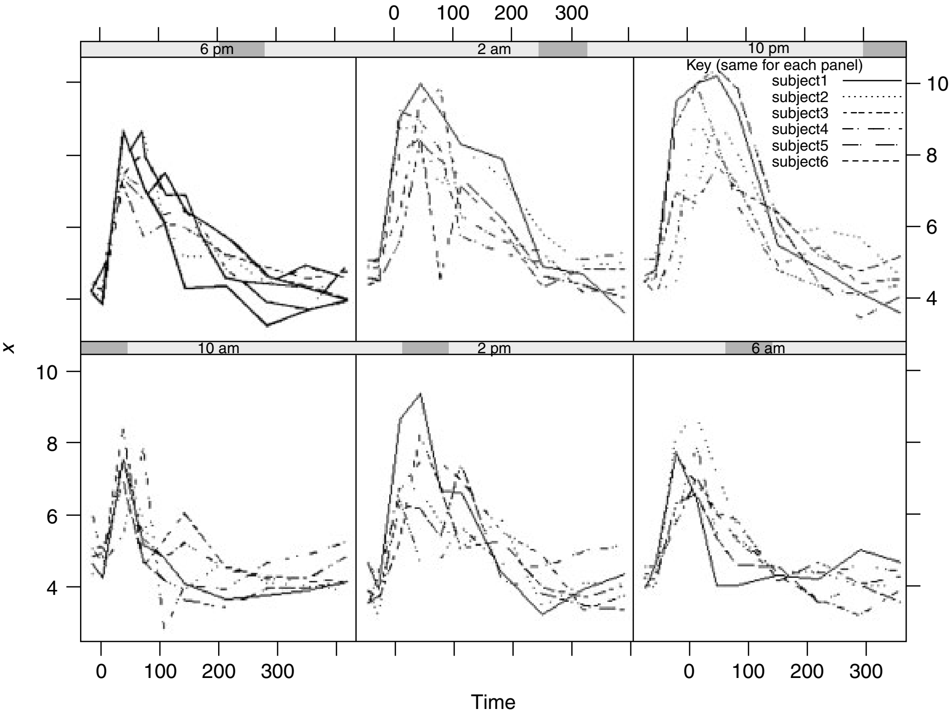


Figure 2 Trellis graphic of glucose level against time conditioned on time of eating test meal

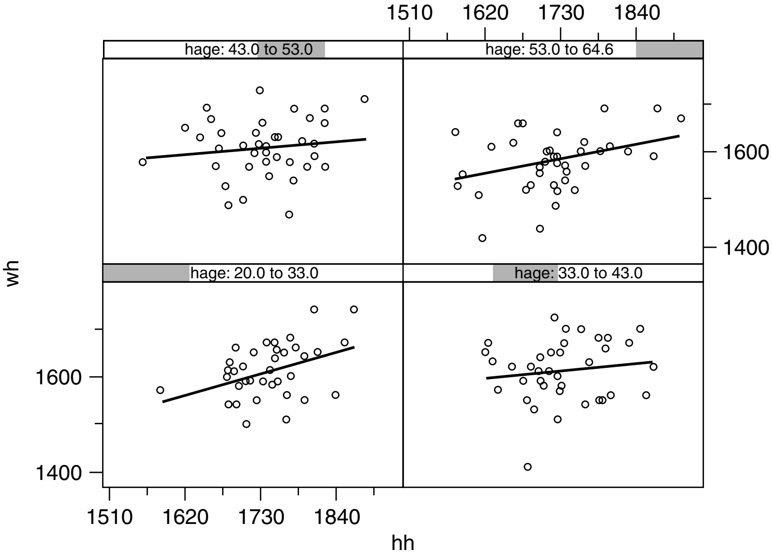


Figure 3 Trellis graphic of height of wife against height of husband conditioned on age of husband

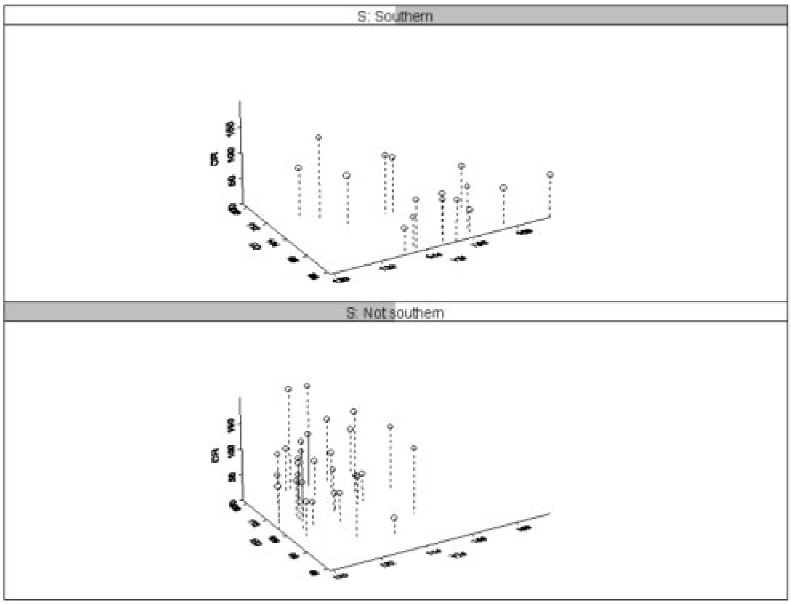


Figure 4 Three-dimensional drop-line plots for crime in southern and non-southern states

is presented. Figure 3 shows a scatterplot of the height of the wife against the height of the husband conditioned on four intervals of the age of the husband. In each of the four panels the fitted least squares regression line is shown. There is some suggestion in this diagram that amongst the youngest and the oldest husbands there is a stronger tendency for taller men to marry taller women than in the two intermediate age groups.

Crime in the USA

Finally we can return to the data set used for Figure 1 to illustrate a further trellis graphic (see Figure 4). Here a three-dimensional 'drop-line' plot is constructed for southern and non-southern states. The variables are CR-crime rate as defined in Figure 1, YM-number of young men as defined in Figure 1, and ED-educational level given by the mean number of years of schooling ×10 of the population 25 years old and over.

Some other examples of trellis graphics are given in [5].

Trellis graphics are available in **S-PLUS** as described in [1].

References

[1] Becker, R.A. & Cleveland, W.S. (1994). *S-PLUS Trellis Graphics User's Manual Version 3.3*, Insightful, Seattle.

[2] Cleveland, W. (1993). *Visualizing Data*, Hobart Press, Summit.

[3] Crowder, M.J. & Hand, D.J. (1990). *Analysis of Repeated Measures*, CRC/Chapman & Hall, London.

[4] Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. & Ostrowski, E. (1993). *Small Data Sets*, CRC/Chapman & Hall, London.

[5] Verbyla, A.P., Cullis, B.R., Kenward, M.G. & Welham, S.J. (1999). The analysis of designed experiments and longitudinal data using smoothing splines, *Applied Statistics* **48**, 269–312.

Brian S. Everitt

Trend Tests for Counts and Proportions

Li Liu, Vance W. Berger and Scott L. Hershberger

Volume 4, pp. 2063–2066

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Trend Tests for Counts and Proportions

Cochran–Armitage Trend Test

In an R by 2 or 2 by C **contingency table**, where one variable is binary and the other is ordinal, the Cochran–Armitage test for trend can be used to test the trend in the contingency table. The binary variable can represent the response, and the ordinal variable can represent an explanatory variable with ordered levels. In an R by 2 table, the trend test can test whether the proportion increases or decreases along the row variable. In a 2 by C table, the trend test can test whether the proportion increases or decreases along the column variable.

For an R by 2 table, the Cochran–Armitage trend statistic [2, 4] is defined as

$$Z^2 = \frac{\left(\sum_{i=1}^{R} n_{i1}(x_i - \bar{x})\right)^2}{p_{+1}p_{+2}\sum_{i=1}^{R} n_{i+}(x_i - \bar{x})^2}, \tag{1}$$

where n_{i1} is the count of response 1 (column 1) for the ith row, n_{i+} is the sum count of the ith row, p_{+1} is the sample proportion of response 1 (column 1), p_{+2} is the sample proportion of response 2 (column 2), x_i is the score assigned to the ith row, and $\bar{x} = \sum_{i=1}^{R} n_{i+}x_i/n$.

The statistic Z^2 has an asymptotic chi-squared distribution with 1 degree of freedom (*see* **Catalogue of Probability Density Functions**). The null hypothesis is that the proportion $p_{i1} = n_{i1}/n_{i+}$ is the same across all levels of the exploratory variable. The alternative hypothesis is that the proportion either increases monotonically or decreases monotonically along the exploratory variable. If we are interested in the direction of the trend, then the statistic Z can be used, which has an asymptotically standard normal distribution under the null hypothesis.

A simple score selection [6] is to use the corresponding row number as the score for that row. Other score selection can be used based on the specific problem.

The trend test is based on the linear probability model, where the response is the **binomial proportion**, and the exploratory variable is the score of each level of the ordinal variable. Let π_{i1} denote the probability of response 1 (column 1), and p_{i1} denote the sample proportion for $i = 1, \ldots, \mathrm{R}$. We have

$$\pi_{i1} = \alpha + \beta(\mathrm{x_i} - \bar{x}). \tag{2}$$

The weighted least squares regression (*see* **Least Squares Estimation**) gives the estimate of α and β, which are

$$\begin{aligned} \hat{\alpha} &= p_{+1}, \\ \hat{\beta} &= \frac{\sum_{i=1}^{R} n_{i+}(p_{i1} - p_{+1})(x_i - \bar{x})}{\sum_{i=1}^{R} n_{i+}(x_i - \bar{x})^2}. \end{aligned} \tag{3}$$

The Pearson statistic for testing independence for an R by 2 table is

$$\begin{aligned} \chi^2 &= \frac{\sum_{i=1}^{R} n_{i+}(p_{i1} - p_{+1})^2}{p_{+1}p_{+2}} \\ &= \frac{\sum_{i=1}^{R} n_{i+}(p_{i1} - \hat{\pi}_{i1} + \hat{\pi}_{i1} - p_{+1})^2}{p_{+1}p_{+2}} \\ &= \frac{\sum_{i=1}^{R} n_{i+}(p_{i1} - \hat{\pi}_{i1})^2 + \sum_{i=1}^{R} n_{i+}(\hat{\pi}_{i1} - p_{+1})^2}{p_{+1}p_{+2}} \\ &= \frac{\sum_{i=1}^{R} n_{i+}(p_{i1} - \hat{\pi}_{i1})^2 + \sum_{i=1}^{R} n_{i+}(p_{+1} + \hat{\beta}(x_i - \bar{x}) - p_{+1})^2}{p_{+1}p_{+2}} \\ &= \frac{\sum_{i=1}^{R} n_{i+}(p_{i1} - \hat{\pi}_{i1})^2}{p_{+1}p_{+2}} + \frac{\sum_{i=1}^{R} n_{i+}\hat{\beta}^2(x_i - \bar{x})^2}{p_{+1}p_{+2}} \end{aligned}$$

$$= \frac{\sum_{i=1}^{R} n_{i+}(p_{i1} - \hat{\pi}_{i1})^2}{p_{+1}p_{+2}} + \frac{\left(\sum_{i=1}^{R} n_{i1}(x_i - \bar{x})\right)^2}{p_{+1}p_{+2}\sum_{i=1}^{R} n_{i+}(x_i - \bar{x})^2}. \tag{4}$$

Basically, we can decompose the Pearson Statistics into two parts. The first part tests the goodness of fit of the linear model, and the second part is the Cochran–Armitage Statistic for testing a linear trend in the proportions [1].

Exact Cochran–Armitage Test for Trend

An Exact test for trend can be used if the asymptotic assumptions are not met (*see* **Exact Methods for Categorical Data**). For example, the sample size may not be large enough, or the data distribution may be sparse, skewed, and so on. Exact test for trend is based on the exact conditional method for contingency tables. Conditional on the row totals and column totals, to test independence against a trend, the exact P value is the sum of the probabilities for those tables having a test statistic larger than or equal to the observed test statistic Z^2. In practice, the sufficient statistic, $T = \sum x_i n_{i1}$, can be used as a test statistic to compute the exact P value.

Jonckheere–Terpstra Trend Test

In an R by C contingency table, where the column variable represents an ordinal response, and the row variable can be nominal or ordinal, sometimes we are interested in testing whether the ordered response follows either an increasing or decreasing trend across the rows. For example, following the omnibus **Kruskal–Wallis** test for differences among doses of a sleeping medication, we might want to determine whether the proportion of subjects who fall asleep within 30 minutes increases as the dose increases. The **Jonckheere–Terpstra trend** test is designed to test the null hypothesis that the distribution of the response variable is the same across the rows [9]. The alternative hypothesis is that

$$s_1 \le s_2 \le \cdots \le s_R \text{ or } s_1 \ge s_2 \ge \cdots \ge s_R, \tag{5}$$

with at least one of the equalities being strict, where s_i represents the ith row effect. Unlike the Cochran–Armitage test, the inequality tested by the Jonckheere–Terpstra test is not necessarily linear. The Jonckheere–Terpstra trend test was proposed independently by Jonckheere [7] and Terpstra [10], and it is a nonparametric test based on the sum of the Mann–Whitney–Wilcoxon (*see* **Wilcoxon–Mann–Whitney Test**) statistic M. To compare row i and row i', we have

$$M_{ii'} = \sum_{j=1}^{n_i} \sum_{j'=1}^{n_{i'}} I(n_{i'j'} - n_{ij}), \tag{6}$$

where $I(x)$ is equal to 0, 1/2, 1 for $x < 0$, $x = 0$, and $x > 0$ respectively. Then the Jonckheere–Terpstra trend test statistic is

$$J = \sum_{i=1}^{R-1} \sum_{i'=i+1}^{R} M_{ii'}. \tag{7}$$

Under the null hypothesis that there is no difference among the rows, the standardized test

$$J^* = \frac{J - u_J}{\sigma_J} \tag{8}$$

is asymptotically distributed as a standard normal variable, where u_J is the expected mean and σ_J is the expected standard deviation under the null. Here

$$u_J = \frac{\left(n^2 - \sum_{i=1}^{R} n_{i+}^2\right)}{4}, \tag{9}$$

and the variance is

$$\sigma_J^2 = \frac{1}{72}\left(n^2(2n+3) - \sum_{i=1}^{R} [n_{i+}^2(2n_{i+}+3)]\right). \tag{10}$$

The Jonckheere–Terpstra test is generally more powerful than the Kruskal–Wallis test, and should be used instead if there is specific interest in a trend in the data.

A modified variance adjusting for the tied values can also be used [8]. We calculate

$$\sigma_J^{*2} = \frac{J_1}{d_1} + \frac{J_2}{d_2} + \frac{J_3}{d_3}, \tag{11}$$

where

$$
\begin{aligned}
J_1 &= n(n-1)(2n+5),\\
&\quad - \sum_{i=1}^{R} n_{i+}(n_{i+}-1)(2n_{i+}+5),\\
&\quad - \sum_{j=1}^{C} n_{+j}(n_{+j}-1)(2n_{+j}+5),\\
J_2 &= \left(\sum_{i=1}^{R} n_{i+}(n_{i+}-1)(n_{i+}-2)\right),\\
&\quad \times \left(\sum_{j=1}^{C} n_{+j}(n_{+j}-1)(n_{+j}-2)\right),\\
J_3 &= \left(\sum_{i=1}^{R} n_{i+}(n_{i+}-1)\right)\left(\sum_{j=1}^{C} n_{+j}(n_{+j}-1)\right)\\
d_1 &= 72,\ \ d_2 = 36n(n-1)(n-2),\ \ d_3 = 8n(n-1).
\end{aligned}
\tag{12}
$$

When there are no ties, all $n_{i+} = 1$ and $J_2 = J_3 = 0$, resulting in $\sigma_J^{*2} = \sigma_J^2$.

The asymptotic Jonckheere–Terpstra test is also equivalent to Kendall's tau.

We can also compute the exact Jonckheere–Terpstra trend test, which is a **permutation test** [3], requiring computation of the test statistic for all permutations of a contingency table.

Example

This example illustrates the use of the Cochran–Armitage trend test. Table 1 is a data set from [5]. This is a retrospective study of the lung cancer and tobacco smoking among patients in hospitals in several English cities. One question of interest is whether subjects with higher numbers of cigarettes daily are more likely to have lung cancer. We use the equal interval scores {1, 2, 3, 4, 5, 6}. The trend test statistic $Z^2 = 129$ for df = 1, and the P value is less than 0.0001. This indicates that there is a strong linear trend along the row variable, the daily average number of cigarettes.

Alternatively, we could compute the Jonckheere–Terpstra test. For the asymptotic test, $J = 10\,90781$, $p < .0001$, and $z = -10.59$; whereas for the exact test, evidence for the inequality of the patient and control distributions is even stronger, $p < .000001$.

Table 1 Retrospective study of lung cancer and tobacco smoking. Reproduced from Doll, R. & Hill, A.B. (1952). A study of the aetiology of carcinoma of the lung, *British Medical Journal* **2**, 1271–1286 [5]

	Disease group	
Daily average Number of cigarettes	Lung cancer patients	Control patient
None	7	61
<5	55	129
5–14	489	570
15–24	475	431
25–49	293	154
50+	38	12

Both the Cochran–Armitage and Jonckheere–Terpstra tests result in the conclusion that the proportion of lung cancer increases as the daily average number of cigarettes increases.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*, John Wiley & Sons, New York.

[2] Armitage, P. (1955). Tests for linear trends in proportions and frequencies, *Biometrics* **11**, 375–386.

[3] Berger, V.W. (2000). Pros and Cons of Permutation Tests in Clinical Trials, *Statistics in Medicine* **19**, 1319–1328.

[4] Cochran, W.G. (1954). Some methods for strengthening the common χ^2 tests, *Biometrics* **10**, 417–451.

[5] Doll, R. & Hill, A.B. (1952). A study of the aetiology of carcinoma of the lung, *British Medical Journal* **2**, 1271–1286.

[6] Ivanova, A. & Berger, V.W. (2001). Drawbacks to Integer Scoring for Ordered Categorical Data, *Biometrics* **57**, 567–570.

[7] Jonckheere, A.R. (1954). A distribution-free k-sample test against ordered alternatives, *Biometrika* **41**, 133–145.

[8] Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods based on Ranks*, Holden-Day, San Francisco.

[9] SAS Institute Inc. (1999). *SAS/STAT User's Guide*, Version 8, SAS Institute Cary.

[10] Terpstra, T.J. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking, *Indigationes Mathematicae* **14**, 327–333.

Li Liu, Vance W. Berger and Scott
L. Hershberger

Trimmed Means

Rand R. Wilcox

Volume 4, pp. 2066–2067

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Trimmed Means

A trimmed mean is computed by removing a proportion of the largest and smallest observations and averaging the values that remain. Included as a special case are the usual sample mean (no trimming) and the median. As a simple illustration, consider the 11 values 6, 2, 10, 14, 9, 8, 22, 15, 13, 82, and 11. To compute a 10% trimmed mean, multiply the sample size by 0.1 and round the result down to the nearest integer. In the example, this yields $g = 1$. Then, remove the g smallest values, as well as the g largest, and average the values that remain. In the illustration, this yields 12. In contrast, the sample mean is 17.45. To compute a 20% trimmed mean, proceed as before; only, now g is 0.2 times the sample sizes rounded down to the nearest integer. Some researchers have considered a more general type of trimmed mean [2], but the description just given is the one most commonly used.

Why trim observations, and if one does trim, why not use the **median**? Consider the goal of achieving a relatively low standard error. Under normality, the optimal amount of trimming is zero. That is, use the untrimmed mean. But under very small departures from normality, the mean is no longer optimal and can perform rather poorly (e.g., [1], [3], [4], [8]). As we move toward situations in which **outliers** are common, the median will have a smaller standard error than the mean, but under normality, the median's standard error is relatively high. So, the idea behind trimmed means is to use a compromise amount of trimming with the goal of achieving a relatively small standard error under both normal and nonnormal distributions. (For an alternative approach, *see* **M Estimators of Location**). Trimming observations with the goal of obtaining a more accurate estimator might seem counterintuitive, but this result has been known for over two centuries. For a nontechnical explanation, see [6].

Another motivation for trimming arises when sampling from a skewed distribution and testing some hypothesis. **Skewness** adversely affects control over the probability of a type I error and **power** when using methods based on means (e.g., [5], [7]). As the amount of trimming increases, these problems are reduced, but if too much trimming is used, power can be low. So, in particular, using a median to deal with a skewed distribution might make it less likely to reject when in fact the null hypothesis is false.

Testing hypotheses on the basis of trimmed means is possible, but theoretically sound methods are not immediately obvious. These issues are easily addressed, however, and easy-to-use software is available as well, some of which is free [7, 8].

References

[1] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., & Stahel, W.A. (1986). *Robust Statistics*, Wiley, New York.

[2] Hogg, R.V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory, *Journal of the American Statistical Association* **69**, 909–922.

[3] Huber, P.J. (1981). *Robust Statistics*, Wiley, New York.

[4] Staudte, R.G. & Sheather, S.J. (1990). *Robust Estimation and Testing*, Wiley, New York.

[5] Westfall, P.H. & Young, S.S. (1993). *Resampling Based Multiple Testing*, Wiley, New York.

[6] Wilcox, R.R. (2001). *Fundamentals of Modern Statistical Methods: Substantially Increasing Power and Accuracy*, Springer, New York.

[7] Wilcox, R.R. (2003). *Applying Conventional Statistical Techniques*, Academic Press, San Diego.

[8] Wilcox, R.R. (2004). (in press). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd Edition, Academic Press, San Diego.

Further Reading

Tukey, J.W. (1960). A survey of sampling from contaminated normal distributions, in *Contributions to Probability and Statistics*, I. Olkin, S. Ghurye, W. Hoeffding, W. Madow & H. Mann, eds, Stanford University Press, Stanford.

RAND R. WILCOX

T-Scores

David Clark-Carter

Volume 4, pp. 2067–2067

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

T-Scores

T-scores are standardized scores. However, unlike **z-scores**, which are standardized to have a mean of 0 and an SD of 1, the T-scoring system (due to McCall [1]) produces a distribution of new scores with a mean of 50 and an SD of 10. The easiest way of calculating a T-score is to find the z-score first and then apply the following simple linear transformation:

$$T = z \times 10 + 50. \tag{1}$$

T-scoring gives the same benefits as any other standardizing system in that, for instance, it makes possible direct comparisons of a person's scores over different, similarly scored, tests. However, it has the additional advantage over z-scores of producing new scores that are easier to interpret since they are always positive and expressed as whole numbers.

If the original scores come from a normal population with a known mean and SD, the resulting T-scores will be normally distributed with mean 50 and SD of 10. Thus, we can convert these back to z-scores and then use standard normal tables to find the percentile point for a given T-score.

T-scoring is the scoring system for several test instruments commonly used in psychology, such as the MMPI.

Reference

[1] McCall, W.A. (1939). *Measurement*, Macmillan Publishers, New York.

David Clark-Carter

Tukey, John Wilder

Sandy Lovie

Volume 4, pp. 2067–2069

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Tukey, John Wilder

Born: June 16, 1915, in Massachusetts, USA.
Died: July 26, 2000, in New Jersey, USA.

It is almost impossible to be a true polymath in modern science, if only because of the intense pressure to specialize as early as possible. John Tukey was one of the few exceptions to this rule in that he made fundamental contributions to many distinct areas of science including statistics, computing, and mathematics, where the latter meant his early love, topology. What seems to characterize all his work is an unwillingness to accept the accepted; thus, his most extensively articulated area of work, **Exploratory Data Analysis** (or EDA), sidesteps twentieth century statistical obsessions with inference to concentrate on extracting new meanings from data. In other words, EDA is perhaps best seen as an attempt to recreate an *inductive* approach to statistics, which may or may not have actually existed in the past. Certainly, Tukey's massive search for novel measures of location in place of the mean using large scale computer simulations seems to reflect his wish to both query the foundations of statistics and to make foundational changes in the discipline [1]. Such a work also mirrors the long running historical debates in the eighteenth and nineteenth centuries over the definition and role of the mean in statistics, where these were most decidedly not the unproblematic issues that they appear today (see [4], [5]). Tukey's invention of novel graphical methods can also be viewed in something like the same light, that is, as both an undermining throwback to the nineteenth century when plots and other displays were central tools for the statistician, and as solutions to modern problems making use of modern technology such as computers, visual displays, and plotters.

Tukey's early life was somewhat isolated and protected in that he was an only child with most of his education coming from his mother, who acted as a private tutor, since, being married, she was unable to practice her original training as a teacher. Tukey's father was also a school teacher, whose *metier* was Latin. Tukey's early degrees were from Brown University (bachelors and masters in chemistry, 1936 and 1937). He then moved to Princeton with the initial aim of obtaining a PhD in chemistry, but saw the error of his ways and switched to mathematics (see the transcript of his early Princeton reminiscence's in [6]). After finishing his graduate work in topology and obtaining his doctorate in 1939, he was appointed to an assistant professorship in the Mathematics Department in Princeton and full professorship in 1950 at the age of 35. Meanwhile, he had discovered statistics when working during the war for Fire Control Research based in Princeton from 1941 to 1945. Other important workers attached to this unit were Tukey's statistical mentor Charles Winsor, Albert Tucker, and **Claude Shannon**. After the war, he alternated between Princeton and the Bell Laboratories at Murray Hill, New Jersey. The earliest work of note is his influential text with Blackman on the analysis of power spectra, which appeared in 1959. But while still maintaining an interest in this field, including the use of computers in the approximation of complex functions (his well regarded paper on fast Fourier transforms, for example, had been published jointly with Cooley in 1964), Tukey had, nevertheless, begun to move into EDA via the study of robustness and regression residuals (*see* **Robust Testing Procedures**; **Robustness of Standard Tests**; **Regression Models**). The public side of this finally emerged in 1977 in the two volumes on EDA [7], and EDA and regression [3], the latter written jointly with Frederick Mosteller.

The initial reaction to these books was strongly positive on many statisticians part, but there were equally strong negative reactions as well. The British statistician Ehrenberg, for example, is quoted as saying that if he had not known who the author was he would have dismissed EDA as a joke. Happily EDA survived and flourished, particularly in the resuscitation and redirecting of regression analysis from a rather fusty nineteenth century area of application of least squares methods into a dynamic and multifaceted technique for the robust exploration of complex data sets. Further, a great many of the displays pioneered by Tukey are now staple fodder in just about all the statistics packages that one can point to, from the hand-holding ones like SPSS and Minitab, to the write-your-own-algorithm environments of S, S-Plus and R. Tukey's defense of this apparent fuzziness is well known, arguing as he did on many occasions that an approximate answer to the correct question is always preferable to a precise one to an incorrect one. Tukey in his long and fruitful career has also inspired generations of students and coworkers, as may be seen from the published volume of collaborative work. In addition, his collected works now run to nine volumes, although I suspect that this is not the final count; while the latest book of his that I can find is one written jointly with Kaye Basford on the graphical analysis of some classic plant breeding trials [2], which he published a year before his death at the age of 84!

References

[1] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. & Tukey, J.W. (1972). *Robust Estimates of Location: Survey and Advances*, Princeton University Press, Princeton.

[2] Basford, K.E. & Tukey, J.W. (1999). *Graphical Analysis of Multiresponse Data: Illustrated with a Plant Breeding Trial*, Chapman & Hall, London.

[3] Mosteller, F. & Tukey, J.W. (1977). *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley, Reading.

[4] Porter, T.M. (1986). *The Rise of Statistical Thinking 1820–1900*, Princeton University Press, Princeton.

[5] Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*, Harvard University Press, Harvard.

[6] Tucker, A. (1985). *The Princeton Mathematics Community in the 1930s: John Tukey*, Princeton University, Transcript No. 41 (PMC41).

[7] Tukey, J.W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading.

SANDY LOVIE

Tukey Quick Test

Shlomo Sawilowsky

Volume 4, pp. 2069–2070

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Tukey Quick Test

The Tukey procedure [4] is a two-independent-samples test of differences in location. It is an alternative to the parametric *t* Test. Despite its simplicity, Neave and Worthington [3] noted that it is 'an entirely valid and distribution-free test'. It has the benefit of easily memorized critical values.

Procedure

The first step is to identify the maximum score in the sample (A_{Max}) with the smaller median and also the minimum score in sample (B_{Min}) with the larger median. The second step is to count the number of scores in sample B that are lower than A_{Max} and also the number of scores in sample A that are higher than B_{Min}.

Hypotheses

The null hypothesis, H_o, is that there is no difference in the two population medians or that the two samples were drawn from the same population. The alternative hypothesis, H_1, is that the populations sampled have different medians or the samples originate from different populations.

Assumptions

Tukey's test assumes there are no tied values, especially at the extreme ends. Neave and Worthington [3] suggested an iterative method of breaking ties in all possible directions, computing the test statistic T for each iteration, and making the decision based on the average value of T. Monte Carlo results by Fay and Sawilowsky [2] and Fay [1] indicated that a simpler procedure for resolving tied values is to randomly assign tied values for or against the null hypothesis.

Test Statistic

The test statistic, T, is the sum of the two counts described above. Critical values for the two-tailed test are easily remembered. As long as the ratio of the two sample sizes is less than 1.5, the critical values for $\alpha = 0.05$ and 0.01 are 7 and 10, respectively. Critical values for the one-tailed test are 6 and 9. Additional tabled critical values appear in [3].

Example

Consider the data from two samples in the table below (Table 1).

Table 1 Sample data

Sample A	Sample B
201	334
333	418
335	419
340	442
420	469
	417

$A_{\text{Max}} = 420$ from sample A, and $B_{\text{Min}} = 334$ from sample B. Four scores in sample B are lower than A_{Max} (334, 418, 419, and 417). Three scores in Group A are higher than B_{Min} (335, 340, and 420). The test statistic is $T = 3 + 4 = 7$, which is significant at $\alpha = 0.050$. Thus, the null hypothesis of no difference in location is rejected in favor of the alternative that Sample B has the larger median.

References

[1] Fay, B.R. (2003). A Monte Carlo computer study of the power properties of six distribution-free and/or nonparametric statistical tests under various methods of resolving tied ranks when applied to normal and nonnormal data distributions, Unpublished doctoral dissertation, Wayne State University, Detroit.

[2] Fay, B.R. & Sawilowsky, S. (2004). The effect on type I error and power of various methods of resolving ties for six distribution-free tests of location. Manuscript under review.

[3] Neave, H.R. & Worthington, P.L. (1988). *Distribution-Free Tests*, Unwin Hyman, London.

[4] Tukey, J.W. (1959). A quick, compact, two-sample test to Duckworth's specifications, *Technometrics* **1**, 31–48.

(*See also* **Distribution-free Inference, an Overview**)

SHLOMO SAWILOWSKY

Tversky, Amos

Sandy Lovie

Volume 4, pp. 2070–2071

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Tversky, Amos

Born: March 16, 1937, in Haifa.
Died: June 2, 1996, in California.

There are two, somewhat different, Amos Tversky's: one is the high profile co-inventor (with Daniel Kahneman) of cognitive biases and heuristics, the other the considerably more retiring mathematical psychologist heavily involved with the development of polynomial conjoint measurement, the recent Great White Hope of psychological measurement [3]. Although Tversky will almost certainly be remembered for his joint attack on the statistical complacency of a generation of psychologists, and the ramifications of both the Law of Small Numbers and a raft of related phenomena, the more rigorous side of his contribution to mathematical psychology should not be overlooked.

Tversky's early education was at the Hebrew University of Jerusalem after a year of army service as a paratrooper, when he had been awarded a military decoration for rescuing a fellow soldier who had got into trouble laying an explosive charge. He also served his country in 1967 and 1973. After graduating from the University in 1961 with a BA in philosophy and psychology, Tversky had immediately moved to the University of Michigan in America to take a PhD with Ward Edwards, then the leading behavioral decision theorist. He was also taught mathematical psychology at Michigan by **Clyde Coombs** whose deep interest in scaling theory had rubbed off on Tversky. On obtaining his PhD in 1965, he had then travelled to the east coast for a year's work at the Harvard Centre for Cognitive Studies. He returned to the Hebrew University in 1966 where he was made full professor in 1972. Interspersed with his time in Israel was a stretch from 1970 as a fellow at Stanford University's Centre for the Advanced Study of the Behavioral Sciences. In 1978, he finally joined the Psychology Department at Stanford where he remained until his death.

In 1969, Tversky had been invited by Daniel Kahneman, his near contemporary at the Hebrew University, to give a series of lectures to the University on the human assessment of probability. From this early collaboration came experimental demonstration of the Law of Small Numbers, and the rest, as they say, is history. Kahneman and Tversky presented a set of deceptively simple stories recounting statistical problems to a series of mathematically and statistically sophisticated audiences. The results showed the poor quality of the statistical intuitions of these groups, leading Kahneman and Tversky to invent an ever expanding set of biases and cognitive short cuts (heuristics), including the best known ones of *representativeness* (the small mirrors the large in all essential elements), and *availability* (the probability that people assign to events is a direct function of how easily they are generated or can be retrieved from memory). Others include the conjunction fallacy, regression to the mean, anchoring and adjustment, and the simulation heuristic, where the latter has triggered a veritable explosion of research into what is termed 'counterfactual' reasoning or thinking, that is, 'but what if...' reasoning (*see* **Counterfactual Reasoning**). The movement also attracted additional studies outside the immediate Tversky–Kahneman, axis, for example, the work of Fischhoff on the hindsight bias ('I always knew it would happen'), while the notion of biases and heuristics seemed to offer a framework for other related studies, such as those on illusory correlation, vividness, and human probability calibration. Indeed, Tversky's later economics-orientated Prospect theory threatened to account for most behavioral studies of risk and gambling! This approach also generated the notion of decision frames, that is, the idea that all choices are made within a context, or, put another way, that people's risky decision making can only be understood if you also understand the setting in which it takes place (classic accounts of the material can be found in [2], with extensive updates contained in [1]). Not too surprisingly, the theme of heuristics and biases has also been taken up enthusiastically by cognitive social psychologists (see [4] for the initial reaction) (*see* **Decision Making Strategies**).

Kahneman was (jointly) awarded the Nobel Prize for Economics in 2003 for his work on what is now termed *Behavioral Economics*. Unfortunately, since the Prize cannot be awarded posthumously, Tversky missed being honoured for his seminal contribution to this new and exciting area.

References

[1] Gilovich, T., Griffin, D. & Kahneman, D., eds (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge University Press, Cambridge.

[2] Kahneman, D., Slovic, P. & Tversky, A., eds (1982). *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge.

[3] Krantz, D.H., Luce, R.D., Suppes, P. & Tversky, A., eds (1971). *Foundations of Measurement*, Vol. 1, Academic Press, New York.

[4] Nisbett, R.E. & Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgement*, Prentice-Hall, Englewood Cliffs.

SANDY LOVIE

Twin Designs

Frank M. Spinath

Volume 4, pp. 2071–2074

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Twin Designs

Introduction

The classical twin study compares the phenotypic resemblances of identical or *monozygotic* (MZ) and fraternal or *dizygotic* (DZ) twins. MZ twins derive from the splitting of one fertilized zygote and therefore inherit identical genetic material. DZ twins are first-degree relatives because they develop from separately fertilized eggs and are 50% genetically identical on average. It follows that a greater within-pair similarity in MZ compared to DZ twins suggests that genetic variance influences the trait under study.

The discovery of the twin method is usually ascribed to **Galton** [9] although it is uncertain whether Galton was aware of the distinction between MZ and DZ twins [22]. It was not until almost 50 years later that explicit descriptions of the classical twin method were published [14, 25].

Terminology

To disentangle and to quantify the contributions that genes and the environment make to human complex traits, data are required either from relatives who are genetically related but who grow up in unrelated environments ('*twin adoption design*' (*see* **Adoption Studies**)), or from relatives who grow up in similar environments but are of differing genetic relatedness ('*twin design*') [1]. Most twin studies that have been conducted over the past 80 years are of the latter type. Only two major studies of the former type have been conducted, one in Minnesota [2] and one in Sweden [17]. These studies have found, for example, that monozygotic twins reared apart from early in life are almost as similar in terms of general cognitive ability as are monozygotic twins reared together, a result suggesting strong genetic influence and little environmental influence caused by growing up together in the same family. These influences are typically called (*see* **Shared Environment**) because they refer to environmental factors contributing to the resemblance between individuals who grow up together [20]. **Nonshared environmental** influences, on the other hand, refer to environmental factors that make individuals who grow up together different from one another.

Twinning

One reason why a predominant number of twin studies have utilized the twin design instead of the twin adoption design is that twins typically grow up together, thus it is much easier to find a large number of participants for the classic twin study. In humans, about 1 in 85 live births are twins. The numbers of identical and same-sex fraternal twins are approximately equal. That is, of all twin pairs, about one third are identical twins, one third are same-sex fraternal twins, and one third are opposite-sex fraternal twins. The rate of twinning differs across countries, increases with maternal age, and may even be inherited in some families. Greater numbers of fraternal twins are the result of the increased use of fertility drugs and *in vitro* fertilization, whereas the rate of identical twinning is not affected by these factors [20].

Zygosity Determination

The best way to determine twin zygosity is by means of DNA markers (polymorphisms in DNA itself). If a pair of twins differs for any DNA marker, they must be fraternal because identical twins are identical genetically. If a reasonable number of markers are examined and no differences are found, it can be concluded that the twin pair is identical. Physical similarity on highly heritable traits such as eye color, hair color, or hair texture as well as reports about twin confusion are also often used for zygosity determination. If twins are highly similar for a number of physical traits, they are likely to be identical. Using physical similarity to determine twin zygosity typically yields accuracy of more than 90% when compared to genotyping data from DNA markers (e.g., [5]).

Deriving Heritability and Environmental Estimates from Twin Correlations

Comparing the phenotypic (*see* **Genotype**) resemblance of MZ and DZ twins for a trait or measure under study offers a first estimate of the extent to which genetic variance is associated with phenotypic variation of that trait. If MZ twins resemble each other to a greater extent than do DZ twins, the **heritability** (***h^2***) of the trait can be estimated by doubling the difference between MZ and DZ correlations, that is, $h^2 = 2(r_{MZ} - r_{DZ})$ [7]. *Heritability* is defined as the proportion of phenotypic differences among individuals that can be attributed to genetic differences in a particular population. Whereas *broad-sense* heritability involves all additive and nonadditive sources of genetic variance, *narrow-sense* heritability is limited to additive genetic variance. The proportion of the variance that is due to the shared environment (c^2) can be derived from calculating $c^2 = r_{MZ} - h^2$ where r_{MZ} is the correlation between MZ twins because MZ similarity can be conceptualized as h^2 (similarity due to genetic influences) $+c^2$ (similarity due to shared environmental influences). Substituting $2(r_{MZ} - r_{DZ})$ for h^2 offers another way of calculating shared environment ($c^2 = 2r_{DZ} - r_{MZ}$). In other words, the presence of shared environmental influences on a certain trait is suggested if DZ twin similarity exceeds half the MZ twin similarity for that trait. Similarly, *nonadditive* genetic effects (dominance and/or epistasis) are implied if DZ twin similarity is less than half the MZ twin correlation. Finally, nonshared environmental influences (e^2) can be estimated from $e^2 = r_{tt} - r_{MZ}$ where r_{tt} is the test-retest reliability of the measure. If $1 - r_{MZ}$ is used to estimate e^2 instead, the resulting nonshared environmental influences are confounded with measurement error. For example, in studies of more than 10,000 MZ and DZ twin pairs on general cognitive ability (g), the average MZ correlation is 0.86, which is near the test-retest reliability of the measures, in contrast to the DZ correlation of 0.60 [21]. Based on this data, application of the above formulae results in a heritability estimate of $h^2 = 0.52$, shared environmental estimate of $c^2 = 0.34$ and nonshared environmental influence/measurement error estimate of $e^2 = 0.14$.

Thus, given that MZ and DZ twin correlations are available, this straightforward set of formulae can be used to derive estimates of genetic and environmental influences on any given trait under study. Instead of the **Pearson product-moment correlation** coefficient, twin similarity is typically calculated using **intra-class correlation** (ICC1.1; [24]) which is identical to the former only if there are no mean or variance differences between the twins.

Requirements and Assumptions

It should be noted that for a meaningful interpretation of twin correlations in the described manner, a number of assumptions have to be met: The absence of assortative mating for the trait in question, the absence of **G**(enotype) – **E**(nvironment) **correlation** and interaction, and the viability of the Equal Environments Assumption. Each of these assumptions will be addressed briefly below.

Assortative mating describes nonrandom mating that results in similarity between spouses and increases correlations and the genetic similarity for first-degree relatives if the trait under study shows genetic influence. Assortative mating can be inferred from spouse correlations which are comparably low for some psychological traits (e.g., personality), yet are substantial for others (e.g., intelligence), with average spouse correlations of about .40 [10]. In twin studies, assortative mating could result in underestimates of heritability because it raises the DZ correlation but does not affect the MZ correlation. If assortative mating were not taken into account, its effects would be attributed to the shared environment.

Gene-Environment Correlation describes the phenomenon that genetic propensities can be correlated with individual differences in experiences. Three types of G × E correlations are distinguished: passive, evocative, and active [19]. Previous research indicates that genetic factors often contribute substantially to measures of the environment, especially the family environment [18]. In the classic twin study, however, G × E correlation is assumed to be zero because it is essentially an analysis of main effects.

G × E interaction (*see* **Gene-Environment Interaction**) is often conceptualized as the genetic control of sensitivity to the environment [11]. Heritability that is conditional on environmental exposure can indicate the presence of a G × E interaction. The classic twin study does not address G × E interaction.

The classic twin model assumes the equality of pre- and postnatal environmental influences within the two types of twins. In other words, the *Equal Environments Assumption* (*EEA*) assumes that environmentally caused similarity is roughly the same for both types of twins reared in the same family. Violations of the EEA because MZ twins experience more similar environments than DZ twins would inflate estimates of genetic influences. The EEA has been tested in a number of studies and even though MZ twins appear to experience more similar environments than DZ twins, it is typically concluded that these differences do not seem to be responsible for the greater MZ twin compared to DZ twin similarity [3]. For example, empirical studies have shown that within a group of MZ twins, those pairs who were treated more individually than others do not behave more differently [13, 15]. Another way of putting the EEA to a test is studying twins who were mislabeled by their parents, that is, twins whose parents thought that they were dizygotic when they were in fact monozygotic and vice versa. Results typically show that the similarity of mislabeled twin pairs reflects their biological zygosity to a much greater extent than their assumed zygosity [12, 23].

In recent years, the *prenatal environment* among twins has received increasing attention (e.g., [4]). About two thirds of MZ twin pairs are monochorionic (MC). All DZ twins are dichorionic (DC). The type of placentation in MZ twins is a consequence of timing in zygotic diversion. If the division occurs at an early stage (up to day 3 after fertilization), the twins will develop separate fetal membranes (chorion and amnion), that is, they will be dichorionic diamniotic. When the division occurs later (between days 4 and 7), the twins will be monochorionic diamniotic. For anthropological measures (such as height), a 'chorion effect' has been documented: Within-pair differences are larger in MC-MZs than in DC-MZs. Findings for cognitive and personality measures are less consistent. If possible, chorionicity should be taken into account in twin studies even if the above example shows that failing to discriminate between MC and DC MZ twin pairs does not necessarily lead to an overestimate of heritability. The importance of the prenatal maternal environment on IQ has been demonstrated in a recent meta-analysis [6].

Structural Equation Modeling

The comparison of intra-class correlations between MZ versus DZ twins can be regarded as a reasonable first step in our understanding of the etiology of particular traits. This approach, however, cannot accommodate the effect of gender on variances and covariances of opposite-sex DZ twins. To model genetic and environmental effects as the contribution of unmeasured (latent) variables to phenotypic differences, **Structural Equation Modelling (SEM)** is required. Analyzing unvariate data from MZ and DZ twins by means of SEM offers numerous advances over the mere use of correlations, including an overall statistical fit of the model, tests of parsimonious submodels, and maximum likelihood confidence intervals for each latent influence included in the model. The true strength of SEM, however, lies in its application to multivariate and multigroup data. During the last decade powerful models and programs to efficiently run these models have been developed [16]. Extended twin designs and the simultaneous analysis of correlated traits are among the most important developments that go beyond the classic twin designs, yet still use the information inherently available in twins [1].

Outlook

Results from classical twin studies have made a remarkable contribution to one of the most dramatic developments in psychology during the past few decades: The increased recognition of the important contribution of genetic factors to virtually every psychological trait [20], particularly for phenotypes such as autism [8]. Currently, worldwide registers of extensive twin data are being established and combined with data from additional family members offering completely new perspectives in a refined behavioral genetic research [1]. Large-scale longitudinal twin studies (*see* **Longitudinal Designs in Genetic Research**) such as the Twins Early Development Study (TEDS) [26] offer opportunities to study the etiology of traits across time and at the extremes and compare it to the etiology across the continuum of trait expression. In this way, twin data remains a valuable and vital tool in the toolbox of behavior genetics.

References

[1] Boomsma, D., Busjahn, A. & Peltonen, L. (2002). Classical twin studies and beyond, *Nature Reviews Genetics* **3**, 872–882.

[2] Bouchard Jr, T.J., Lykken, D.T., McGue, M., Segal, N.L. & Tellegen, A. (1990). Sources of human psychological differences: the Minnesota study of twins reared apart, *Science* **250**, 223–228.

[3] Bouchard Jr, T.J. & Propping, P. (1993). *Twins as a Tool of Behavioral Genetics*, John Wiley & Sons, New York.

[4] Carlier, M. & Spitz, E. (2000). The twin method, in *Neurobehavioral Genetics: Methods and Applications*, B. Jones & P. Mormède eds, CRC Press, pp. 151–159.

[5] Chen, W.J., Chang, H.W., Lin, C.C.H., Chang, C., Chiu, Y.N. & Soong, W.T. (1999). Diagnosis of zygosity

by questionnaire and polymarker polymerase chain reaction in young twins, *Behavior Genetics* **29**, 115–123.

[6] Devlin, B., Daniels, M. & Roeder, K. (1997). The heritability of IQ, *Nature* **388**, 468–471.

[7] Falconer, D.S. (1960). *Introduction to Quantitative Genetics*, Ronald Press, New York.

[8] Folstein, S. & Rutter, M. (1977). Genetic influences and infantile autism, *Nature* **265**, 726–728.

[9] Galton, F. (1876). The history of twins as a criterion of the relative powers of nature and nurture, *Royal Anthropological Institute of Great Britain and Ireland Journal* **6**, 391–406.

[10] Jensen, A.R. (1998). *The g Factor: The Science of Mental Ability*, Praeger, London.

[11] Kendler, K.S. & Eaves, L.J. (1986). Models for the joint effects of genotype and environment on liability to psychiatric illness, *American Journal of Psychiatry* **143**, 279–289.

[12] Kendler, K.S., Neale, M.C., Kessler, R.C., Heath, A.C. & Eaves, L.J. (1993). A test of the equal-environment assumption in twin studies of psychiatric illness, *Behavior Genetics* **23**, 21–27.

[13] Loehlin, J.C. & Nichols, J. (1976). *Heredity, Environment and Personality*, University of Texas, Austin.

[14] Merriman, C. (1924). The intellectual resemblance of twins, *Psychological Monographs* **33**, 1–58.

[15] Morris Yates, A., Andrews, G., Howie, P. & Henderson, S. (1990). Twins: a test of the equal environments assumption, *Acta Psychiatrica Scandinavica* **81**, 322–326.

[16] Neale, M.C., Boker, S.M., Xie, G. & Maes, H. (1999). *Mx: Statistical Modeling*, 5th Edition, VCU Box 900126, Department of Psychiatry, Richmond, 23298.

[17] Pedersen, N.L., McClearn, G.E., Plomin, R. & Nesselroade, J.R. (1992). Effects of early rearing environment on twin similarity in the last half of the life span, *British Journal of Developmental Psychology* **10**, 255–267.

[18] Plomin, R. (1994). *Genetics and Experience: The Interplay Between Nature and Nurture*, Sage Publications, Thousand Oaks.

[19] Plomin, R., DeFries, J.C. & Loehlin, J.C. (1977). Genotype-environment interaction and correlation in the analysis of human behavior, *Psychological Medicine* **84**, 309–322.

[20] Plomin, R., DeFries, J.C., McClearn, G.E. & McGuffin, P. (2001). *Behavioral Genetics*, 4th Edition, Worth Publishers, New York.

[21] Plomin, R. & Spinath, F.M. (2004). Intelligence: genetics, genes, and genomics, *Journal of Personality and Social Psychology* **86**, 112–129.

[22] Rende, R.D., Plomin, R. & Vandenberg, S.G. (1990). Who discovered the twin method? *Behavior Genetics* **20**, 277–285.

[23] Scarr, S. & Carter-Saltzman, L. (1979). Twin method: defense of a critical assumption, *Behavior Genetics* **9**, 527–542.

[24] Shrout, P.E. & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability, *Psychological Bulletin* **86**, 420–428.

[25] Siemens, H.W. (1924). *Die Zwillingspathologie: Ihre Bedeutung, ihre Methodik, Ihre Bisherigen Ergebnisse. (Twin Pathology: Its Importance, its Methodology, its Previous Results)*, Springer, Berlin.

[26] Trouton, A., Spinath, F.M. & Plomin, R. (2002). Twins early development study (TEDS): a multivariate, longitudinal genetic investigation of language, cognition and behaviour problems in childhood, *Twin Research* **5**, 444–448.

FRANK M. SPINATH

Twins Reared Apart Design

NANCY L. SEGAL

Volume 4, pp. 2074–2076

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Twins Reared Apart Design

Studies of identical twins raised apart from birth provide direct estimates of genetic influence on behavior. This is because when twins are brought up separately, in uncorrelated environments and with minimal or no contact until adulthood, their trait similarity is associated with their shared genes. Fraternal, or nonidentical, twins reared apart offer investigators an important control group, as well as opportunities for tests of various interactions between genotypes and environments. Separated twins are rare, relative to twins reared together. However, there have been seven studies of reared apart twins, conducted in six countries. They include the United States, Great Britain, Denmark, Japan, Sweden, and Finland [9]. Identifying reared apart twins has been easier in Scandinavian nations where extensive population registries are maintained.

Heritability, the portion of population variance associated with genetic differences in measured traits can be calculated more efficiently using twins reared apart than twins reared together. For example, 400 to 500 identical and 400 to 500 fraternal twin pairs reared together allow heritability to be estimated with the same degree of confidence as 50 MZ twin pairs reared apart [5]. This is because heritability is estimated directly from reared apart twins and indirectly from reared together twins.

Findings from reared apart twin studies have been controversial. Four key objections have been raised: (a) twins reared by relatives will be more similar than twins reared by unrelated families; (b) twins separated late will be more alike than twins separated early; (c) twins meeting before they are tested will be more similar than twins meeting later because of their increased social contact; and (d) similarities in twins' separate homes will be associated with their measured similarities [12]. These objections have, however, been subjected to testing and can be ruled out [1, 2].

A large number of analyses can be conducted by using the family members of reared apart twins in ongoing studies. It is possible to compare twin-spouse similarity, spouse–spouse similarity and similarity between the two sets of offspring who are 'genetic half-siblings' [8]. Another informative addition includes the unrelated siblings with whom the twins were raised; these comparisons offer tests of the extent to which shared environments are associated with behavioral resemblance among relatives [11].

A large number of reared apart twin studies have demonstrated genetic influence on psychological, physical, and medical characteristics [1]. One of the most provocative findings concerns personality development. A personality study combining four twin groups (identical twins raised together and apart; fraternal twins raised together and apart) found that the degree of resemblance was the same for identical twins raised together and identical twins raised apart [13]. The shared environment of the twins reared together did not make them more similar than their reared apart counterparts; thus, personality similarity in family members is explained by their similar genes, not by their similar environments. However, twins reared together are somewhat more alike in general intelligence than twins reared apart; the **intraclass correlations** are 0.86 and 0.75, respectively [6]. These values may be misleading because most twins reared together are measured as children (when the modest effects of the shared family environment on ability are operative), while twins reared apart are measured as adults. It is possible that adult twins reared together would be as similar in intelligence as adult twins reared apart.

Studies of identical reared apart twins are natural **co-twin control studies**, thus providing insights into environmental effects on behavior. For example, it is possible to see if differences in twins' rearing histories are linked to differences in their current behavior. An analysis of relationships between IQ and rearing family measures (e.g., parental socioeconomic status, facilities in the home) did not find meaningful associations [2]. Case studies of selected pairs are also illustrative in this regard. Twins in one set of identical British women were raised by parents who provided them with different educational opportunities. Despite these differences, the twins' ability levels were quite close and both twins were avid readers of the same type of books.

Twins reared apart can also be used to examine evolutionary-based questions and hypotheses [10]. A finding that girls raised in father-absent homes undergo early sexual development and poor social relationships has attracted attention [4]. It has been suggested that studies of MZ female cotwins reared separately in father-absent and father-present homes could clarify factors underlying this behavioral pattern [3]. Further discussion of the potential role of reared apart twins in evolutionary psychological work is available in Mealey [7].

References

[1] Bouchard Jr, T.J. (1997). IQ similarity in twins reared apart: findings and responses to critics, in *Intelligence: Heredity and Environment*, R.J. Sterberg & E.L. Grigorenko eds. Cambridge University Press, New York. pp. 126–160.

[2] Bouchard Jr, T.J., Lykken, D.T., McGue, M., Segal, N.L. & Tellegen, A. (1990). Sources of human psychological differences: the minnesota study of twins reared apart. *Science* **250**, 223–228.

[3] Crawford, C.B. & Anderson, J.L. (1989). Sociobiology: an environmentalist discipline? *American Psychologist* **44**, 1449–1459.

[4] Ellis, B.J. & Garber, J. (2000). Psychosocial antecedents of variation in girls' pubertal timing: maternal depression, stepfather presence, and marital and family stress, *Child Development* **71**, 485–501.

[5] Lykken, D.T., Geisser, S. & Tellegen, A. (1981). Heritability estimates from twin studies: The efficiency of the MZA design, Unpublished manuscript.

[6] McGue, M., & Bouchard Jr, T.J. (1998). Genetic and environmental influences on human behavioral differences, *Annual Review of Neuroscience*, **21**, 1–24.

[7] Mealey, L. (2001). Kinship: the tie that binds (disciplines), in *Conceptual Challenges in Evolutionary Psychology: Innovative Research Strategies*, Holcomb III, H.R. ed., Kluwer Academic Publisher, Dordrecht, pp. 19–38.

[8] Segal, N.L. (2000). *Entwined Lives: Twins and What they Tell us About Human Behavior*, Plume, New York.

[9] Segal, N.L. (2003). Spotlights: reared apart twin researchers, *Twin Research* **6**, 72–81.

[10] Segal, N.L. & Hershberger, S.L. (1999). Cooperation and competition in adolescent twins: findings from a prisoner's dilemma game, *Evolution and Human Behavior* **20**, 29–51.

[11] Segal, N.L., Hershberger, N.L. & Arad, S. (2003). Meeting one's twin: perceived social closeness and familiarity, *Evolutionary Psychology* **1**, 70–95.

[12] Taylor, H.F. (1980). *The IQ Game: A Methodological Inquiry into the Heredity-Environment Controversy*, Rutgers University Press, New Brunswick.

[13] Tellegen, A., Lykken, D.T., Bouchard Jr, , T.J., Wilcox, K.J., Segal, N.L. & Rich, S. (1988). Personality similarity in twins reared apart and together, *Journal of Personality and Social Psychology*, **54**, 1031–1039.

NANCY L. SEGAL

Two by Two Contingency Tables

Li Liu and Vance W. Berger

Volume 4, pp. 2076–2081

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Two by Two Contingency Tables

In a two by two **contingency table**, both the row variable and the column variable have two levels, and each cell represents the count for that specific condition. This type of table arises in many different contexts, and can be generated by different sampling schemes. We illustrate a few examples.

- Randomly sample subjects from a single group and cross classify each subject into four categories corresponding to the presence or absence of each of two conditions. This yields a multinomial distribution, and each cell count is considered to be the result of an independent Poisson process. For example, the 1982 General Social Survey [1] was used to sample and classify individuals by their opinions on gun registration and the death penalty. This sampling scheme yields a multinomial distribution with four outcomes that can be arranged as a 2 × 2 table, as in Table 1. The row variable represents the opinion on gun registration, and the column variable represents the opinion on the death penalty.
- Randomly sample subjects from each of two groups, and classify each of them by a single binary variable. This results in two independent binomial distributions (*see* **Catalogue of Probability Density Functions**). For example, a study was designed to study whether cigarette smoking is related to lung cancer [6]. Roughly equal numbers of lung cancer patients and controls (without lung cancer) were asked whether they smoked or not (see Table 2).
- Randomly assign subjects (selected with or without randomization) to one of two treatments, and then classify each subject by a binary response. For example, a preliminary randomized, placebo-controlled trial (*see* **Clinical Trials and Intervention Studies**) was conducted to determine the antidepressant efficacy of pramipexole in a group of 22 treatment-resistant bipolar depression outpatients [8]. The data are as in Table 3.

Table 1 Opinions on the death penalty cross-classified with opinions on gun registration

Gun	Death penalty		
registration	Favor	Oppose	Total
Favor	784	236	1020
Oppose	311	66	377
Total	1095	302	1397

Table 2 Lung cancer and smoking

	Lung cancer		
Smoker	Yes	No	Total
Yes	647	622	1269
No	2	27	29
Total	649	649	1298

Table 3 Treatment and depression improvement

	Depression improved?		
Treatment	Yes	No	Total
Pramipexole	8	4	12
Placebo	2	8	10
Total	10	12	22

Under the null hypothesis that the two treatments produce the same response in any given subject (that is, that response is an attribute of the subject, independent of the treatments, so that there are some patients destined to respond and others destined not to) [2], the column totals are fixed and the cell counts follow the hypergeometric distribution.

In a two by two table, we are interested in studying whether there is a relationship between the row variable and column variable. If there is an association, then we also want to know how strong it is, and how the two variables are related to each other. The following topics in the paper will help us understand these questions. We use Table 4 to represent any arbitrary two by two table,

Table 4 Generic two-by-two table

	Row variable		
Column variable	Level 1	Level 2	Total
Level 1	n_{11}	n_{12}	n_{1+}
Level 2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

where $n_{11}, n_{12}, n_{21}, n_{22}$ represent the cell counts, $n_{1+}, n_{2+}, n_{+1}, n_{+2}$ represent the row totals and column totals, respectively, and n represents the total count in the table.

χ^2 Test of Independence

For a two by two contingency table, an initial question is whether the row variable and the column variable are independent or not, and the χ^2 test of independence is one method that can be used to answer this question. Given that all the marginal totals (row totals and column totals) are fixed under the null hypothesis that the row variable and column variable are independent, the expected value of n_{ij}, assuming independence of rows and columns, is

$$m_{ij} = \frac{n_{i+}n_{+j}}{n}, \tag{1}$$

and Pearson proposed the test statistic

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}. \tag{2}$$

If the cell counts are large enough, then χ^2 has approximately a chi-square distribution with one degree of freedom. Large values of χ^2 lead to the rejection of the null hypothesis of no association. For contingency Table 1, $\chi^2 = 5.15$ with one degree of freedom, and the P value is 0.0232, which suggests that there is an association between one's opinion on gun registration and one's opinion on the death penalty. Generally, the conventional wisdom is that all the expected cell counts m_{ij} should be larger than five for the test to be valid, and that if some cell counts are small, then Fisher's exact test (*see* **Exact Methods for Categorical Data**) can be used instead. This is not a very sensible plan, however, because it would be quite difficult to justify the use of an approximate test given the availability of the exact test it is trying to approximate [2, 4]. Moreover, it has been demonstrated that even if all expected cell counts exceed five, the approximate test can still give different results from the exact test. Just as it is a better idea to wear a seat belt in all weather rather than just in inclement weather, the safe approach is to select an exact test all the time. Hence Fisher's exact test should be used instead of the chi-square test, for any expected cell counts.

Difference in Proportions

If there is an association based on the chi-square test of independence, or preferably Fisher's exact test, then we may be interested in knowing how the two variables are related. One way is to study the proportions for the two groups, and see how they differ. The difference in proportions is used to compare the conditional (on the row) distributions of a column response variable across the two rows. For these measures, the rows are treated as independent binomial samples. Consider Table 1, as an example. Let π_1 and π_2 represent the probabilities of favoring death penalty for those favoring and opposing gun registration from the population, respectively. Then we are interested in estimating the difference between π_1 and π_2. The sample proportion (using the notation of Table 4) of those favoring the death penalty, among those favoring gun registration, is $p_1 = n_{11}/n_{1+}$, and has expectation π_1 and variance $\pi_1(1 - \pi_1)/n_{1+}$, and the sample proportion of those favoring the death penalty among those opposing gun registration can be computed accordingly. Thus, the difference in proportions has expectation of

$$E(p_1 - p_2) = \pi_1 - \pi_2, \tag{3}$$

and variance (using the notation of Table 4)

$$\sigma^2(p_1 - p_2) = \frac{\pi_1(1 - \pi_1)}{n_{1+}} + \frac{\pi_2(1 - \pi_2)}{n_{2+}}, \tag{4}$$

and the estimated variance (using the notation of Table 4) is

$$\hat{\sigma}^2(p_1 - p_2) = \frac{p_1(1 - p_1)}{n_{1+}} + \frac{p_2(1 - p_2)}{n_{2+}}. \tag{5}$$

Then a $100(1 - \alpha)$ % **confidence interval** for $\pi_1 - \pi_2$ is

$$p_1 - p_2 \pm z_{\alpha/2}\hat{\sigma}(p_1 - p_2). \tag{6}$$

For Table 1, we have

$$\begin{aligned} p_1 - p_2 &= \frac{784}{1020} - \frac{311}{377} = 0.7686 - 0.8249 \\ &= -0.0563, \end{aligned} \tag{7}$$

and

$$\hat{\sigma}^2(p_1 - p_2) = \frac{0.7686(1 - 0.7686)}{1020} + \frac{0.8249(1 - 0.8249)}{377} = 0.000557. \quad (8)$$

So the 95% confidence interval for the true difference is $-0.0563 \pm 1.96(0.0236)$, or $(-0.010, -0.103)$. Since this interval contains only negative values, we can conclude that $\pi_1 - \pi_2 < 0$, so people who favor gun registration are less likely to favor the death penalty.

Relative Risk and Odds Ratio

To compare how the two groups differ, we can use the difference of the two proportions. Naturally, we can also use the ratio of the two proportions, and this is called the **relative risk**. The estimated relative risk is

$$RR = \frac{p_1}{p_2} = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}, \quad (9)$$

and the estimated variance of log (*RR*) is

$$\hat{\sigma}^2(\log(RR)) = \frac{1}{n_{11}} - \frac{1}{n_{1+}} + \frac{1}{n_{21}} - \frac{1}{n_{2+}}. \quad (10)$$

Thus a $100(1 - \alpha)$ % confidence interval for the relative risk π_1/π_2 is

$$\exp(RR \pm z_{\alpha/2}\hat{\sigma}(\log(RR))). \quad (11)$$

Instead of computing the ratio of the proportion of yes for group 1 versus group 2, we can also compute the ratio of the odds of yes for group 1 versus for group 2. This is called the **odds ratio**. The estimated odds ratio is

$$OR = \frac{p_1/(1 - p_1)}{p_2/(1 - p_2)} = \frac{n_{11}n_{22}}{n_{12}n_{21}}, \quad (12)$$

and the estimated variance of log(*OR*) is

$$\hat{\sigma}^2(\log(OR)) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}. \quad (13)$$

Then a $100(1 - \alpha)$ % confidence interval for odds ratio $\pi_1/(1 - \pi_1)/\pi_2/(1 - \pi_2)$ is

$$\exp(OR \pm z_{\alpha/2}\hat{\sigma}(\log(OR))). \quad (14)$$

The sample odds ratio is either to 0 or ∞ if any $n_{ij} = 0$, and it is undefined if both entries are 0 for a row or column. To solve this problem, a modified estimate can be used by adding 1/2 to each cell count, and its variance can be estimated accordingly. The relationship between the odds ratio and the relative risk is shown in the following formula

$$RR = OR \times \frac{1 - p_1}{1 - p_2}. \quad (15)$$

Unlike the relative risk, the odds ratio can be used to measure an association no matter how the data were collected. This is very useful for rare disease retrospective studies such as the study in Table 2. In such a study, we cannot obtain the relative risk directly, however, we can still compute the odds ratio. Since $1 - p_1 \approx 1$ and $1 - p_2 \approx 1$ for rare diseases, the relative risk and odds ratio are numerically very close in such studies, and so the odds ratio can be used to estimate the relative risk. For example, if $p_1 = 0.01$ and $p_2 = 0.001$, then we have

$$RR = \frac{p_1}{p_2} = 10, \quad \text{and}$$

$$OR = \frac{p_1/(1 - p_1)}{p_2/(1 - p_2)} = \frac{0.01/0.99}{0.001/0.999} = 10.09, \quad (16)$$

which are very close. For the data in Table 2, we have

$$OR = \frac{P(E/D)/P(\bar{E}/D)}{P(E/\bar{D})/P(\bar{E}/\bar{D})} = \frac{P(E \cap D)/P(\bar{E} \cap D)}{P(E \cap \bar{D})/P(\bar{E} \cap \bar{D})}$$
$$= \frac{P(D/E)/P(D/\bar{E})}{P(\bar{D}/E)/P(\bar{D}/\bar{E})} \approx \frac{P(D/E)}{P(D/\bar{E})} = RR, \quad (17)$$

since $P(\bar{D}/E)$and$P(\bar{D}/\bar{E})$ are almost 1 for rare disease. So

$$RR \approx OR = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{647 \times 27}{622 \times 2} = 14.04. \quad (18)$$

This statistic indicates that the risk of getting lung cancer is much higher for smokers than it is for nonsmokers.

Sensitivity, Specificity, False Positive Rate, and False Negative Rate

These measures are commonly used when evaluating the efficacy of a screening test for a disease outcome.

Table 5 VAP diagnosis results

	Chest radiograph		
Disease status	Diagnosis +	Diagnosis −	Total
VAP	12	1	13
No VAP	8	4	12
Total	20	5	25

Table 5 contains a study assessing the accuracy of chest radiograph (used to detect radiographic infiltrates) for the diagnosis of ventilator associated pneumonia (VAP) [7]. The sensitivity is the true proportion of positive diagnosis results among the VAP patients, and the specificity is the true proportion of negative diagnosis results among those without VAP. Both the sensitivity and the specificity can be estimated by

$$\begin{aligned}\text{Sensitivity} &= \frac{n_{11}}{n_{1+}} = \text{P(Diagnosis} + |\text{VAP)}\\ \text{Specificity} &= \frac{n_{22}}{n_{2+}} = \text{P(Diagnosis} - |\text{No VAP)}\end{aligned} \tag{19}$$

For the data in Table 5, sensitivity = 12/13 = 0.92 and specificity = 4/12 = 0.33.

Sensitivity is also called the *true positive rate*, and specificity is also called the *true negative rate*. They are closely related to two other two rates, specifically the false positive rate and the false negative rate. The false negative rate is the true proportion of negative diagnosis results among the VAP patients, and the false positive rate is the true proportion of positive diagnosis results among those without VAP. The false positive rate and false negative can be estimated by

$$\begin{aligned}\text{False positive rate} &= \frac{n_{21}}{n_{2+}}\\ &= \text{P(Diagnosis} + |\text{No VAP)}\\ &= 1 - \text{specificity}\\ \text{False negative rate} &= \frac{n_{12}}{n_{1+}}\\ &= \text{P(Diagnosis} - |\text{VAP)}\\ &= 1 - \text{sensitivity.,}\end{aligned} \tag{20}$$

Another useful term is the false discovery rate, which is the proportion of subjects without VAP among all the positive diagnosis results.

Fisher's Exact Test

When the sample size is small, the χ^2 test based on large samples may not be valid, and Fisher's exact test can be used to test the independence of a contingency table. For given row and column totals, the value n_{11} determines the other three cell counts (there is but one degree of freedom). Under the null hypothesis of independence, the probability of a particular value n_{11} given the marginal totals is

$$P(n_{11}) = \frac{(n_{1+}!n_{2+}!)(n_{+1}!n_{+2}!)}{n!(n_{11}!n_{12}!n_{21}!n_{22}!)}, \tag{21}$$

which is the hypergeometric probability function (*see* **Catalogue of Probability Density Functions**). One would enumerate all possible tables of counts consistent with the row and column totals n_{i+} and n_{+j}. For each one, the associated conditional probability can be calculated using the above formula, and the sum of these probabilities must be one. To test independence, the P value is the sum of the hypergeometric probabilities for those tables at least as favorable to the alternative hypothesis as the observed one. That is, for a given two by two table, the P value of the Fisher exact test is the sum of all the conditional probabilities that correspond to tables that are as extreme as or more extreme than the observed table. Consider Table 3, for example. The null distribution of n_{11} is the hypergeometric distribution defined for all the two by two tables having row totals and column totals (12,10) and (10,12). The potential values for n_{11} are $(0, 1, 2, 3, \ldots, 10)$. First, we can compute the probability of the observed table as

$$P(4) = \frac{(12!10!)(10!12!)}{22!(8!4!2!8!)} = 0.0344. \tag{22}$$

Other possible two by two tables and their probabilities are

$$\begin{aligned}&\begin{bmatrix}10 & 2\\ 0 & 10\end{bmatrix} \text{Prob} = 0.0001 \quad \begin{bmatrix}9 & 3\\ 1 & 9\end{bmatrix} \text{Prob} = 0.0034\\ &\begin{bmatrix}7 & 5\\ 3 & 7\end{bmatrix} \text{Prob} = 0.1470 \quad \begin{bmatrix}6 & 6\\ 4 & 6\end{bmatrix} \text{Prob} = 0.3001\\ &\begin{bmatrix}5 & 7\\ 5 & 5\end{bmatrix} \text{Prob} = 0.3086 \quad \begin{bmatrix}4 & 8\\ 6 & 4\end{bmatrix} \text{Prob} = 0.1608\\ &\begin{bmatrix}3 & 9\\ 7 & 3\end{bmatrix} \text{Prob} = 0.0408 \quad \begin{bmatrix}2 & 10\\ 8 & 2\end{bmatrix} \text{Prob} = 0.0046\\ &\begin{bmatrix}1 & 11\\ 9 & 1\end{bmatrix} \text{Prob}=0.0002 \quad \begin{bmatrix}0 & 12\\ 10 & 0\end{bmatrix} \text{Prob}=0.0000015\end{aligned} \tag{23}$$

Together with the observed probability, these probabilities sum up to one. The sum of the probabilities of the tables in the two-sided rejection region is 0.0427. This rejects the null hypothesis of independence (at the customary 0.05 level), and suggests that treatment and improvement are correlated. The one-sided rejection region would consist of the tables with upper-left cell counts of 10, 9, and 8, and the one-sided P value is $0.0001 + 0.0034 + 0.0344 = 0.0379$. Fisher's exact test can be conservative, and so one can use the **mid-*P* value** or the P value interval [3].

Simpson's Paradox

For a $2 \times 2 \times 2$ contingency table, **Simpson's Paradox** refers to the situation in which a marginal association has a different direction from the conditional associations (*see* **Measures of Association**). In a $2 \times 2 \times 2$ contingency table, if we cross classify two variables X and Y at a fixed level of binary variable Z, we can obtain two 2×2 tables with variables X and Y and they are called *partial tables*. If we combine the two partial tables, we can obtain one 2×2 table and this is called the *marginal table*. The associations in the partial tables are called *partial associations*, and the association in the marginal table is called *marginal association*. Table 6 is a $2 \times 2 \times 2$ contingency table that studied the relationship between urinary tract infections (UTI) and antibiotic prophylaxis (ABP) [9].

The last section of Table 6 displays the marginal association. As we can see, 3.28% of the patients who used antibiotic prophylaxis got UTI, and 4.64% of patients who did not use antibiotic prophylaxis got UTI. Clearly, the probability of UTI is lower for those who used antibiotic prophylaxis than it is for those who did not. This is consistent with previous findings that antibiotic prophylaxis is effective in preventing UTI. But now consider the first two partial tables in Table 6, which display the conditional associations between antibiotic prophylaxis and UTI. When the patients were from four hospitals with low incidence of UTI, the probability of UTI is higher for those who used antibiotic prophylaxis $1.80\% - 0.70\% = 1.10\%$. It might seem that to compensate for this reversed effect, the patients from the four hospitals with high incidence of UTI would have shown a very strong trend in the direction of higher UTI incidence for those without the antibiotic prophylaxis. But alas this was not the case. In fact, the probability of UTI is higher for those who used antibiotic prophylaxis $13.25\% - 6.51\% = 6.74\%$.

Thus, controlling for hospital type, the probability of UTI is higher for those who used antibiotic prophylaxis. The partial association gives a different direction of association compared to the marginal associations. This is called *Simpson's Paradox* [9]. The reason that the marginal association and partial associations have different directions is related to the association between the control variable, hospital type, and the other two variables. First, consider the association between hospital type and the usage of antibiotic prophylaxis based on the marginal table with these two variables. The odds ratio equals

$$\frac{1113 \times 1520}{(720 \times 166)} = 14.15, \tag{24}$$

which indicates a strong association between hospital type and the usage of antibiotic prophylaxis. Patients from the four hospitals with low incidence of UTI were more likely to have used antibiotic prophylaxis. Second, the probability of UTI is tautologically higher for the four high incidence hospitals. The odds

Table 6 Urinary tract infections (UTI) and antibiotic prophylaxis (ABP)

Hospital	Antibiotic prophylaxis	UTI? Yes	UTI? No	Percentage
Patients from four hospitals with	Yes	20	1093	1.80%
Low incidence of UTI ($\leq 2.5\%$)	No	5	715	0.70%
Patients from four hospitals with	Yes	22	144	13.25%
High incidence of UTI ($> 2.5\%$)	No	99	1421	6.51%
Total	Yes	42	1237	3.28%
	No	104	2136	4.64%

ratio based on the marginal table of hospital and UTI is

$$\frac{25 \times 1565}{(1808 \times 121)} = 0.18. \tag{25}$$

An explanation of the contrary results in Simpson's Paradox is that there are other confounding variables that may have been unrecognized.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*, John Wiley & Sons, New York.

[2] Berger, V.W. (2000). Pros and cons of permutation tests in clinical trials, *Statistics in Medicine* **19**, 1319–1328.

[3] Berger, V.W. (2001). The p-Value Interval as an Inferential Tool, *Journal of the Royal Statistical Society D (The Statistician) 50*, **1**, 79–85.

[4] Berger, V.W., Lunneborg, C., Ernst, M.D. & Levine, J.G. (2002). Parametric analyses in randomized clinical trials, *Journal of Modern Applied Statistical Methods* **1**, 74–82.

[5] Berger, V.W. (2004),. Valid Adjustment of Randomized Comparisons for Binary Covariates, *Biometrical Journal* **46**(5), 589–594.

[6] Doll, R. & Hill, A.B. (1950). Smoking and carcinoma of the lung; preliminary report, *British Medical Journal* **2**(4682), 739–748.

[7] Fabregas, N., Ewig, S., Torres, A., El-Ebiary, M., Ramirez, J., de La Bellacasa, J.P., Bauer, T., Cabello, H., (1999). Clinical diagnosis of ventilator associated pneumonia revisited: comparative validation using immediate post-mortem lung biopsies, *Thorax* **54**, 867–873.

[8] Goldberg, J.F., Burdick, K.E. & Endick, C.J. (2004). Preliminary randomized, double-blind, placebo-controlled trial of pramipexole added to mood stabilizers for treatment-resistant bipolar depression, *The American Journal of Psychiatry* **161**(3), 564–566.

[9] Reintjes, R., de Boer, A., van Pelt, W. & Mintjes-de Groot, J. (2000). Simpson's paradox: an example from hospital epidemiology, *Epidemiology* **11**(1), 81–83.

Li Liu and Vance W. Berger

Two-mode Clustering

Iven Van Mechelen, Hans-Hermann Bock and Paul de Boeck

Volume 4, pp. 2081–2086

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Two-mode Clustering

Data in the behavioral sciences often can be written in the form of a rectangular matrix. Common examples include observed performances denoted in a person by task matrix, questionnaire data written in a person by item matrix, and semantic information written in a concept by feature matrix. Rectangular matrix data imply two sets of entities or *modes*, the row mode and the column mode.

Within the family of two-mode data, several subtypes may be distinguished. Common subtypes include case by variable data that denote the value (data entry) that each of the variables (columns) under study takes for each case (row) under study, contingency table data (*see* **Contingency Tables**) that denote the frequency (data entry) with which each combination of a row entity and a column entity is being observed, and categorical predictor/criterion data that denote the value of some criterion variable (data entry) observed for each pair of values of two categorical predictor variables (corresponding to the elements of the row and column modes).

Two-mode data may be subjected to various kinds of clustering methods. Two major subtypes are *indirect* and *direct approaches*. Indirect clustering methods presuppose the conversion of the two-mode data into a set of (one-mode) proximities (similarities or dissimilarities *see* **Proximity Measures**) among the elements of either the row or the column mode; this conversion is usually done as a separate step prior to the actual **cluster analysis**. Direct clustering methods, on the other hand, directly operate on the two-mode data without a preceding proximity transformation. Several direct clustering methods yield a clustering of the elements of *only one of the modes* of the data. Optionally, such methods can be applied twice to yield successively a clustering of the row entities and a clustering of the column entities.

As an alternative, one may wish to rely on direct *two-mode clustering methods*. Such methods yield clusterings of rows and columns *simultaneously* rather than successively. The most important advantage of simultaneous approaches is that they may reveal information on the linkage between the two modes of the data.

As most methods of data analysis, two-mode clustering methods imply a reduction of the data, and, hence, a loss of information. The goal of the clustering methods, however, is that the loss is as small as possible with regard to the particular subtype of information that constitutes the target of the clustering method under study, and into which the method aims at providing more insight. For case by variable data, the target information typically consists of the actual values the variables take for each of the cases; two-mode clustering methods for this type of data aim at reconstructing these values as well as possible. Furthermore, for contingency table data, the target information typically pertains to the amount of dependence between the row and column modes, whereas for categorical predictor/criterion data it consists of the amount of interaction as implied by the prediction of the data entries on the basis of the categorical row and column variables.

Basic Two-mode Clustering Concepts

Since the pioneering conceptual and algorithmic work by Hartigan [3] and Bock [1], a large number of quite diverse simultaneous clustering methods has been developed. Those range from heuristic *ad hoc* procedures, over deterministic structures estimated in terms of some objective or loss function, to fully stochastic model-based approaches. A structured overview of the area can be found in [4].

To grasp the multitude of methods, two conceptual distinctions may be useful:

1. The *nature of the clusters*: a cluster may be a set of row elements (*row cluster*), a set of column elements (*column cluster*), or a Cartesian product of a set of row elements and a set of column elements (*data cluster*). One may note that each data cluster as obtained from a two-mode clustering procedure always implies a row and a column cluster; the reverse, however, does not necessarily hold.
2. The *set-theoretical structure* of a particular set of clusters or clustering (see also Figure 1, to be discussed below): this may be (a) a partitioning, (b) a nested clustering (i.e., a clustering that includes intersecting clusters, albeit such that intersecting clusters are always in a subset-superset relation), and (c) an overlapping clustering (i.e., a clustering that includes intersecting, nonnested clusters).

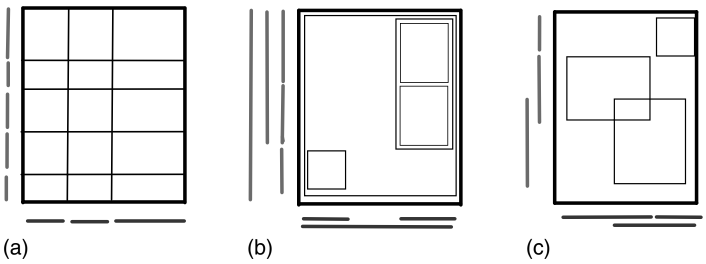


Figure 1 Schematic representation of hypothetical examples of three types of two-mode clustering: (a) partitioning, (b) nested clustering, (c) overlapping clustering

The data clustering constitutes the cornerstone of any two-mode cluster analysis. As such it is the key element in the representation of the value, dependence, or interaction information in the data. For case by variable data, this representation (and, more in particular, the reconstruction of the data values) further relies on parameters (scalar or vector) associated with the data clusters. Otherwise, we will limit the remainder of this exposition to case by variable data.

In general, two-mode clustering methods may be considered that yield row, column and data clusterings with different set-theoretical structures as distinguished above. However, here we will consider further only methods that yield row, column, and data clusterings of the same type. Taking into account the three possible set-theoretical structures, we will therefore focus on the three types of clustering as schematically presented in Figure 1.

In what follows, we will briefly present one instance of each type of clustering method. Each instance will further be illustrated making use of the same 14 × 11 response by situation data matrix obtained from a single participant who was asked to rate the applicability of each of 14 anxiety responses to each of 11 stressful situations, on a 5-point scale ranging from 1 (=not applicable at all) to 5 (=applicable to a strong extent).

Partitioning

Two-mode partitioning methods of a data matrix **X** imply a partitioning of the Cartesian product of the row and column modes that is obtained by fully crossing a row and a column partitioning (see leftmost panel in Figure 1). In one common instance of this class of methods, each data cluster A × B is associated with a (scalar) parameter $\mu_{A,B}$. The clustering and parameters are further such that all entries x_{ab} (with $a \in A, b \in B$) are as close as possible to the corresponding value $\mu_{A,B}$ (which acts as the reconstructed data value). This implies that row entries of the same row cluster behave similarly across columns, that column entries of the same column cluster behave similarly across rows, and that the data values are as homogeneous as possible within each data cluster.

The result of a two-mode partitioning of the anxiety data is graphically represented in Figure 2. The representation is one in terms of a so-called *heat map*, with the estimated $\mu_{A,B}$-parameters being represented in terms of grey values. The analysis reveals, for instance, an avoidance behavior class (third row cluster) that is associated fairly strongly with a class of situations that imply some form of psychological assessment (third column cluster).

Nested Clustering

One instance in this class of methods is *two-mode ultrametric tree modeling*. In this method, too, each data cluster A × B is associated with a (scalar) parameter $\mu_{A,B}$. As for all two-mode clustering methods for case by variable data that are not partitions, ultrametric tree models include a rule for the reconstruction of data entries in intersections of different data clusters. In the ultrametric tree model, the data reconstruction rule makes use of the Maximum (or Minimum) operator. In particular, the clustering and $\mu_{A,B}$-parameters are to be such that all entries x_{ab} are as close as possible to the maximum (or minimum) of the $\mu_{A,B}$-values of all data clusters $A \times B$ to which (a, b) belongs.

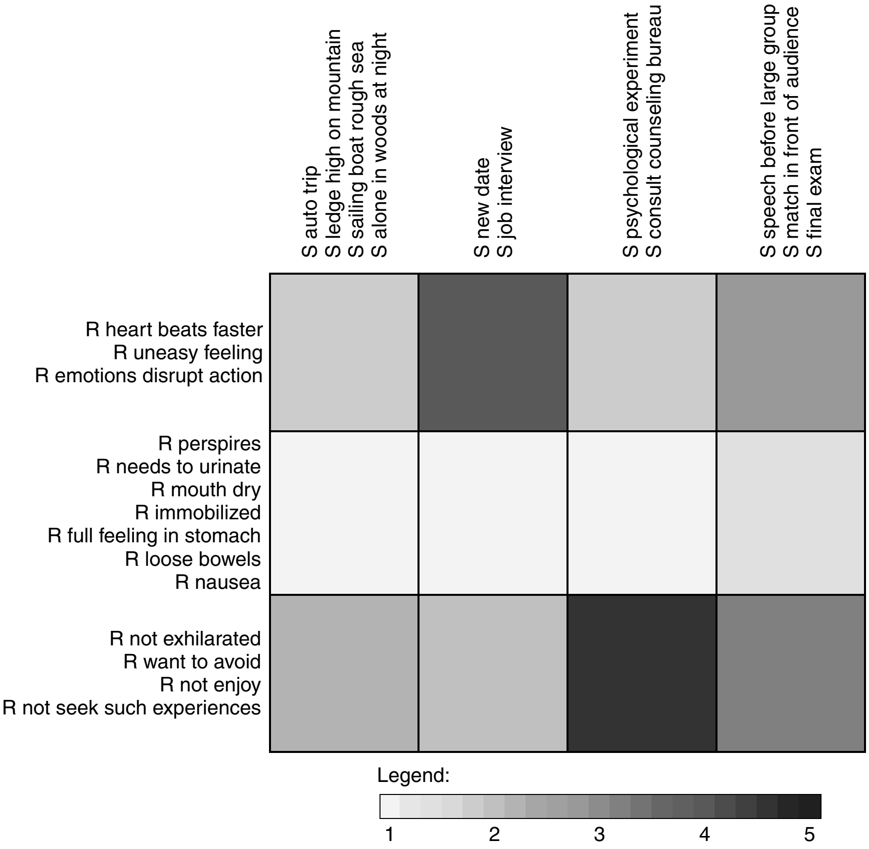


Figure 2 Two-mode partitioning of anxiety data

Optionally, a data cluster can be interpreted as a feature that applies to the row and column entries involved in it.

Part of the result of a two-mode ultrametric tree analysis of the anxiety data (making use of a constrained optimization algorithm) is graphically represented as a tree diagram in Figure 3. (The representation is limited to a subset of situations and responses to improve readability.) In Figure 3, the data clusters correspond to the vertical lines and comprise all leaves at the right linked to them. The $\mu_{A,B}$-values can further be read from the hierarchical scale at the bottom of the figure. As such, one can read from the lower part of Figure 3 that the two psychological assessment-related situations ‘consult counseling bureau’ and ‘psychological experiment’ elicit ‘not enjoy’ and ‘not feel exhilarated’ with strength 5, and ‘not seek experiences like this’ with strength 4.5.

Overlapping Clustering

Hierarchical classes models [2] are overlapping two-mode clustering methods for case by variable data that make use of a data reconstruction rule with a Maximum (or Minimum) operator. In case of positive real-valued data, the hierarchical classes model can be considered a generalization of the two-mode ultrametric tree, with each data cluster again being associated with a $\mu_{A,B}$-parameter, and with the clustering and parameters being such that all entries x_{ab} are as close as possible to the maximum of the $\mu_{A,B}$-values of all data clusters $A \times B$ to which (a, b) belongs. The generalization implies that row, column, and data clusters are allowed to overlap. The row and column clusters of the model then can be considered to be (possibly overlapping) types that are associated with a value of association strength as specified in the model. A distinctive feature of

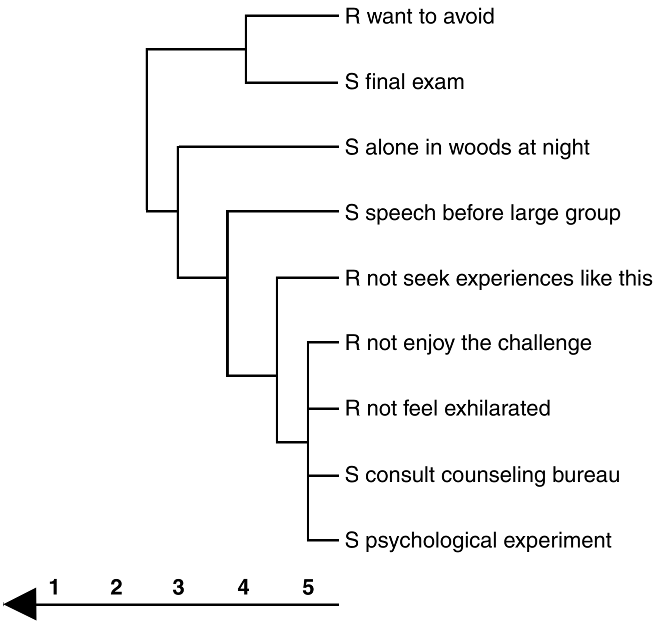

R want to avoid
S final exam
S alone in woods at night
S speech before large group
R not seek experiences like this
R not enjoy the challenge
R not feel exhilarated
S consult counseling bureau
S psychological experiment
1
2
3
4
5

Figure 3 Part of a two-mode ultrametric tree for anxiety data

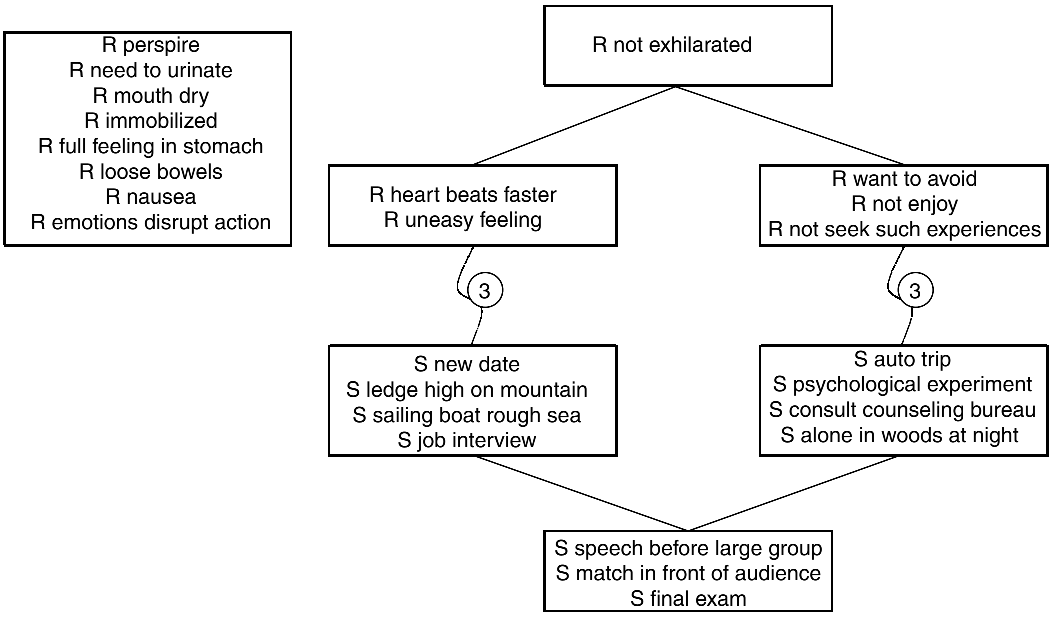


Figure 4 Overlapping clustering (HICLAS-R model) of anxiety data

hierarchical classes models further reads that they represent implicational if-then type relations among row and among column elements.

The result of a hierarchical classes analysis of the anxiety data is presented in Figure 4. A data cluster can be read from this figure as the set of all situations and responses linked to a particular zigzag; the $\mu_{A,B}$-value associated with this cluster is further attached as a label to the zigzag. The represented model, for instance, implies that the two psychological assessment-related situations elicit 'not exhilarated' and 'not enjoy' with value 3 (encircled number). From the representation, one may further read that, for instance, if a situation elicits 'not enjoy' (with some association value) then it also elicits 'not exhilarated' (with at least the same value of association strength) (*see* **Cluster Analysis: Overview**; **Overlapping Clusters**).

Software

Two-mode clustering methods mostly have not been incorporated in general purpose statistical packages. Exceptions include variants of a nested method (two-way joining) due to Hartigan as incorporated within *Statistica*™ and *Systat*™, and an overlapping method (Boolean factor analysis) as incorporated within *BMDP*™. A number of two-mode clustering methods are available from their developers as stand-alone programs. Finally, more specific packages are being developed within specialized areas such as bioinformatics, in particular, for applications in microarray data analysis (*see* **Microarrays**).

References

[1] Bock, H.-H. (1968). *Stochastische Modelle für die einfache und doppelte Klassifikation von normalverteilten Beobachtungen*, Dissertation, University of Freiburg, Germany.

[2] De Boeck, P. & Rosenberg, S. (1988). Hierarchical classes: model and data analysis, *Psychometrika* **53**, 361–381.

[3] Hartigan, J. (1972). Direct clustering of a data matrix, *Journal of the American Statistical Association* **67**, 123–129.

[4] Van Mechelen, I., Bock, H.-H. & De Boeck, P. (2004). Two-mode clustering methods: a structured overview, *Statistical Methods in Medical Research* **13**, 363–394.

IVEN VAN MECHELEN, HANS-HERMANN BOCK
AND PAUL DE BOECK

Two-way Factorial: Distribution-Free Methods

Larry E. Toothaker

Volume 4, pp. 2086–2089

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Two-way Factorial: Distribution-Free Methods

The classical tests for main effects and interaction in a $J \times K$ two-way **factorial design** are the F tests in the two-way fixed-effects **analysis of variance (ANOVA)** (*see* **Fixed and Random Effects**), with the assumptions of normality, independence, and equal variances of errors in the $J \times K$ cells. When all assumptions are met and the null hypotheses are true, the **sampling distribution** of each F is distributed as a theoretical F distribution with appropriate degrees of freedom, in which the F-critical values cut off exactly the set α in the upper tail.

When any assumption is violated, the sampling distributions might not be well-fit by the F distributions, and the F-critical values might not cut off exactly α. Also, violation of assumptions can lead to poor **power** properties, such as when the power of F decreases as differences in means increase. The extent to which the F-statistics are resistant to violations of the assumptions is called *robustness*, and is often measured by how close the true α in the sampling distribution is to the set α, and by having power functions where power increases as mean differences increase. Also, power should be compared for different statistical methods in the same circumstances, with preference being given to those methods that maintain control of α and have the best power.

Alternatives to relying on the robustness of the F tests might be available in the area of nonparametric methods. These methods are so named because their early ancestors were originally designed to test hypotheses that had no parameters, but were tests of equality of entire distributions. Many of the modern nonparametric methods test hypotheses about parameters, although often not the means nor treatment effects of the model. They also free the researcher from the normality assumption of the ANOVA (*see* **Distribution-free Inference, an Overview**).

Such alternative methods include **permutation tests** [5], **bootstrap** tests [6], the rank transform [7], aligned ranks tests [9], tests on **trimmed means** [21], and a rank-based ANOVA method that allows heteroscedastic variances, called the *BDM method* (after the surnames of the authors of [4]). Some of these are more general procedures that can be useful when combined with any statistic, for example, the bootstrap method might be combined with tests on trimmed means, using the bootstrap to make the decision on the hypothesis rather than using a standard critical value approach.

Permutation Tests

Permutation tests (*see* **Linear Models: Permutation Methods**) rely on permuting the observations among treatments in all ways possible, given the design of the study and the treatments to be compared. For each permutation, a treatment comparison statistic is computed, forming a null reference distribution. The proportion of reference statistics equal to or more extreme than that computed from the original data is the P value used to test the null hypothesis. If the number of permissible permutations is prohibitively large, the P value is computed from a large random sample of the permutations.

Using a simple 2×2 example with $n = 2$ per cell, if the original data were as in Table 1, then one permutation of the scores among the four treatment combinations would be as in Table 2.

Permutation tests rely on the original sampling, so the permutation illustrated above would be appropriate for a completely randomized two-way design. As the random sample of subjects was randomly assigned two to each cell, any permutation of the results, two scores to each cell would be permissible. Observations may be exchanged between any pair of cells. For randomized block designs, however, the only permissible permutations would be those in

Table 1 Original data

1 2	3 4
5 6	7 8

Table 2 Permutation of data

3 2	1 7
8 6	4 5

Table 3 Bootstrap sample of data

3	1
1	7
8	6
8	5

which scores are exchanged between treatments but within the same block. Software for permutation tests includes [5, 12, 14, 16, and 17].

Bootstrap Tests

Bootstrap tests follow a pattern similar to permutation tests except that the empirical distribution is based on B random samples from the original data rather than all possible rearrangements of the data. Put another way, the bootstrap uses sampling with replacement. So from the original data above, a possible sample with replacement would be as in Table 3.

A generally useful taxonomy is given by [13] for some of these methods that use an empirically generated sampling distribution. Consider these methods as resampling methods, and define the sampling frame from which the sample is to be selected as the set of scores that was actually observed. Then, resampling can be done without replacement (the permutation test) or with replacement (the bootstrap). Both of these use the original sample size, n, as the sample size for the resample (using a subsample of size $m < n$ leads to the **jackknife**).

The percentile bootstrap method for the two-way **factorial design** is succinctly covered in [21, p. 350]. For broader coverage, [6] and [10] give book-length coverage of the topic, including the percentile bootstrap method. Software for bootstrap tests includes [12], and macros from `www.sas.com` [15].

Rank Transform

The concept behind the rank transform is to substitute ranks for the original observations and compute the usual statistic (*see* **Rank Based Inference**). While this concept is simple and intuitively appealing, and while it works in some simpler settings (including the correlation, the two-independent-sample case, and the one-way design), it has some problems when applied to the two-way factorial design. Notably, the ranks are not independent, resulting in F tests that do not maintain α-control. Put another way, the linear model and the F tests are not invariant under the rank transformation. Ample evidence exists that the rank transform should not be considered for the two-way factorial design [2, 18, 19, and 20] but it persists in the literature because of suggestions like that in [15].

Indeed, documentation for the SAS programming language [15] suggest that a wide range of **linear model** hypotheses can be tested nonparametrically by taking ranks of the data (using the RANK procedure) and using a regular parametric procedure (such as GLM or ANOVA) to perform the analysis. It is likely that these tests are as suspect in the wider context of linear models as for the $J \times K$ factorial design.

Other authors [11] have proposed analogous rank-based tests that rely on chi-squared distributions, rather than those of the t and F random variables. However, [20] establishes that these methods have problems similar to those that employ F tests.

Aligned Ranks Tests

When one or more of the estimated sources in the linear model are first subtracted from the scores, and the subsequent residuals are then ranked, the scores are said to have been aligned, and the tests are called *aligned ranks tests* [9]. Results for the two-way factorial design show problems with liberal α for some F tests in selected factorial designs with cell sizes of 10 or fewer [20]. However, results from other designs, such as the factorial **Analysis of Covariance (ANCOVA)**, are promising for larger cell sizes, that is, 20 or more [8].

Tests on Trimmed Means

For a method that is based on robust estimators, tests on 20% **trimmed means** are an option. For the ordered scores in each cell, remove 20% of both the largest and smallest scores, and average the remaining 60% of the scores, yielding a 20% trimmed mean. For example, if $n = 10$, then $0.2(n) = 0.2(10) = 2$. If the data for one cell are

$$23\ 24\ 26\ 27\ 34\ 35\ 38\ 45\ 46\ 56,$$

then the process to get the 20% trimmed mean would be to remove the 23 and 24 and the 46 and 56. Then the 20% trimmed mean would sum 26 through 45 to get 205, then divide by $0.6n$ to get $205/6 = 34.17$.

The numerators of the test statistics for the main effects and interactions are functions of these trimmed means. The denominators of the test statistics are functions of 20% Winsorized variances (*see* **Winsorized Robust Measures**).

To obtain Winsorized scores for the data in each cell, instead of removing the 20% largest and smallest scores, replace the 20% smallest scores with the next score up in order, and replace the 20% largest scores with the next score down in order. Then compute the unbiased sample variance on these Winsorized scores.

For the data above, the 20% Winsorized scores would be

26 26 26 27 34 35 38 45 45 45

and the 20% Winsorized variance would be obtained by computing the unbiased sample variance of these $n = 10$ Winsorized scores, yielding 68.46.

For complete details on heteroscedastic methods using 20% trimmed means and 20% Winsorized variances, see [21]. While t Tests or F tests based on trimmed means and Winsorized variances are not nonparametric methods, they are robust to nonnormality and unequal variances, and provide an alternative to the two-way ANOVA. Of course, when combined with, say, bootstrap sampling distributions, tests based on trimmed means take on the additional property of being nonparametric.

BDM

Heteroscedastic nonparametric F tests for factorial designs that allow for tied values and test hypotheses of equality of distributions were developed by [4]. This method, called the BDM method (after the authors of [4]), is based on F distributions with estimated degrees of freedom generalized from [3]. All of the N observations are pooled and ranked, with tied ranks resolved by the mid-rank solution where the tied observations are given the average of the ranks among the tied values. Computation of the subsequent ANOVA-type rank statistics (BDM tests) is shown in [21, p. 572]. Simulations by [4] showed these BDM tests to have adequate control of α and competitive power.

Hypotheses

The hypotheses tested by the factorial ANOVA for the A main effect, B main effect, and interaction, are, respectively,

$$\begin{aligned}
&\mathrm{H_o}\colon \alpha_j = \mu_j - \mu = 0 \text{ for all } j = 1 \text{ to } J\\
&\mathrm{H_o}\colon \alpha_k = \mu_k - \mu = 0 \text{ for all } k = 1 \text{ to } K\\
&\mathrm{H_o}\colon \alpha\beta_{jk} = \mu_{jk} - \mu_j - \mu_k + \mu = 0 \text{ for all}\\
&\qquad j = 1 \text{ to } J, \text{ for all } k = 1 \text{ to } K. \qquad (1)
\end{aligned}$$

Some of the rank procedures might claim to test hypotheses where rank mean counterparts are substituted for the appropriate μ's in the above hypotheses, but in reality they test truly nonparametric hypotheses. Such hypotheses are given as a function of the cumulative distribution for each cell, $F_{jk}(x)$, see [1]. $F_{j.}$ is the average of the $F_{jk}(x)$ across the K levels of B, $F_{.k}$ is the average of the $F_{jk}(x)$ across the J levels of A, and $F_{..}$ is the average of the $F_{jk}(x)$ across the JK cells. Then the hypotheses tested by these nonparametric methods for the A main effect, B main effect, and interaction, are, respectively,

$$\begin{aligned}
&\mathrm{H_o}\colon \mathrm{A}_j = F_{j.} - F_{..} = 0 \text{ for all } j = 1 \text{ to } J\\
&\mathrm{H_o}\colon \mathrm{B}_k = F_{.k} - F_{..} = 0 \text{ for all } k = 1 \text{ to } K\\
&\mathrm{H_o}\colon \mathrm{AB}_{jk} = F_{jk}(x) - F_{j.} - F_k + F_{..} = 0 \text{ for all}\\
&\qquad j = 1 \text{ to } J, \text{ for all } k = 1 \text{ to } K. \qquad (2)
\end{aligned}$$

So the permutation tests, bootstrap tests, aligned ranks tests, and BDM all test this last set of hypotheses. The tests based on trimmed means test hypotheses about population trimmed means, unless, of course, they are used in the bootstrap. Practically, the nonparametric tests allow a researcher to say that the distributions are different, without specifying a particular parameter for the difference.

References

[1] Akritas, M.G., Arnold, S.F. & Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs, *Journal of the American Statistical Association* **92**, 258–265.

[2] Blair, R.C., Sawilowsky, S.S. & Higgins, J.J. (1987). Limitations of the rank transform statistic in tests for interaction, *Communications in Statistics-Simulation and Computation* **16**, 1133–1145.

[3] Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I: effect of inequality of variance in the one-way classification, *The Annals of Mathematical Statistics* **25**, 290–302.

[4] Brunner, E., Dette, H. & Munk, A. (1997). Box-type approximations in nonparametric factorial designs, *Journal of the American Statistical Association* **92**, 1494–1502.

[5] Edgington, E.S. (1987). *Randomization Tests*, Marcel Dekker, New York.

[6] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.

[7] Headrick, T.C. & Sawilowsky, S.S. (2000). Properties of the rank transformation in factorial analysis of covariance, *Communications in Statistics: Simulation and Computation* **29**, 1059–1088.

[8] Headrick, T.C. & Vineyard, G. (2001). An empirical investigation of four tests for interaction in the context of factorial analysis of covariance, *Multiple Linear Regression Viewpoints* **27**, 3–15.

[9] Hettmansperger, T.P. (1984). *Statistical Inference Based on Ranks*, Wiley, New York.

[10] Lunneborg, C.E. (2000). *Data Analysis by Resampling: Concepts and Applications*, Duxbury Press, Pacific Grove.

[11] Puri, M.L. & Sen, P.K. (1985). *Nonparametric Methods in General Linear Models*, Wiley, New York.

[12] Resampling Stats [Computer Software]. (1999). Resampling Stats, Arlington.

[13] Rodgers, J. (1999). The bootstrap, the jackknife, and the randomization test: a sampling taxonomy, *Multivariate Behavioral Research* **34**(4), 441–456.

[14] RT [Computer Software]. (1999). West, Cheyenne.

[15] SAS [Computer Software]. (1999). SAS Institute, Cary.

[16] StatXact [Computer Software]. (1999). Cytel Software, Cambridge.

[17] Testimate [Computer Software]. (1999). SciTech, Chicago.

[18] Thompson, G.L. (1991). A note on the rank transform for interactions, *Biometrika* **78**, 697–701.

[19] Thompson, G.L. (1993). A correction note on the rank transform for interactions, *Biometrika* **80**, 711.

[20] Toothaker, L.E. & Newman, D. (1994). Nonparametric competitors to the two-way ANOVA, *Journal of Educational and Behavioral Statistics* **19**, 237–273.

[21] Wilcox, R.R. (2003). *Applying Contemporary Statistical Techniques*, Academic Press, San Diego.

LARRY E. TOOTHAKER

Type I, Type II and Type III Sums of Squares

Scott L. Hershberger

Volume 4, pp. 2089–2092

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Type I, Type II and Type III Sums of Squares

Introduction

In structured experiments to compare a number of treatments, two factors are said to be *balanced* if each level of the first factor occurs as frequently at each level of the second. One example of this is a **randomized block design**, where each level of J treatments occurs only once in each of the K blocks; thus, every possible block-treatment combination occurs once. In balanced designs, blocks and treatments are *orthogonal* factors.

Orthogonal factors are linearly independent: When graphed in n-dimensional space, orthogonal factors lie at a right angle (90°) to one another. As in other linearly independent relationships, orthogonal factors are often, but not always, uncorrelated with one another [4]. Orthogonal factors become correlated if, upon centering the factors, the angle between the factors deviates from 90°.

In balanced designs, (1) least-squares estimates (*see* **Least Squares Estimation**) of treatment parameters (each of the factor's levels) are simply the contrast of the levels' means, and (2) the sum of squares for testing the main effect of A depends only on the means of the factor's levels and does not involve elimination of blocks [3]. Property (2) implies that each level of the blocking variable contributes equally to the estimation of the main effect. Whenever properties (1) and (2) are true, the factor and blocking variable are orthogonal, and as a result, the factor's main effect is estimated independently of the blocking variable. Similarly, in a two-way ANOVA design in which cells have equal numbers of observations and every level of factor A is crossed once with every level of factor B, the A and B main effects are independently estimated, neither one affecting the other (*see* **Factorial Designs**).

Properties (1) and (2) also hold in another type of orthogonal design, the **Latin square**. In an $n \times n$ Latin square, each of the n treatments occurs only once in each row and once in each column. Here, treatments are orthogonal to both rows and columns. Indeed, rows and columns are themselves orthogonal. Thus, when we say that a Latin square design is an orthogonal design, we mean that it is orthogonal for the estimation of row, column, and main effects.

In *nonorthogonal* designs, the estimation of the main effect of one factor is determined in part by the estimation of the main effects of other factors. Nonorthogonal factors occur when the combination of their levels is *unbalanced.* Generally speaking, unbalanced designs arise under one of two circumstances: either one or more factor level combinations are missing because of the complete absence of observations for one or more cells or factor level combinations vary in number of observations.

Whenever nonorthogonality is present, the effects in an experiment are **confounded** (yoked). Consider the two 3×3 designs below in which each cell's sample size is given.

Design I

			B		
		1	2		3
	1	5	5	5	
A	2	5	5		5
	3	5	5		0

Design II

		B			
		1	2		3
	1	5	5	5	
A	2	5	5		5
	3	5	5	2	

In both designs, factors A and B are confounded, and thus, nonorthogonal. In Design I, nonorthogonality is due to the zero frequency of cell a_3b_3: (*see* **Structural Zeros**) main effect B (i.e., comparing levels $B1$, $B2$, and $B3$) is evaluated by collapsing within each level of B rows $A1$, $A2$, and $A3$. However, while $A1$, $A2$, and $A3$ are available to be collapsed within $B1$ and $B2$, only $A1$ and $A2$ are available within $B3$. As a result, the main effect hypothesis for factor B cannot be constructed so that its marginal means are based on cell means that have all of the same levels of factor A. Consequently, the test of B's marginal means is confounded and dependent on factor A's marginal means. Similarly, the test of A's marginal means is confounded and dependent on factor B's marginal means. In Design II, factors A and B are confounded because of the smaller sample size of cell a_3b_3: the main effect hypothesis for factor B

(and A) cannot be constructed so that each of factor A's (and B's) levels are weighted equally in testing the B (and A) main effect.

Sum of Squares

In an orthogonal or balanced ANOVA, there is no need to worry about the decomposition of sums of squares. Here, one ANOVA factor is independent of another ANOVA factor, so a test for, say, a sex effect is independent of a test for, say, an age effect. When the design is unbalanced or nonorthogonal, there is not a unique decomposition of the sums of squares. Hence, decisions must be made to account for the dependence between the ANOVA factors in quantifying the effects of any single factor. The situation is mathematically equivalent to a multiple regression model where there are correlations among the predictor variables. Each variable has direct and indirect effects on the dependent variable. In an ANOVA, each factor will have direct and indirect effects on the dependent variable. Four different types of sums of squares are available for the estimation of factor effects. In an orthogonal design, all four will be equal. In a nonorthogonal design, the correct sums of squares will depend upon the logic of the design.

Type I SS

Type I SS are order-dependent (hierarchical). Each effect is adjusted for all other effects that appear earlier in the model, but not for any effects that appear later in the model. For example, if a three-way ANOVA model was specified to have the following order of effects,

$$A, B, A \times B, C, A \times C, B \times C, A \times B \times C,$$

the sums of squares would be calculated with the following adjustments:

Effect	Adjusted for
A	–
B	A
$A \times B$	A, B
C	$A, B, A \times B$
$A \times C$	$A, B, C, A \times B$
$B \times C$	$A, B, C, A \times B, A \times C$
$A \times B \times C$	$A, B, C, A \times B, A \times C, A \times B \times C$

Type I SS are computed as the decrease in the error SS (SSE) when the effect is added to a model. For example, if SSE for $Y = A$ is 15 and SSE for $Y = A \times B$ is 5, then the Type I SS for B is 10. The sum of all of the effects' SS will equal the total model SS for Type I SS – this is not generally true for the other types of SS (which exclude some or all of the variance that cannot be unambiguously allocated to one and only one effect). In fact, specifying effects hierarchically is the only method of determining the unique amount of variance in a dependent variable explained by an effect. Type I SS are appropriate for balanced (orthogonal, equal n) analyses of variance in which the effects are specified in proper order (main effects, then two-way interactions, then three-way interactions, etc.), for trend analysis where the powers for the quantitative factor are ordered from lowest to highest in the model statement, and the **analysis of covariance** (ANCOVA) in which covariates are specified first. Type I SS are also used for hierarchical step-down nonorthogonal analyses of variance [1] and hierarchical regression [2]. With such procedures, one obtains the particular SS needed (adjusted for some effects but not for others) by carefully ordering the effects. The order of effects is usually based on temporal priority, on the causal relations between effects (an outcome should not be added to the model before its cause), or on theoretical importance.

Type II SS

Type II SS are the reduction in the SSE as a result of adding the effect to a model that contains all other effects except effects that contain the effect being tested. An effect is contained in another effect if it can be derived by deleting terms in that effect – for example, A, B, C, $A \times B$, $A \times C$, and $B \times C$ are all contained in $A \times B \times C$. The Type II SS for our example involve the following adjustments:

Effect	Adjusted for
A	$B, C, B \times C$
B	$A, C, A \times C$
$A \times B$	$A, B, C, A \times C, B \times C$
C	$A, B, A \times B$
$A \times C$	$A, B, C, A \times B, B \times C$
$B \times C$	$A, B, C, A \times B, A \times C$
$A \times C$	$A, B, C, A \times B, A \times C, B \times C$

When the design is balanced, Type I and Type II *SS* are identical.

Type III SS

Type III SS are identical to those of Type II *SS* when the design is balanced. For example, the sum of squares for A is adjusted for the effects of B and for $A \times B$. When the design is unbalanced, these are the *SS* that are approximated by the traditional unweighted means ANOVA that uses **harmonic mean** sample sizes to adjust cell totals: Type III *SS* adjusts the sums of squares to estimate what they might be if the design were truly balanced. To illustrate the difference between Type II and Type III *SS*, consider factor A to be a dichotomous variable such as gender. If the data contained 60% females and 40% males, the Type II sums of squares are based on those percentages. In contrast, the Type III *SS* assume that the sex difference came about because of sampling and tries to generalize to a population in which the number of males and females is equal.

Type IV SS

Type IV SS differ from Types I, II, and III *SS* in that it was developed for designs that have one or more empty cells, that is, cells that contain no observations. (*see* **Structural Zeros**) Type IV *SS* evaluate marginal means that are based on equally weighted cell means. They yield the same results as a Type III *SS* if all cells in the design have at least one observation. As a result, with Type IV *SS*, marginal means of one factor are based on cell means that have all of the same levels of the other factor, avoiding the confounding of factors that would occur if cells were empty.

References

[1] Applebaum, M.I. & Cramer, E.M. (1974). Some problems in the nonorthogonal analysis of variance, *Psychological Bulletin* **81**, 335–343.

[2] Cohen, J. & Cohen, P. (1983). *Applied Multiple Regression/correlation Analysis for the Behavioral Sciences*, 2nd Edition. Erlbaum, Hillsdale.

[3] Milliken, G.A. & Johnson, D.E. (1984). *Analysis of Messy Data*, Vol. 1, Wadsworth, Belmont.

[4] Rodgers, J.L., Nicewander, W.A. & Toothaker, L. (1984). Linearly independent, orthogonal, and uncorrelated variables, *American Statistician* **38**, 133–134.

Scott L. Hershberger

Ultrametric Inequality

Fionn Murtagh

Volume 4, pp. 2093–2094

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Ultrametric Inequality

The triangular inequality holds for a metric space: $d(x, z) \leq d(x, y) + d(y, z)$ for any triplet of points x, y, z. In addition, the properties of symmetry $(d(x, y) = d(y, x))$ and positive definiteness ($d(x, y) \geq 0$ with $x = y$ if $d(x, y) = 0$) are respected.

The 'strong triangular inequality' or ultrametric inequality is $d(x, z) \leq \max\ \{d(x, y), d(y, z)\}$ for any triplet x, y, z.

An ultrametric space implies respect for a range of stringent properties. For example, the triangle formed by any triplet is necessarily isosceles, with the two large sides equal, or is equilateral.

Consider the dissimilarity data shown on the left side of Table 1. (For instance, this could be the similarity of performance between five test subjects on a scale of $1 =$ very similar, to $9 =$ very dissimilar.) The single link criterion was used to construct the dendrogram shown (Figure 1). On the right side of Table 1, the ultrametric distances defined from the sequence of agglomeration values are given.

Among the ultrametric distances, consider for example, $d(2, 3)$, $d(3, 4)$, and $d(4, 1)$. We see that $d(2, 3) \leq \max\ \{d(3, 4), d(4, 1)\}$ since here we have $5 \leq \max\ \{5, 4\}$. We can turn around any way we like what we take as x, y, z but with ultrametric distances, we will always find that $d(x, z) \leq \max\ \{d(x, y), d(y, z)\}$.

Table 1 Left: original pairwise dissimilarities. Right: derived ultrametric distances

	1	2	3	4	5	1	2	3	4	5
1	0	4	9	5	8	0	4	5	4	5
2	4	0	6	3	6	4	0	5	3	5
3	9	6	0	6	3	5	5	0	5	3
4	5	3	6	0	5	4	3	5	0	5
5	8	6	3	5	0	5	5	3	5	0

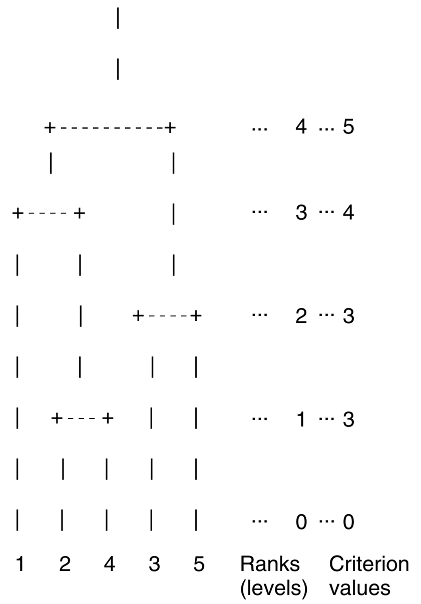


Figure 1 Resulting dendrogram

In data analysis, the ultrametric inequality is important because a **hierarchical clustering** is tantamount to defining ultrametric distances on the objects under investigation. More formally, we say that in clustering, a bijection is defined between a rooted, binary, ranked, indexed tree, called a *dendrogram* (*see* **Hierarchical Clustering**), and a set of ultrametric distances ([1], representing work going back to the early 1960s; [2]).

References

[1] Benzécri, J.P. (1979). *La Taxinomie*, 2nd Edition, Dunod, Paris.

[2] Johnson, S.C. (1967). Hierarchical clustering schemes, *Psychometrika* **32**, 241–254.

Fionn Murtagh

Ultrametric Trees

Fionn Murtagh

Volume 4, pp. 2094–2095

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Ultrametric Trees

Let us call the row of a data table an observation vector. In the following, we will consider the case of n rows. The set of rows will be denoted I.

Hierarchical agglomeration on n observation vectors, $i \in I$, involves a series of pairwise agglomerations of observations or clusters (*see* **Hierarchical Clustering**). In Figure 1 we see that observation vectors x_1 and x_2 are first agglomerated, followed by the incorporation of x_3, and so on. Since an agglomeration is always pairwise, we see that there are precisely $n - 1$ agglomerations for n observation vectors.

We will briefly look at three properties of a hierarchical clustering tree (referred to as conditions (i), (ii) and (iii)). A hierarchy, H, is a set of sets, q. Each q is a subset of I. By convention, we include I itself as a potential subset, but we do not include the empty set. Our input data (observation vectors), denoted $x_1, x_2, \ldots$ are also subsets of I. All subsets of I are collectively termed the power set of I. The notation sometimes used for the power set of I is 2^I. Condition (i) is that the hierarchy includes the set I, and this corresponds to the top (uppermost) level in Figure 1. Condition (ii) requires that $x_1, x_2, \ldots$ are in H, and this corresponds to the bottom (lowermost) level in Figure 1. Finally, condition (iii) is that different subsets do not intersect unless one is a member of the other that is, they do not overlap other than through 100% inclusion of one in the other.

In Figure 1, the observation set $I = \{x_1, x_2, \ldots, x_8\}$. The $x_1, x_2, \ldots$ are called *singleton sets* (as opposed to clusters), and are associated with terminal nodes in the dendrogram tree.

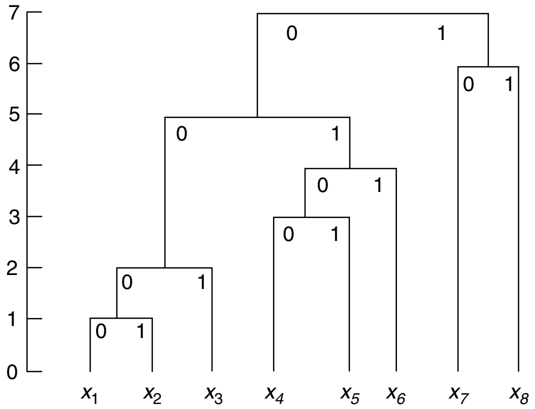


Figure 1 Labeled, ranked dendrogram on 8 terminal nodes. Branches labeled 0 and 1

A dendrogram, displayed in Figure 1, is the name given to a binary, ranked tree, which is a convenient representation for the set of agglomerations (*see* **Hierarchical Clustering**).

Now we will show the link with the **ultrametric inequality**. An indexed hierarchy is the pair (H, ν) where the positive function defined on H, that is, $\nu : H \to \mathbb{R}^+$, ($\mathbb{R}^+$ denotes the set of positive real numbers) satisfies $\nu(i) = 0$ if $i \in H$ is a singleton; and $q \subset q' \Longrightarrow \nu(q) < \nu(q')$. Function ν is the agglomeration level. Typically the index, or agglomeration level, is defined from the succession of agglomerative values that are used in the creation of the dendrogram. In practice, this means that the hierarchic clustering algorithm used to produce the dendrogram representation yields the values of ν.

The distance between any two clusters is based on the 'common ancestor', that is, how high in the dendrogram we have to go to find a cluster that contains both. Take $q \subset q'$, let $q \subset q''$ and $q' \subset q''$, and let q'' be the lowest level cluster for which this is true. Then if we define $D(q, q') = \nu(q'')$, it can be shown that D is an ultrametric.

In practice, we start with our given data, and a Euclidean or other dissimilarity, we use some compactness agglomeration criterion such as minimizing the change in variance resulting from the agglomerations, and then define $\nu(q)$ as the dissimilarity associated with the agglomeration carried out.

The standard agglomerative hierarchical clustering algorithm developed in the early 1980s is based on the construction of nearest neighbor chains, followed by agglomeration of reciprocal nearest neighbors. For a survey, see Murtagh [1, 2].

References

[1] Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms, *The Computer Journal* **26**, 354–359.

[2] Murtagh, F. (1985). *Multidimensional Clustering Algorithms*, COMPSTAT Lectures Volume 4, Physica-Verlag, Vienna.

Further Reading

Benzécri, J.P. (1976). *L'Analyse des Données. Tome 1. La Taxinomie*, 2nd Edition, Dunod, Paris.

Fionn Murtagh

Unidimensional Scaling

Jan de Leeuw

Volume 4, pp. 2095–2097

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Unidimensional Scaling

Unidimensional scaling is the special one-dimensional case of **multidimensional scaling**. It is often discussed separately, because the unidimensional case is quite different from the general multidimensional case. It is applied in situations in which we have a strong reason to believe there is only one interesting underlying dimension, such as time, ability, or preference. In the unidimensional case, we do not have to choose between different metrics, such as the Euclidean metric, the City Block metric, or the Dominance metric (*see* **Proximity Measures**). Unidimensional scaling techniques are very different from multidimensional scaling techniques, because they use very different algorithms to minimize their loss functions.

The classical form of unidimensional scaling starts with a symmetric and nonnegative matrix $\Delta = \{\delta_{ij}\}$ of *dissimilarities* and another symmetric and nonnegative matrix $W = \{w_{ij}\}$ of *weights*. Both W and Δ have a zero diagonal. Unidimensional scaling finds *coordinates* x_i for n points on the line such that

$$\sigma(x) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(\delta_{ij} - |x_i - x_j|)^2 \qquad (1)$$

is minimized. Those n coordinates in x define the scale we were looking for.

To analyze this unidimensional scaling problem in more detail, let us start with the situation in which we know the order of the x_i, and we are just looking for their scale values. Now $|x_i - x_j| = s_{ij}(x)(x_i - x_j)$, where $s_{ij}(x) = \mathrm{sign}(x_i - x_j)$. If the order of the x_i is known, then the $s_{ij}(x)$ are known numbers, equal to either -1 or $+1$ or 0, and thus our problem becomes minimization of

$$\sigma(x) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(\delta_{ij} - s_{ij}(x_i - x_j))^2 \qquad (2)$$

over all x such that $s_{ij}(x_i - x_j) \geq 0$. Assume, without loss of generality, that the weighted sum of squares of the dissimilarities is one. By expanding the sum of squares we see that

$$\sigma(x) = 1 - t'Vt + (x - t)'V(x - t). \qquad (3)$$

Here V is the matrix with off-diagonal elements $v_{ij} = -w_{ij}$ and diagonal elements $v_{ii} = \sum_{j=1}^{n} w_{ij}$. Also, $t = V^{+}r$, where r is the vector with elements $r_i = \sum_{j=1}^{n} w_{ij}\delta_{ij}s_{ij}$, and V^{+} is a generalized inverse of V. If all the off-diagonal weights are equal, we simply have $t = r/n$.

Thus, the unidimensional scaling problem, with a known scale order, requires us to minimize $(x - t)'V(x - t)$ over all x, satisfying the order restrictions. This is a **monotone regression** problem, possibly with a nondiagonal weight matrix, which can be solved quickly and uniquely by simple quadratic programming methods.

Now for some geometry. The vectors x satisfying the same set of order constraints form a polyhedral convex cone $\mathcal{K}$ in $\mathbb{R}^n$. Think of $\mathcal{K}$ as an ice-cream cone with its apex at the origin, except for the fact that the ice-cream cone is not round, but instead bounded by a finite number of hyperplanes. Since there are $n!$ different possible orderings of x, there are $n!$ cones, all with their apex at the origin. The interior of the cone consists of the vectors without ties, intersections of different cones are vectors with at least one tie. Obviously, the union of the $n!$ cones is all of $\mathbb{R}^n$.

Thus, the unidimensional scaling problem can be solved by solving $n!$ monotone regression problems, one for each of the $n!$ cones [2]. The problem has a solution, which is at the same time very simple and prohibitively complicated. The simplicity comes from the $n!$ subproblems, which are easy to solve, and the complications come from the fact that there are simply too many different subproblems. Enumeration of all possible orders is impractical for $n > 10$, although using combinatorial programming techniques makes it possible to find solutions for n as large as 30 [6].

Actually, the subproblems are even simpler than we suggested above. The geometry tells us that we solve the subproblem for cone $\mathcal{K}$ by finding the closest vector to t in the cone, or, in other words, by projecting t on the cone. There are three possibilities. Either t is in the interior of its cone, or on the boundary of its cone, or outside its cone. In the first two cases, t is equal to its projection, in the third case, the projection is on the boundary. The general result in [1] tells us that the loss function σ cannot have a local minimum at a point in which there are ties, and, thus, local minima can only occur in the interior of the cones. This means that we can only have a local minimum if t is in the interior of its cone [4], and it also means that we actually never have to compute the monotone regression. We just have to verify if t

is in the interior, if it is not, then σ does not have a local minimum in this cone.

There have been many proposals to solve the combinatorial optimization problem of moving through the $n!$ cones until the global optimum of σ has been found. A recent review is [7].

We illustrate the method with a simple example, using the vegetable paired-comparison data from [5, p. 160]. Paired comparison data are usually given in a matrix P of proportions, indicating how many times stimulus i is preferred over stimulus j. P has 0.5 on the diagonal, while corresponding elements p_{ij} and p_{ji} on both sides of the diagonal add up to 1.0. We transform the proportions to dissimilarities by using the probit transformation $z_{ij} = \Phi^{-1}(p_{ij})$ and then defining $\delta_{ij} = |z_{ij}|$. There are nine vegetables in the experiment, and we evaluate all $9! = 362\,880$ permutations. Of these cones, 14 354 or 4% have a local minimum in their interior. This may be a small percentage, but the fact that σ has 14 354 isolated local minima indicates how complicated the unidimensional scaling problem is. The global minimum is obtained for the order given in Guilford's book, which is Turnips < Cabbage < Beets < Asparagus < Carrots < Spinach < String Beans < Peas < Corn. Since there are no weights in this example, the optimal unidimensional scaling values are the row averages of the matrix with elements $s_{ij}(x)\delta_{ij}$. Except for a single sign change of the smallest element (the Carrots and Spinach comparison), this matrix is identical to the probit matrix Z. And because the Thurstone Case V scale values are the row averages of Z, they are virtually identical to the unidimensional scaling solution in this case.

The second example is quite different. It has weights and incomplete information. We take it from an early paper by Fisher [3], in which he studies crossover percentages of eight genes on the sex chromosome of *Drosophila willistoni*. He takes the crossover percentage as a measure of distance, and supposes the number n_{ij} of crossovers in N_{ij} observations is binomial. Although there are eight genes, and thus $\binom{8}{2} = 28$ possible dissimilarities, there are only 15 pairs that are actually observed. Thus, 13 of the off-diagonal weights are zero, and the other weights are set to the inverses of the standard errors of the proportions. We investigate all $8! = 40\,320$ permutations, and we find 78 local minima. The solution given by **Fisher**, computed by solving linearized likelihood equations, has Reduced < Scute < Peach < Beaded < Rough < Triple < Deformed < Rimmed. This order corresponds with a local minimum of σ equal to 40.16. The global minimum is obtained for the permutation that interchanges Reduced and Scute, with value 35.88. In Figure 1, we see the scales for the two local minima, one corresponding with Fisher's order and the other one with the optimal order.

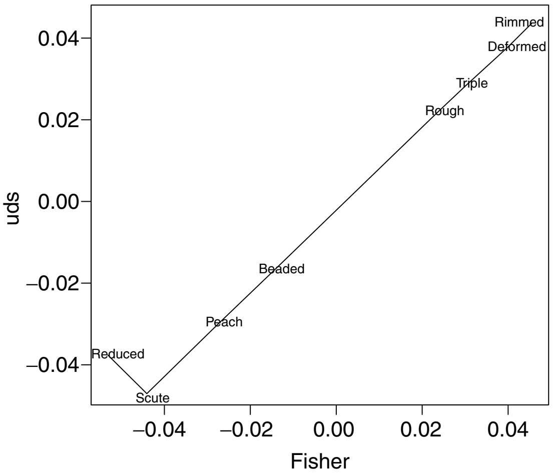


Figure 1 Genes on Chromosome

In this entry, we have only discussed least squares metric unidimensional scaling. The first obvious generalization is to replace the least squares loss function, for example, by the least absolute value or $\mathcal{L}_1$ loss function. The second generalization is to look at nonmetric unidimensional scaling. These generalizations have not been studied in much detail, but in both, we can continue to use the basic geometry we have discussed. The combinatorial nature of the problem remains intact.

References

[1] De Leeuw, J. (1984). Differentiability of Kruskal's stress at a local minimum, *Psychometrika* **49**, 111–113.

[2] De Leeuw, J. & Heiser, W.J. (1977). Convergence of correction matrix algorithms for multidimensional scaling, in *Geometric Representations of Relational Data*, J.C., Lingoes, ed., Mathesis Press, Ann Arbor, pp. 735–752.

[3] Fisher, R.A. (1922). The systematic location of genes by means of crossover observations, *American Naturalist* **56**, 406–411.

[4] Groenen, P.J.F. (1993). The majorization approach to multidimensional scaling: some problems and extensions, Ph.D. thesis, University of Leiden.

[5] Guilford, J.P. (1954). *Psychometric Methods*, 2nd Edition, McGraw-Hill.

[6] Hubert, L.J. & Arabie, P. (1986). Unidimensional scaling and combinatorial optimization, in *Multidimensional Data Analysis*, J. De Leeuw, W. Heiser, J. Meulman & F. Critchley, eds, DSWO-Press, Leiden.

[7] Hubert, L.J., Arabie, P. & Meulman, J.J. (2002a). Linear unidimensional scaling in the L_2-Norm: basic optimization methods using MATLAB, *Journal of Classification* **19**, 303–328.

JAN DE LEEUW

Urban, F M

Helen Ross

Volume 4, pp. 2097–2098

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Urban, F M

Born: December 28, 1878, Brünn, Moravia (now Austria).
Died: May 4, 1964, Paris, France.

F. M. Urban's main claim to fame was the probability weightings that he added to the calculation of the psychometric function. In 1968, he was ranked the 14th most famous psychologist, but is now almost forgotten – so much so that a professor who was equipping a new psychological laboratory reputedly attempted to order a set of Urban weights. There are few biographical sources about him, and even his usual first name is in doubt (see [1]).

Friedrich Johann Victor Urban was born into a German-speaking family in Austro-Hungary. He is known by the initials F. M. because he adopted Friedrich Maria as a pen name, which became Frederick Mary, or perhaps Francis M., in the United States. He studied philosophy at the University of Vienna from 1897, obtaining a Ph.D. in 1902 in aesthetics. He also studied probability under Wilhelm Wirth at the University of Leipzig. He taught psychological acoustics at Harvard University in 1904–1905. His main work on statistics was conducted from 1905 at the University of Pennsylvania, where he rose to assistant professor in 1910. He published in German and English in 1908–1909 [3, 4]. He was elected a member of the American Psychological Association in 1906, the American Association for the Advancement of Science in 1907, and a Fellow of the latter in 1911. Urban's life in America ended in 1914, when he returned home and married Adele Königsgarten. Lacking US citizenship, he was not allowed to return to the United States on the outbreak of war. He moved to Sweden, and spent 1914 to 1917 in Götheborg and Stockholm, where he did research at the Kungliga Vetenskapliga Akademien. He returned to Brünn in 1917, and served in the Austrian army. Moravia became part of the Czechoslovak Republic in 1918. Urban then worked as an insurance statistician, his Czech language ability being inadequate for a university post. He corresponded with his colleagues abroad, and continued to publish in German and English on psychometry and psychophysics. He also translated into German J. M. Keynes's *A Treatise on Probability* (1926) and J. L. Coolidge's *An Introduction to Mathematical Probability* (1927). He and his Jewish wife stayed in Brünn throughout the Second World War, suffering Hitler's invasion, the bombing of their house, the Russian 'liberation', and the expulsion of German speakers by the returning Czech regime. He was put in a concentration camp first by the Russians and then by the Czechs. He was released owing to pleas from abroad, but was forced to leave Czechoslovakia in 1948. He and his wife went to join their elder daughter first in Norway, and then in Brazil (1949–52), where he lectured on **factor analysis** at São Paulo University. In 1952, they moved in with their younger daughter in France, living in Toulon and, finally, Paris.

Urban's many contributions to statistics are discussed in [2]. The Müller–Urban weights are a refinement to the least squares solution for the psychometric function, when P values are transformed to z values. Müller argued that the proportions near 0.5 should be weighted most, while Urban argued that those with the smallest mean-square error should be weighted most. The combined weightings were laborious to calculate, and made little difference to the final threshold values. The advent of computers and new statistical procedures has relegated these weights to history.

References

[1] Ertle, J.E., Bushong, R.C. & Hillix, W.A. (1977). The real F.M.Urban, *Journal of the History of the Behavioral Sciences* **13**, 379–383.

[2] Guilford, J.P. (1954). *Psychometric Methods*, 2nd Edition, McGraw-Hill, New York.

[3] Urban, F.M. (1908). *The Application of Statistical Methods to Problems of Psychophysics*, Psychological Clinic Press, Philadelphia.

[4] Urban, F.M. (1909). Die psychophysischen Massmethoden als Grundlagen empirischer Messungen, *Archiv für die Gesamte Psychologie* **15**, 261–355; **16**, 168–227.

HELEN ROSS

Utility Theory

Daniel Read and Mara Airoldi

Volume 4, pp. 2098–2101

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Utility Theory

The term utility is commonly used in two ways. First, following the usage of the utilitarian philosophers, it refers to the total amount of pleasure (or pain) that an action or a good can bring to those affected by it. According to this view, utility is a function of experience. If an apple has more utility than an orange, this means that it brings more pleasure to the person eating it. A second view of utility is that it is a way of representing behaviors, such as choices or expressed preferences, without any reference to the experience arising from what is chosen. For instance, if someone is offered either an apple or an orange, and they choose an apple, we would describe this preference by assigning a higher utility number to the apple than to the orange. This number would say nothing about whether that person would gain more or less pleasure from the apple, but only that they chose the apple. The first use of the term is most closely associated with utilitarian philosophers, pre-twentieth-century economists, and psychologists. The second use is associated with twentieth-century economists.

Daniel Bernoulli is the father of utility theory. He developed the theory to solve the St Petersburg paradox. The paradox revolves around the value placed on the following wager: A fair coin is to be tossed until it comes up heads, at which point those who have paid to play the game will receive 2^n ducats, where n is the number of the toss when heads came up. For instance, if the first head comes up on the third toss the player will receive 8 ducats. Prior to Bernoulli, mathematicians held the view that a gamble was worth its *expected monetary value*, meaning the sum of all possible outcomes multiplied by their probability (*see* **Expectation**). Yet although the expected monetary value of the St Petersburg wager is infinite:

$$\sum_{n=1}^{\infty} \frac{1}{2^n} 2^n = 1 + 1 + 1 + \cdots = \infty, \qquad (1)$$

nobody would be willing to pay more than a few ducats for it.

Bernoulli argued that people value this wager according to its expected *utility* rather than its expected *monetary value*, and proposed a specific relationship between utility and wealth: the 'utility resulting from any small increase in wealth will be inversely proportional to the quantity of goods previously possessed' [3]. This implies that the utility of wealth increases logarithmically with additional increments: $\mathrm{u(w)} = \log(\mathrm{w})$. With this *utility function*, the St Petersburg paradox was resolved because:

$$\sum_{n=1}^{\infty} \frac{1}{2^n} \log(2^n) = 2\log 2 < \infty. \qquad (2)$$

As Menger showed [8], this solution does not hold for all versions of the St Petersburg gamble. For instance, if the payoff is e^{2^n}, the expected utility of the wager is still infinite even with a logarithmic utility function. The paradox can be completely solved only with a bounded utility function. Nonetheless, through this analysis Bernoulli introduced the three major themes of utility theory: first, the same outcome has different utility for different people; second, the relationship between wealth and utility can be described mathematically; third, utility is marginally diminishing, so that the increase in utility from each additional ducat is less than that from the one before.

It is clear that Bernoulli viewed utility as an index of the *benefit* or *good* that a person would get from their income. This view was central to the utilitarians, such as Godwin and Bentham, who considered utility to be a quantity reflecting the disposition to bring pleasure or pain. These philosophers, who wrote in the eighteenth and nineteenth centuries, maintained that the goal of social and personal action is to maximize the sum of the utility of all members of society. Many utilitarian philosophers (and the economists who followed them, such as Alfred Marshall) pointed out that if utility were a logarithmic function of wealth, then transferring wealth from the rich to the poor would maximize the total utility of society. This argument was based on two assumptions that have been difficult to maintain: utility is measurable, and interpersonal comparisons of utility are possible [1, 4].

At the turn of the twentieth century, economists realized that these assumptions were not needed for economic analysis, and that they could get by with a strictly *ordinal* utility function [5, 9]. The idea is that through their choices consumers show which of the many possible allocations of their income they prefer (*see* **Demand Characteristics**). In general, for any pair of allocations, a consumer will either be indifferent between them, or prefer one to the other.

By obtaining binary choices between allocations, it is possible to draw *indifference curves*. To predict such choices, we just need to know which allocations are on the highest indifference curve. This change in thinking was not only a change in mathematics from interval to ordinal measurement (*see* **Scales of Measurement**; **Measurement: Overview**) but also a conceptual revolution. Utility, as used by economists, lost its connection with 'pleasure and pain' or other measures of benefit.

One limitation of ordinal utility was that it dealt only with choices under certainty – problems such as the St Petersburg paradox could not be discussed using ordinal utility language. In the mid-twentieth century, Von Neumann and Morgenstern [11] reintroduced the concept of choice under uncertainty with expected utility theory. They showed that just as indifference maps can be drawn from consistent choices between outcomes, so cardinal utility functions (i.e., those measurable on an interval scale) can be derived from consistent choices between gambles or lotteries. Von Neumann and Morgenstern, however, did not reinstate the link between utility and psychology. They did not view utility as a *cause*, but a *reflection*, of behavior: we do not choose an apple over an orange because it has more utility, but it has more utility because we chose it. They showed that if choices between lotteries meet certain formal requirements then preferences can be described by a cardinal utility function. The following assumptions summarize these requirements:

Assumption 1 *Ordering of alternatives*. For each pair of prizes or lotteries an agent will either be indifferent ($\sim$) between them or prefer one to the other ($\succ$). That is, either A $\succ$ B, B $\succ$ A or A$\sim$B.

Assumption 2 *Transitivity*. If A $\succ$ B and B $\succ$ C, then A $\succ$ C.

Assumption 3 *Continuity*. If an agent prefers A to B to C, then there is some probability p which makes a lottery offering A with probability p, and C with probability $(1 - p)$ equal in utility to B. That is B $\sim (p, \mathrm{A}; 1 - p, \mathrm{C})$.

Assumption 4 *Substitutability*. Prizes in lotteries can be replaced with lotteries having the same value. Using the outcomes from Assumption 3, the lottery $(q, \mathrm{X}; 1 - q, \mathrm{B})$ is equivalent to $(q, \mathrm{X}; 1 - q, (p, \mathrm{A}; 1 - p, \mathrm{C}))$.

Assumption 5 *Independence of irrelevant alternatives*. If two lotteries offer the same outcome with identical probabilities, then the preferences between the lotteries will depend only on the other (unshared) outcomes. If A $\succ$ B, then $(p, \mathrm{A}; 1 - p, \mathrm{C}) \succ (p, \mathrm{B}; 1 - p, \mathrm{C})$.

Assumption 6 *Reduction of compound lotteries*. A compound lottery is a lottery over lotteries. The reduction of compound lotteries condition means that such a lottery is equivalent to one that has been reduced using the standard rules of probability. For instance, consider the two lotteries, Z and Z′, in Figure 1: Assumption 6 states that Z $\sim$ Z′.

If these assumptions hold then a cardinal utility function, unique up to a linear transformation, can be derived from preferences over lotteries. Researchers view these assumptions as 'axioms of rationality' because they seem, intuitively, to be how a rational person would behave [10].

One way of doing so is via the *certainty equivalent method* of calculating utilities [12]. In this method, the decision maker first ranks all relevant prizes or outcomes from best to worst. An arbitrary utility value, usually 0 and 1, is given to the worst (W) and best (B) outcome. Then for each intermediate outcome X the decision maker specifies the probability p that would make them indifferent between X for

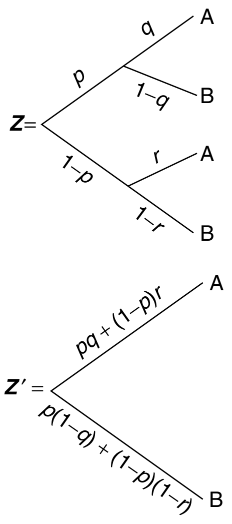


Figure 1 Assumption 6 implies indifference between Z and Z′ (Z $\sim$ Z′)

sure and the lottery $(p, \mathrm{B}; 1-p, W)$. The utility of X is then found by replacing $u(\mathrm{W})=0$ and $u(\mathrm{B})=1$ in the calculation:

$$u(\mathrm{X})=pu(\mathrm{B})+(1-p)\,u(\mathrm{W})=p. \qquad (3)$$

If different values had been assigned to $u(\mathrm{W})$ and $u(\mathrm{B})$, the resultant utility scale would be a linear transformation of this one.

A further development of utility theory is subjective expected utility theory, which incorporates subjective probabilities. Ramsey, **de Finetti** and **Savage** [10] showed that probabilities as well as utilities could be axiomatized, with choices that reflect probabilities being used to derive a subjective utility function.

Current Thinking About Utility

Neither expected utility nor subjective expected utility theory has proved to be a good *descriptive* theory of choice. An early challenge came from Maurice Allais [2] who showed that independence of irrelevant alternatives could be routinely violated. The challenge, known as 'Allais Paradox', shows that the *variance* of the probability distribution of a gamble affects preferences. In particular there is strong empirical evidence that subjects attach additional psychological value or utility to an outcome if it has *zero* variance (*certainty effect* [6]). Allais rejected the idea that behaviors inconsistent with the standard axioms are irrational and claimed that 'rationality can be defined experimentally by observing the actions of people who can be regarded as acting in a rational manner' [2]. One consequence of this has been a number of *non-expected utility* theories that either relax or drop some of the assumptions, or else substitute them with empirically derived generalizations. The most influential of these is *prospect theory* [6], which is based on observations of how decision makers actually behave. It differs from expected utility theory in that the probability function is replaced by a nonlinear *weighting function*, and the utility function with a *value function*. The weighting function puts too much weight on low probabilities, too little on moderate to high probabilities, and has a discontinuity for changes from certainty to uncertainty. The value function is defined over deviations from current wealth. It is concave for increasing gains, and convex for losses. This leads to the *reflection effect*, which means that if a preference pattern is observed for gains, the opposite pattern will be found for losses. For example, people are generally risk averse for gains ($10 for sure is preferred to a 50/50 chance of $20 or nothing), but risk seeking for losses (a 50/50 chance of losing $20 is preferred to a sure loss of $10).

One issue that is currently the focus of much attention is whether it is possible to find a measure of utility that reflects happiness or pleasure, as originally envisaged by the utilitarians. People often choose options that appear objectively 'bad' for them, such as smoking or procrastinating, yet if utility is derived from choice behavior it means the utility of these bad options is greater than that of options that seem objectively better. If we are interested in questions of welfare, there is a practical need for a measure of utility that would permit us to say that 'X is better than Y, because it yields more utility'. One suggestion comes from Daniel Kahneman, who argues for a distinction between 'experienced utility' (the pain or pleasure from an outcome, the definition adopted by utilitarians), and 'decision utility' (utility as reflected in decisions, the definition adopted by modern economists) [7]. Through such refinements in the definition of utility, researchers like Kahneman aspire to reintroduce Bernoulli and Bentham's perspective to the scientific understanding of utility.

Acknowledgment

The writing of this article was assisted by funding from the Leverhulme Foundation Grant F/07004/Y.

References

[1] Alchian, A. (1953). The meaning of utility measurement, *The American Economic Review* **43**, 26–50.

[2] Allais, M. (1953). Le comportement de l'homme rationnel devant le risqué, critique des postulats et axioms de l'école Américaine, *Econometrica* **21**, 503–546; English edition: Allais, M. & Hagen O. eds. (1979). *Expected Utility Hypothesis and the Allais' Paradox*, Reidel Publishing, Dordrecht, 27–145.

[3] Bernoulli, D. (1738). *Specimen Theoriae Novae de Mensura Sortis*, in *Commentarii of Sciences in Petersburg, Vol. V*, 175–192; English edition (1954). Exposition of a new theory on the measurement of risk, *Econometrica* **22**(1), 23–36.

[4] Ellsberg, D. (1954). Classic and current notions of "measurable utility", *The Economic Journal* **64**, 528–556.

[5] Fisher, I. (1892). Mathematical investigations in the theory of value and prices, *Transaction of the Connecticut Academy of Art and Science* **IX**(1), 1–124.

[6] Kahneman, D. & Tversky, A. (1979). Prospect theory: an analysis of decision under risk, *Econometrica* **47**, 263–291.

[7] Kahneman, D., Wakker, P. & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility, *The Quarterly Journal of Economics* **112**, 375–406.

[8] Menger, K. (1934). Das unsicherheitsmoment in der wertlehre. Betrachtungen im anschluß an das sogenannte Petersburger Spiel, *Zeitschrift Für Nationalökonomie* **5**, 459–485.

[9] Pareto, V. (1909). *Manuale di Economia Politica: con una Introduzione Alla Scienza Sociale*, Società editrice libraria, Milan; English edition (1972). *Manual of Political Economy*, Macmillan, London.

[10] Savage, L.J. (1954). *The Foundations of Statistics*, Wiley, New York.

[11] von Neumann, J. & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*, 2nd Edition, Princeton University Press, Princeton.

[12] von Winterfeldt, D. & Edwards, W. (1986). *Decision Analysis and Behavioral Research*, Cambridge University Press, Cambridge.

DANIEL READ AND MARA AIROLDI

Validity Theory and Applications

Stephen G. Sireci

Volume 4, pp. 2103–2107

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Validity Theory and Applications

Measurement in the behavioral sciences typically refers to measuring characteristics of people such as attitudes, knowledge, ability, or psychological functioning. Such characteristics, called *constructs*, are hard to measure because unlike objects measured in the physical sciences (such as weight and distance), they are not directly observable. Therefore, the validity of measures taken in the behavioral sciences is always suspect and must be defended from both theoretical and empirical perspectives.

In this entry, I discuss validity theory and describe current methods for validating assessments used in the behavioral sciences. I begin with a description of a construct, followed by descriptions of construct validity and other validity nomenclature. Subsequently, the practice of test validation is described within the context of Kane's [8] argument-based approach to validity and the current *Standards for Educational and Psychological Testing* (American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education, [1]).

Constructs, Validity, and Construct Validity

The term *construct* has an important meaning in testing and measurement because it emphasizes the fact that we are not measuring tangible attributes (*see* **Latent Variable**). Assessments attempt to measure unobservable attributes such as attitudes, beliefs, feelings, knowledge, skills, and abilities. Given this endeavor, it must be assumed that (a) such attributes exist within people and (b) they are measurable. Since we do not know for sure if such intangible attributes or proficiencies really exist, we admit they are 'constructs'; they are hypothesized attributes we believe exist within people. Hence, we 'construct' these attributes from educational and psychological theories. To measure these constructs, we typically define them operationally through the use of test specifications and other elements of the assessment process [6].

Cronbach and Meehl [5] formally defined 'construct' as 'some postulated attribute of people, assumed to be reflected in test performance' (p. 283). The current version of the *Standards for Educational and Psychological Testing* [1] defines a construct as 'the concept or characteristic that a test is designed to measure' (p. 173). Given that a construct is invoked whenever behavioral measurement exists [9, 10], the *validity* of behavioral measures are often defined within the framework of *construct validity*. In fact, many validity theorists describe *construct validity* as equivalent to validity in general. The *Standards* borrow from Messick [10] and other validity theorists to underscore the notion that validity refers to inferences about constructs that are made on the basis of test scores.

According to the *Standards*, construct validity is:

> A term used to indicate that the test scores are to be interpreted as indicating the test taker's standing on the psychological construct measured by the test. A construct is a theoretical variable inferred from multiple types of evidence, which might include the interrelations of the test scores with other variables, internal test structure, observations of response processes, as well as the content of the test. In the current standards, all test scores are viewed as measures of some construct, so the phrase is redundant with validity. The validity argument establishes the construct validity of a test. (AERA et al. [1], p. 174)

The notion that construct validity is validity in general asserts that all other validity modifiers, such as content or criterion-related validity, are merely different ways of 'cutting validity evidence' ([10], p. 16). As Messick put it, 'Construct validity is based on an integration of any evidence that bears on the interpretation or meaning of the test scores' (p. 17).

Before the most recent version of the *Standards*, validity was most often discussed using three categories – content validity, criterion-related validity, and construct validity [11]. Although many theorists today believe terms such as content and criterion-related validity are misnomers, an understanding of these traditional ways of describing the complex nature of validity is important for understanding contemporary validity theory and how to validate inferences derived from test scores. We turn to a brief description of these traditional categories. Following this description, we summarize the key aspects of contemporary validity theory, and then describe current test validation practices.

Traditional Validity Categories: Content, Criterion-related, and Construct Validity

Prior to the 1980s, most discussions of validity described it as comprising the three aforementioned types or aspects – content, criterion-related, and construct. *Content validity* refers to the degree to which an assessment represents the content domain it is designed to measure. There are four components of content validity – domain definition, domain relevance, domain representation, and appropriate test construction procedures [11]. The domain definition aspect of content validity refers to how well the description of the test content, particularly the test specifications, is regarded as adequately describing what is measured and the degree to which it is consistent with other theories and pragmatic descriptions of the targeted construct. Domain relevance refers to the relevance of the items (tasks) on a test to the domain being measured. Domain representation refers to the degree to which a test adequately represents the domain being measured. Finally, appropriate test construction procedures refer to all processes used when constructing a test to ensure that test content faithfully and fully represents the construct intended to be measured and does not measure irrelevant material.

The term *content domain* is often used in describing content validity, but how this domain differs from the construct is often unclear, which is one reason why many theorists believe content validity is merely an aspect of construct validity. Another argument against the use of the term content validity is that it refers to properties of a test (e.g., how well do these test items represent a hypothetical domain?) rather than to the inferences derived from test scores. Regardless of the legitimacy of the term content validity, it is important to bear in mind that the constructs measured in the behavioral sciences must be defined clearly and unambiguously and evaluations of how well the test represents the construct must be made. Validating test content is necessary to establish an upper limit on the validity of inferences derived from test scores because if the content of a test cannot be defended as relevant to the construct measured, then the utility of the test scores cannot be trusted.

The second traditional validity category is *criterion-related validity*, which refers to the degree to which test scores correlate with external criteria that are considered to be manifestations of the construct of interest. There are two subtypes of criterion validity – *predictive validity* and *concurrent validity*. When the external criterion data are gathered long after a test is administered, such as when subsequent college grades are correlated with earlier college admissions test scores, the validity information is of the predictive variety. When the external criterion data are gathered about the same time as the test data, such as when examinees take two different test forms, the criterion-related validity is of the concurrent variety. The notion of criterion-related validity has received much attention in the validity literature, including the importance of looking at correlations between test scores and external criteria *irrelevant* to the construct measured (i.e., discriminant validity) as well as external criteria commensurate with the construct measured (i.e., convergent validity, [2]).

The last category of validity is construct validity, and as mentioned earlier, it is considered to be the most comprehensive form of validity. Traditionally, construct validity referred to the degree to which test scores are indicative of a person's standing on a construct. After introducing the term, Cronbach and Meehl [5] stated 'Construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured' (p 282). Given that a finite universe of content rarely exists and valid external criteria are extremely elusive, it is easy to infer that construct validity is involved whenever validity is investigated. Therefore, contemporary formulations of validity tend to describe content and criterion-related validity not as types of validity, but as ways of accumulating construct validity *evidence*. Using this perspective, evidence that the content of a test is congruent with the test specifications and evidence that test scores correlate with relevant criteria are taken as evidence that the construct is being measured. But construct validity is more than content and criterion-related validity. As described in the excerpt from the *Standards* given earlier, studies that evaluate the internal structure of a test (e.g., **factor analysis** studies), **differential item functioning** (item bias), cultural group differences (test bias), and unintended and intended consequences of a testing program all provide information regarding construct validity.

Fundamental Characteristics of Validity

The preceding section gave an historical perspective on validity and stressed the importance of construct validity. We now summarize fundamental characteristics of contemporary validity theory, borrowing from Sireci [13]. First, validity is *not* an intrinsic property of a test. A test may be valid for one purpose, but not for another, and so what we seek to validate in judging the worth of a test is the inferences derived from the test scores, not the test itself. Therefore, an evaluation of test validity starts with an identification of the specific purposes for which test scores are being used. When considering inferences derived from test scores, the validator must ask 'For what purposes are tests being used?' and 'How are the scores being interpreted?'

Another important characteristic of contemporary validity theory is that evaluating inferences derived from test scores involves several different types of qualitative and quantitative evidence. There is not one study or one statistical test that can validate a particular test for a particular purpose. Test validation is continuous, with older studies paving the way for additional research and newer studies building on the information learned in prior studies.

Finally, it should be noted that although test developers must provide evidence to support the validity of the interpretations that are likely to be made from test scores, ultimately, it is the responsibility of the *users* of a test to evaluate this evidence to ensure that the test is appropriate for the purpose(s) for which it is being used.

Messick succinctly summarized the fundamental characteristics of validity by defining validity as 'an integrated evaluative judgment of the degree to which evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment' (p. 13). By describing validity as 'integrative', he championed the notion that validity is a unitary concept centered on construct validity. He also paved the way for the *argument-based approach* to validity that was articulated by Kane [8], which is congruent with the *Standards*. We turn now to a discussion of test validation from this perspective.

Test Validation: Validity Theory Applied in Practice

To make the task of validating inferences derived from test scores both scientifically sound and manageable, Kane [8] suggested developing a defensible validity 'argument'. In this approach, the validator builds an argument on the basis of empirical evidence to support the use of a test for a particular purpose. Although this validation framework acknowledges that validity can never be established absolutely, it requires evidence that (a) the test measures what it claims to measure, (b) the test scores display adequate reliability, and (c) test scores display relationships with other variables in a manner congruent with its predicted properties. Kane's practical perspective is congruent with the *Standards*, which provide detailed guidance regarding the types of evidence that should be brought forward to support the use of a test for a particular purpose. For example, the *Standards* state that

> A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses... Ultimately, the validity of an intended interpretation... relies on all the available evidence relevant to the technical quality of a testing system. This includes evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all examinees... (AERA et al. [1], p. 17)

Two factors guiding test validation are evaluating *construct underrepresentation* and *construct-irrelevant variance*. As Messick [10] put it, 'Tests are imperfect measures of constructs because they either leave out something that should be included... or else include something that should be left out, or both' ([10], p. 34). Construct underrepresentation refers to the situation in which a test measures only a portion of the intended construct (or content domain) and leaves important knowledge, skills, and abilities untested. Construct-irrelevant variance refers to the situation in which the test measures proficiencies irrelevant to the intended construct. Examples of construct-irrelevant variance undermining test score interpretations are when computer proficiency affects performance on a computerized mathematics test, or when familiarity with a particular item format (e.g., multiple-choice items) affects performance on a reading test.

To evaluate construct underrepresentation, a test evaluator searches for content validity evidence. A preliminary question to answer is 'How is the content domain defined?' In employment, licensure, and certification testing, job analyses often help determine the content domain to be tested. Another important question to answer is 'How well does the test represent the content domain?' In educational testing, subject matter experts (SMEs) are used to evaluate test items and judge their congruence with test specifications and their relevance to the constructs measured (see [4] or [12] for descriptions of methods for evaluating content representativeness). Thus, traditional studies of content validity remain important in contemporary test validation efforts (see [14] for an example of content validation of psychological assessments).

Evaluating construct-irrelevant variance involves ruling out extraneous behaviors measured by a test. An example of construct-irrelevant variance is 'method bias', where test scores are contaminated by the mode of assessment. Campbell and Fiske [2] proposed a **multitrait-multimethod** framework for studying construct representation (e.g., convergent validity) and construct-irrelevant variance due to method bias.

Investigation of differential item functioning (DIF) is another popular method for evaluating construct-irrelevant variance. DIF refers to a situation in which test takers who are considered to be of equal proficiency on the construct measured, but who come from different groups, have different probabilities of earning a particular score on a test item. DIF is a statistical observation that involves matching test takers from different groups on the characteristic measured and then looking for performance differences on an item. Test takers of equal proficiency who belong to different groups should respond similarly to a given test item. If they do not, the item is said to function differently across groups and is classified as a DIF item (see [3], or [7] for more complete descriptions of DIF theory and methodology). Item bias is present when an item has been statistically flagged for DIF and the reason for the DIF is traced to a factor irrelevant to the construct the test is intended to measure. Therefore, for item bias to exist, a characteristic of the item that is unfair to one or more groups must be identified. Thus, a determination of item bias requires subjective judgment that a statistical observation (i.e., DIF) is due to some aspect of an item that is irrelevant to the construct measured. That is, difference observed across groups in performance on an item is due to something unfair about the item.

Another important area of evaluation in contemporary validation efforts is the analysis of the fairness of a test with respect to consistent measurement across identifiable subgroups of examinees. One popular method for looking at such consistency is analysis of *differential predictive validity*. These analyses are relevant to tests that have a predictive purpose such as admissions tests used for college, graduate schools, and professional schools. In differential predictive validity analyses, the predictive relationships across test scores and criterion variables are evaluated for consistency across different groups of examinees. The typical groups investigated are males, females, and ethnic minority groups. Most analyses use **multiple linear regression** to evaluate whether the regression slopes and intercepts are constant across groups [15].

Summary

In sum, contemporary test validation is a complex endeavor involving a variety of studies aimed toward demonstrating that a test is measuring what it claims to measure and that potential sources of invalidity are ruled out. Such studies include dimensionality analyses to ensure the structure of item response data is congruent with the intended test structure, differential item functioning analyses to rule out item bias, content validity studies to ensure the relevance and appropriateness of test content, criterion-related validity studies to evaluate hypothesized relationships among test scores and external variables, and surveys of invested stakeholders such as test takers and test administrators. Relatively recent additions to test validation are studies focusing on social considerations associated with a testing program including unintended consequences such as narrowing the curriculum to improve students' scores on educational tests. It should also be noted that evidence of adequate test score reliability is a prerequisite for supporting the use of a test for a particular purpose since inconsistency in measurement due to content sampling, task specificity, ambiguous scoring rubrics, the passage of time, and other factors adds construct-irrelevant variance (i.e., error variance) to test scores.

References

[1] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington.

[2] Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological Bulletin* **56**, 81–105.

[3] Clauser, B.E. & Mazor, K.M. (1998). Using statistical procedures to identify differentially functioning test items, *Educational Measurement: Issues and Practice* **17**(1), 31–44.

[4] Crocker, L.M., Miller, D. & Franks, E.A. (1989). Quantitative methods for assessing the fit between test and curriculum, *Applied Measurement in Education* **2**, 179–194.

[5] Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests, *Psychological Bulletin* **52**, 281–302.

[6] Downing, S.M. & Haladyna, T.M., eds (2005). *Handbook of Testing*, Lawrence Erlbaum, Mahwah.

[7] Holland, P.W. & Wainer, H. eds (1993). *Differential Item Functioning*, Lawrence Erlbaum, Hillsdale.

[8] Kane, M.T. (1992). An argument based approach to validity, *Psychological Bulletin* **112**, 527–535.

[9] Loevinger, J. (1957). Objective tests as instruments of psychological theory, *Psychological Reports* **3**,(Monograph Supplement 9), 635–694.

[10] Messick, S. (1989). Validity, in *Educational Measurement*, 3rd Edition, R. Linn, ed., American Council on Education, Washington, pp. 13–100.

[11] Sireci, S.G. (1998a). The construct of content validity, *Social Indicators Research* **45**, 83–117.

[12] Sireci, S.G. (1998b). Gathering and analyzing content validity data, *Educational Assessment* **5**, 299–321.

[13] Sireci, S.G. (2003). Validity, *Encyclopedia of Psychological Assessment*, Sage Publications, London, pp. 1067–1069.

[14] Vogt, D.S., King, D.W. & King, L.A. (2004). Focus groups in psychological assessment: enhancing content validity by consulting members of the target population, *Psychological Assessment*, **16**, 231–243.

[15] Wainer, H. & Sireci, S.G. (2005). Item and test bias, in *Encyclopedia of Social Measurement Volume 2*, Elsevier, San Diego, pp. 365–371.

Stephen G. Sireci

Variable Selection

Stanley A. Mulaik

Volume 4, pp. 2107–2110

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Variable Selection

In selecting variables for study, it is necessary to have a clear idea of what a variable is. For mathematicians, it is a quantity that may assume any one of a set of values, such as the set of integers, the set of real numbers, the set of positive numbers, and so on. But when mathematics is used to model the world, the schema of a mathematical variable must be put in correspondence to something in the world. This is possible because the schema of an object is that an object comes bearing attributes or properties. For example, a particular person will have a certain color for the eyes, another for the hair, a certain weight, a certain height, a certain age: ‘John has brown eyes, blonde hair, is 180 cm tall, weighs 82 kg, and is 45 years old’. Eye color, hair color, height, weight, and age are all variables (in a more general sense). For example, eye color pertains to the set (blue, brown, pink, green, hazel, gray, black). No person’s eye color can assume more than one of the members of this set in any one eye at any one time. Eye color is not a quantity, but it varies from person to person and no person’s eye can be more than one of the members of this set. One might map eye colors to integers, but such a mapping might not represent anything particularly interesting other than a way to distinguish individuals’ eye color with numerical names. On the other hand, weight is a quantity and pertains to the set of weights in some unit of measurement represented by positive real numbers. How a number gets assigned to the person to be his or her weight involves measurement, and this is a complex topic of it own (*see* **Measurement: Overview** and [2]). In behavioral and social sciences, variables may be identified with the responses an individual makes to items of a questionnaire. These may be scaled to represent measurements of some attribute that persons have. We will assume that the variables to be studied are measurements.

A frequent failing of researchers who have not mastered working with quantitative concepts is that in their theorizing they often fail to think of their theoretical constructs as variables. But then they seek to study them statistically using methods that require that they work with variables. A common mistake that may mislead one’s thinking is not to think of them as quantities or by names of quantities. For example, a researcher may hypothesize a causal relationship between ‘leader–follower exchange’ and ‘productivity’, or between ‘academic self-concept’ and ‘achievement’. These do not refer to quantities, except indirectly. ‘Leader–follower exchange’ does not indicate which aspect of an exchange between leader and follower is being measured and how it is a quantity. Is it ‘the *degree* to which the leader distrusts the follower to carry out orders’? Is it ‘the *frequency* with which the follower seeks advice from the leader’? Is it ‘the *extent* to which the leader allows the follower to make decisions on his own’? And what is ‘productivity’? What is produced? Can it be quantified? As for ‘academic self-concept’, there are numerous variables by which one may describe oneself. So, which is it? Is it ‘the *degree* to which the individual feels confident with students of the opposite sex’? Is it ‘the*frequency* with which the individual reports partying with friends during a school year’? Is it ‘the number of hours the individual reports he/she studies each week’? And is ‘achievement’ measured by ‘GPA’ or ‘final exam score’, or ‘postgraduate income’? When one names variables, one should name them in a way that accurately describes what is measured. Forcing one to use words like ‘degree’, ‘extent’, ‘frequency’, ‘score’, ‘number of’ in defining variables will help in thinking concretely about what quantities have influence on other quantities, and improve the design of studies.

Selecting variables for regression. Suppose Y is a criterion (dependent) variable to be predicted and X_i is a predictor (independent variable) in a set of p predictors. It is known that the regression coefficient between variable X_i and variable Y *is*

$$\beta_{Yi} = \rho_{Yi|12\cdots(i)\cdots p}\frac{\sigma_{Y|12\cdots(i)\cdots p}}{\sigma_{i|12\cdots(i)\cdots p}}, \qquad (1)$$

where $\rho_{Yi|12\cdots(i)\cdots p}$ is the partial correlation between the criterion Y and predictor X_i, with the predicted effects of all other predictors subtracted out of them. (Note ‘$12\cdots(i)\cdots p$’ means ‘the variables X_1, X_2 through to X_p with variable X_i not included’.) $\sigma_{Y|12\cdots(i)\cdots p}$ is the conditional standard deviation of the dependent variable Y, with the predicted effects of all predictor variables except variable X_i subtracted out. $\sigma_{i|12\cdots(i)\cdots p}$ is the conditional standard deviation of X_i, with the predicted effects of all other predictors subtracted from it. So, the regression weight represents the degree to which the part of X_i that is relatively unique with respect to the other predictors is still related to a part of the criterion not predicted by the other variables. This means that an ideal predictor is one that has little in common with the other predictors but much in common with the criterion Y. In an ideal extreme case, the predictors have zero correlations among themselves, so that they have nothing in common among them, but each of the predictors has a strong correlation with the criterion.

There are several posthoc procedures for selecting variables as predictors in regression analysis. The method of *simultaneous regression* estimates the regression coefficients of all of the potential predictor variables given for consideration simultaneously. Thus, the regression weight of any predictor is relative to the other predictors included, because it concerns the relationship of the predictor to the criterion variable, with the predicted effects of other predictors subtracted out. The success of the set of predictors is given by the multiple correlation coefficient R (*see* **Multiple Linear Regression**). This quantity varies between 0 and 1, with one indicating perfect prediction by the predictors. The squared multiple correlation R^2 gives the proportion of the variance of the criterion variable accounted for by the full set of predictors. One way of assessing the importance of a variable to prediction is to compute the relative gain in the proportion of variance accounted for by adding the variable to the other variables in the prediction set. Let W be the full set of p predictors including X_j. Let V be the set of $p - 1$ predictors $W - \{X_j\}$. Then $r^2_{yj\cdot V} = (R^2_{y\cdot W} - R^2_{y\cdot V})/(1 - R^2_{y\cdot V})$ is the relative gain in proportion of variance accounted for due to including variable X_j with the predictors in set V. Here, $r^2_{yj\cdot V}$ is also the squared partial correlation of variable X_j with criterion variable Y, holding constant the variables in $V.R^2_{y\cdot W}$ is the squared multiple correlation for predicting Y from the full set of predictors $W.R^2_{y\cdot V}$ is the squared multiple correlation for predicting Y from the reduced set $V.r^2_{yj\cdot V}$ can be computed for each variable and relative importance compared among the predictors. It is possible also to test whether the absolute incremental gain in proportion of variance accounted for due to variable X_j is significantly different from zero. This is given by the formula $F = (R^2_{y\cdot W} - R^2_{y\cdot V})/[(1 - R^2_{y\cdot W})(N - p - 1)]$. F is distributed as chi squared with 1 and $(N - p - 1)$ degrees of freedom. (Source: [1], pp. 719–720).

Another method is the *hierarchical* method. The predictor variables are arranged in a specific rational order based on the research question. We begin by entering the first variable in the order and determine the degree to which it explains variance in the dependent variable. We then examine how much the next variable in the order adds to the proportion of variance accounted for beyond the first, then how much the next variable afterwards adds to the proportion of variance accounted for beyond the first two, and so on, to include finally the pth predictor beyond the first $p - 1$ predictors. In the end, the R^2 for the full set will be the same as in the simultaneous method, and the regression weights will be the same. But this method can suggest at what point what variables might be dropped from consideration (Source: [1], pp. 731–732).

In addition to this procedure based on rational ordering, entry can be based on empirical ordering. The method of *stepwise regression* begins with no preset order to the variables. Usually, one begins with the variable that has the largest squared correlation with the criterion. Then one pairs the first variable with each of the other variables and computes a squared multiple correlation for each pair with the criterion. One adds the variable from the remaining variables that produces the largest squared multiple correlation among the pairs. One also computes the relative gain in proportion of variance accounted for. Then one seeks a third variable from the remaining predictors, which when added to the first two, produces the largest squared multiple correlation with the criterion. Again one also computes the relative gain in proportion of variance accounted for. And one keeps on until either one has selected all of the variables or comes to a point where no additional variable adds a meaningful or significant increment in proportion of variance accounted for. If one accompanies this with significance tests for the gain, it can involve many tests which are not statistically independent (Source: [1], pp. 732–735).

One can also proceed in a *backward* or *step-down* direction, beginning with all the variables and eliminating, one at a time, variables that successively account for the least decrement in the squared multiple correlation at each step until one gets to a step where any additional elimination of variables would seriously decrease the squared multiple correlation at that point.

An important disadvantage of empirical entry/selection is that it capitalizes on chance. This is particularly important when predictors are highly related to one another. In a given data set, any of the empirical entry procedures might lead to a certain order of entry, whereas in another data set, with values drawn from the same populations of values, the order might be quite different.

References

[1] Hays, W.L. (1994). *Statistics*. Fort Worth, TX: Harcourt Brace.

[2] Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement*, Lawrence Erlbaum, Hillsdale.

Stanley A. Mulaik

Variance

DAVID CLARK-CARTER

Volume 4, pp. 2110–2110

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Variance

The variance is defined as the mean of the squared deviations of a set of numbers about their mean. If a population consists of the set of values $X_1, X_2, \ldots, X_n (i = 1, 2, \ldots, n)$, then the population variance, usually denoted by σ^2, is

$$\sigma^2 = \frac{\sum_i (X_i - M)^2}{n}, \tag{1}$$

where M is the population **mean**.

This formula is also appropriate for finding the variance of a sample in the unlikely event that we know the population mean. However, if we are dealing with a sample and we are also obliged to calculate the sample mean, $\bar{X}$, then dividing the sum of squared deviations by a value fewer than the number of scores in the set $(n - 1)$ produces the best estimate of the variance in the population from which the sample was obtained. This form, often denoted by S^2, is called the *sample variance* and is defined as

$$S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n - 1}. \tag{2}$$

It is this version of the variance that is usually reported by statistical software. Notice, though, that as the sample size increases, there will be less and less difference between the values of the variance obtained using n or $(n - 1)$. Even for small samples, if the variance is required only for descriptive purposes, it is usually immaterial as to which divisor is used.

Table 1 shows the number of words recalled by a hypothetical sample of six participants in a study of short-term memory. The mean recall for the sample is 7.5 words. The sum of the squared deviations is 23.5.

Table 1 Variance calculations for word recall data

Participant	Words recalled	Deviation from mean	Squared deviation
1	5	−2.5	6.25
2	7	−0.5	0.25
3	6	−1.5	2.25
4	9	1.5	2.25
5	11	3.5	12.25
6	7	−0.5	0.25

The version of the variance treating the data as a population (that is dividing this sum of the squared deviations by 6) is 3.92, while the version of the variance that *estimates* the population variance (dividing by 5) is 4.7.

Calculating variances using the above method can be tedious, so it is worth noting that there are easier computational forms (see, for example, [1]).

As with other summary statistics that rely equally on all the numbers in the set, the variance can be severely affected by extreme scores. Nonetheless, the variance underlies inferential procedures such as the **analysis of variance**, while other methods, such as standardizing a set of data, draw on its close relative, the **standard deviation**.

Reference

[1] Howell, D.C. (2004). *Fundamental Statistics for the Behavioral Sciences*, 5th Edition, Duxbury Press, Pacific Grove.

David Clark-Carter

Variance Components

Nicholas T. Longford

Volume 4, pp. 2110–2113

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Variance Components

Variability is a ubiquitous feature of characteristics and attributes studied in human populations. Its prominence is a result of our interest in the details of behavioral, attitudinal, economic, and medical outcomes, aspects in which human subjects differ widely and with limited predictability. Without variability, there is not much to study about a population, because one or a few of its members then inform about the population completely. On the other hand, explanation of variability, in the form of a mechanism describing its causes, is the ultimate and often unachievable research goal. Such goals are pursued by various regression models and their extensions, and the commonly used terminology in connection with imperfect model fit, such as 'unexplained variation', implies a failure of the research effort. Such a pessimistic view is often poorly supported. In many situations, we have to resign ourselves to a description of variability less complete than by regression. *Variance components* are a key device for such a description, and they are associated with factors or *contexts* involved in the analyzed study. The contexts may be introduced deliberately, or be an unavoidable nuisance feature. For example, when essays (examination papers) are graded by raters, the raters are such a context. Their training and instruction aims to reduce the differences among the raters, but eradicating them completely is not possible [4].

In this example, the elementary observations (essays) are clustered within raters; the essays graded by a rater form clusters or level-2 units (*see* **Clustered Data**). Even if the essays are assigned to raters uninformatively (at random, or with no regard for any factor associated with their quality), the scores assigned to them have an element of similarity brought on by being graded by the same rater. Inferences are desired for each essay and an 'average' or typical rater; we wish to *generalize* our findings from the incidental to the universal. Hence the term *generalizability theory* introduced by [1] (*see* **Generalizability**).

In practice, we come across several contexts 'interfering' with our observations simultaneously. For example, intellectual or academic abilities can be studied only indirectly, by observing individuals' performances on tasks that represent the domain of abilities, and assessing the performances by raters using scales constructed with simplicity in mind. Here, the assessment (testing) instrument, or the individual tasks, the setting (occasion) of the text or examination, and the raters are factors (or contexts) associated with variation. That is, if a different testing instrument were used (and the settings of all other contexts held fixed), the assessments of the (same) subjects would be different. If a different rater graded a particular response, the score assigned may be different. Ability, with the appropriate qualification, is regarded as a characteristic of an individual, whereas the individual's performances vary around the ability, depending on the momentary disposition, the balance of questions, tasks, or the like, in the assessment instrument and the rater's judgment, which may be inconsistent over replications and may differ among the raters. Thus, apart from the variation in the ability, that is of key inferential interest, several other sources contribute to the variation of the outcomes (scores) – in a hypothetical replication of the study, different scores would be recorded. These *nuisance* contexts are unavoidable, or have to be introduced, because assessment (measurement) cannot be conducted in a contextual vacuum. The levels (or settings) of a context can be regarded as a *population*, thus introducing the sampling-design issue of good representation of a context in the study. This clarifies the scope of the inference – for what range of settings the conclusions are intended and are appropriate.

For the case of a single context, let y_{ij} be the outcome for elementary-level unit (say, subject) i in setting j (say, group, cluster, or unit at level 2). The simplest nontrivial model that describes y_{ij} is

$$y_{ij} = \mu + \delta_j + \varepsilon_{ij}, \qquad (1)$$

where μ is the overall mean that corresponds to (population-related) averaging over settings and elements, δ_j is the deviation specific to setting j, $j = 1, \ldots, J$, and ε_{ij} represents the deviation of subject i, $i = 1, \ldots, n_j$, from the setting-specific mean $\mu + \delta_j$. The deviations δ_j and ε_{ij} are mutually independent random samples from centered distributions with respective variances σ_1^2 and σ_2^2. Usually, these distributions are assumed to be normal, often as a matter of expedience because the normal distribution is closed with respect to addition: that is, if a and b are normally distributed random variables, then so is their total $a + b$, with mean $\mathrm{E}(a) + \mathrm{E}(b)$ and variance $\mathrm{var}(a) + \mathrm{var}(b) + 2\mathrm{cov}(a, b)$.

As $\mathrm{var}(y_{ij}) = \sigma_1^2 + \sigma_2^2$, it is meaningful to call σ_1^2 and σ_2^2 the variance components associated with subjects and settings, respectively. The deviations δ_j cause the observations y_{ij} to be correlated – observations within settings are more similar than observations in general. We have

$$\mathrm{cor}(y_{ij}, y_{i'j}) = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \qquad (2)$$

for any pair of observations $i \neq i'$ in the same setting. If a single setting j were studied, this correlation could not be recognized (identified); δ_j and μ would be confounded:

$$y_{ij} = \mu_j + \varepsilon_{ij} \qquad (3)$$

and, as μ_j is a constant (unchanged in replications), the outcomes are mutually independent, each with variance σ_1^2. That is, $\mathrm{var}(y_{ij}|j) = \sigma_1^2$. (Instead of conditioning on setting j we can condition on the expectation μ_j.) If the setting-specific expectations μ_j were the observations, σ_2^2 would be identified as their variance. It is essential to distinguish between μ_j and μ, even when the study involves a single setting j of the context. Otherwise, an unjustified generalization is made that all the settings of the context are identical. Also, the parameter μ should be qualified by the population (class) of contexts; different classes of a context are associated with different values of μ.

When several contexts are present each of them is represented by a variance (component). The contexts may be nested, as with individuals within households, streets, towns, and countries, or crossed, as with examination papers rated separately by two raters, each drawn from the same pool (*see* **Cross-classified and Multiple Membership Models**). In general, nested contexts are much easier to handle analytically, although estimation and other forms of inference simplify substantially when the numbers of observations within the combinations of levels of the contexts are equal (*balanced* design).

When the context-specific deviations are additive, as in

$$y_{hij} = \mu + \delta_h + \gamma_i + \varepsilon_{hij} \qquad (4)$$

(contexts h and i are crossed), either a suitable notation has to be introduced, or the model supplemented with a description of how the contexts appear in the data and in the relevant populations. For example, each level h of one context can occur with each level i of the other context. The covariance of two observations is equal to the total of the variance components for the contexts shared by the two observations. For example, $\mathrm{cov}(y_{hij}, y_{hi'j'}) = \sigma_\delta^2$.

Variance component models can be combined with regression. The model in (1) has the extension

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \delta_j + \varepsilon_{ij}, \qquad (5)$$

where $\mathbf{x}_{ij}$ are the covariates associated with the element ij. Each covariate may be defined for the elements ij or for the setting j (of the context); a more detailed model formulation is

$$y_{ij} = \mathbf{x}_{1,ij}\boldsymbol{\beta}_1 + \mathbf{x}_{2,j}\boldsymbol{\beta}_2 + \delta_j + \varepsilon_{ij}, \qquad (6)$$

in which the covariates $\mathbf{x}_{1,ij}$ are defined for the elements and $\mathbf{x}_{2,j}$ for the settings. A variable defined for the elements may be constant within settings; $x_{ij} \equiv x_j^*$; this may be the case only in the sample (it would not be the case in each replication), or in the relevant population. Conversely, each context-level variable x_j can be regarded as an elementary-level variable by defining $x_{ij}^\dagger \equiv x_j$.

In most research problems, the role of the covariates is to reduce the variance components $\mathrm{var}(\varepsilon_{ij}) = \sigma_1^2$ and $\mathrm{var}(\delta_j) = \sigma_2^2$. Adding a context-level covariate to $\mathbf{x}$ can reduce only the context-level variance σ_2^2, whereas the addition of an elementary-level variable to $\mathbf{x}$ can reduce both σ_1^2 and σ_2^2. A variable with the same distribution at each setting of the context reduces only the within-context variation. Variables with both within- and between-setting components of variation can reduce both variance components. Counterintuitive examples in which variance components are increased when the model is expanded by one or several variables are given in [6].

The model in (5) involves several restrictive assumptions. First, linearity is often adopted as a matter of analytical convenience. It can be dispensed with by replacing the predictor $\mathbf{x}\boldsymbol{\beta}$ with a general function $f(\mathbf{x}; \boldsymbol{\theta})$ which would, nevertheless, involve a functional form of f. Next, the variance components are constant. Heteroscedasticity of the context can be introduced by assuming a functional form for $\mathrm{var}(\delta_j)$, dependent on some variables (*see* **Heteroscedasticity and Complex Variation**). No generality is lost by assuming that these variables are a subset of $\mathbf{x}$. The simplest nontrivial example assumes that the levels of the context belong to a small number of categories (subpopulations), each with its own variance. In another example, $\mathrm{var}(\delta_j) = \exp(m_j)$, where m_j is a function of the population size of context j.

The deviation of the context from the average regression $\mathbf{x}\boldsymbol{\beta}$ need not be constant. A natural way of introducing this feature is by random coefficients, or by a more general pattern of between-context variation:

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\delta_j + \varepsilon_{ij}, \qquad (7)$$

where $\mathbf{z}$ is a subset of the variables in $\mathbf{x}$. Now $\mathrm{var}(\mathbf{z}_{ij}\boldsymbol{\delta}_j) = \mathbf{z}_{ij}\boldsymbol{\Sigma}\mathbf{z}_{ij}^{\top}$, where $\boldsymbol{\Sigma} = \mathrm{var}(\boldsymbol{\delta}_j)$. If all variables in $\mathbf{z}$ are defined for elements the model in (7) has an interpretation in terms of varying within-context regressions. But the *variation part* of the model, $\mathbf{z}_{ij}\boldsymbol{\delta}_j$, may, in principle, involve any variables. Including a context-level variable in $\mathbf{z}$ is meaningful only when the context has many levels (categories); otherwise estimation of the elements of the variance matrix $\boldsymbol{\Sigma}$ is an ill-conditioned problem.

Finally, the deviations of the contexts and elements within contexts, after an appropriate regression adjustment, need not be additive. The random deviations are multiplicative in the model

$$y_{ij} = f(\mathbf{x};\boldsymbol{\beta})\,\delta_j\,\varepsilon_{ij}, \qquad (8)$$

where δ_j and ε_{ij} are random samples from nonnegative distributions with unit means. This model reduces to additive deviations (and variance components) by the log-transformation. Although nontrivial nonadditive models are easy to specify, they are applied in practice infrequently, because additive random terms are much more tractable and correspond to addition of their variances. Moreover, when the deviations are normally distributed, so are their linear combinations, including totals.

Generalized linear models provide a vehicle for substantial extension of variance component models within the realm of linearity and normality of the deviations. Thus a model

$$g\{\mathrm{E}(y_{ij})\} = \mathbf{x}_{ij}\boldsymbol{\beta}, \qquad (9)$$

with a suitable link function g and a distributional assumption for y (say, logit function g and binary y) is extended to

$$g\{\mathrm{E}(y_{ij}\,|\,\delta_j)\} = \mathbf{x}_{ij}\boldsymbol{\beta} + \delta_j, \qquad (10)$$

with the obvious further extensions to varying within-setting regressions (*see* **Generalized Linear Mixed Models**).

A related way of defining more general variance component models is by assuming that an ordinal or dichotomous observed variable y is the *manifest* version of a normally distributed *latent* variable y^* that satisfies an additive variance component model. A generic approach to fitting such models is by the EM algorithm (*see* **History of Intelligence Measurement**) [2], regarding the latent outcomes as the missing information. See [5].

A suitable process for the conversion of latent values to their manifest versions has to be defined. As many key variables in behavioral research are ordinal categorical, a coarsening process [3] can be posited in many settings. It specifies that there is a small number of cut-points $-\infty = c_0 < c_1 < \cdots < c_{K-1} < c_K = +\infty$ and all latent values that fall into the interval (c_{k-1}, c_k) convert to manifest value k. Note that the decomposition of the variance to its context-related components is meaningful only for the latent outcomes. The pattern of dependence or the covariance structure among the values of the observed outcome y is usually not related in any straightforward way to the covariance structure on the latent scale. The connection can usually be explored only by simulations.

References

[1] Cronbach, L.J. Gleser, G.C. Nanda, H. & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements. The Theory of Generalizability of Scores and Profiles*, Wiley and Sons, New York.

[2] Dempster, A.P. Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society Series B* **39**, 1–38.

[3] Heitjan, D.F. & Rubin, D.B. (1990). Inference from coarse data via multiple imputation with application to age heaping, *Journal of the American Statistical Association* **85**, 304–314.

[4] Longford, N.T. (1995). *Models for Uncertainty in Educational Testing*, Springer-Verlag, New York.

[5] McKelvey, R. & Zavoina, W. (1975). A statistical model for the analysis of ordinal dependent variables, *Journal of Mathematical Sociology* **4**, 103–120.

[6] Snijders, T.A.B. & Bosker, R. (1999). *Multilevel Analysis*, Sage Publications, London.

(*See also* **Linear Multilevel Models**)

NICHOLAS T. LONGFORD

Walsh Averages

Clifford E. Lunneborg

Volume 4, pp. 2115–2116

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Walsh Averages

Let $(x_1, x_2, \dots, x_n)$ be a random sample from a symmetric distribution with unknown median θ. The set of Walsh averages [4] is the collection of $n(n+1)/2$ pairwise averages, each of the form $(x_i + x_j)/2$ and computed for all $i = 1, 2, \dots, n$ and for all $j = 1, 2, \dots, n$.

The sample (2, 5, 7, 11), for example, gives rise to the 10 Walsh averages shown in Table 1.

The median of the set of Walsh averages is the one-sample **Hodges-Lehmann estimator**, an efficient point estimator of the median of the sampled distribution. For our example, with 10 Walsh averages, the median estimate is the average of the fifth and sixth smallest Walsh averages, $(6 + 6.5)/2 = 6.25$.

Walsh averages also can be used to find a **confidence interval** (CI) for the median. This usage follows the logic of the Wilcoxon **signed-rank test**; that is, the resulting $(1 - \alpha)100\%$ CI includes those values of θ that would not be rejected at the α level by the Wilcoxon signed-rank test.

One approach to the median CI is described in [1]. We illustrate here the mechanics, using our $n = 4$ example. Let $L_{\alpha/2}$ be a signed-rank sum such that, for a sample of size n, the probability is $\alpha/2$ of a signed-rank sum that size or smaller under the null hypothesis. Such values are tabulated and widely available. Large-sample approximations have been developed as well, for example, [1] and [2].

The lower and upper limits to the $(1 - \alpha)100\%$ CI are given by the $L_{\alpha/2}$th smallest and the $L_{\alpha/2}$th largest Walsh average, respectively.

For $n = 4$, the tabled [3] probability is 0.125 that the signed-rank sum will be 1 or smaller. Thus, a 75% CI for the median is bounded by the smallest and largest Walsh averages. For our example, this yields a 75% CI bounded below by 2 and above by 11. We have 75% confidence that the interval from 2 to 11 contains the population median, θ. A slightly different algorithm for the CI is given in [2], where it is referred to as a Tukey CI.

References

[1] Conover, W.J. (1999). *Practical Nonparametric Statistics*, 3rd Edition, Wiley, New York.

[2] Hollander, M. & Wolfe, D.A. (1999). *Nonparametric Statistical Methods*, 2nd EditionWiley, New York.

[3] Larsen, R.J. & Marx, M.L. (2001). *An Introduction to Mathematical Statistics and its Applications*, 3rd Edition, Prentice-Hall, Upper Saddle River.

[4] Walsh, J.E. (1949). Some significance tests for the median which are valid under very general conditions, *SAnnals of Mathematical Statistics* **20**, 64–81.

CLIFFORD E. LUNNEBORG

Table 1 Computation of Walsh averages

	2	5	7	11
2	$(2 + 2)/2 = 2$	$(2 + 5)/2 = 3.5$	$(2 + 7)/2 = 4.5$	$(2 + 11)/2 = 6.5$
5		$(5 + 5)/2 = 5$	$(5 + 7)/2 = 6$	$(5 + 11)/2 = 8$
7			$(7 + 7)/2 = 7$	$(7 + 11)/2 = 9$
11				$(11 + 11)/2 = 11$

Wiener, Norbert

Brian S. Everitt

Volume 4, pp. 2116–2117

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Wiener, Norbert

Born: November 26, 1894, in Columbia, USA.
Died: March 18, 1964, in Stockholm, Sweden.

The son of Russian émigré, Leo Wiener, Norbert Wiener was a child prodigy who entered Tufts College at the age of 11, graduating three years later. He then entered Harvard to begin graduate studies at the age of 14. Beginning with zoology, Wiener soon changed to mathematical philosophy, receiving his Ph.D. from Harvard at the age of 18, with a dissertation on mathematical logic. From Harvard, the 18-year-old Wiener travelled first to Cambridge, England, to study under Russell and Hardy, and then on to Gottingen to work on differential equations under Hilbert. At the end of World War I, Wiener took up a mathematics post at the Massachusetts Institute of Technology (MIT), where he was eventually to become professor in 1932, a post he held until 1960.

Wiener's mathematical work included functions of a real variable, mathematical logic, relativity, quantum theory, Brownian motion, and the Fourier integral and many of its applications. During World War II, he worked on guided missiles. It was during this period that Wiener studied the handling of information by electronic devices, based on the feedback principle, phenomena that were later compared with human mental processes in his most famous book, *Cybernetics*, first published in 1947 [1]. Cybernetics was generally defined as 'the science of control and communication in the animal and the machine', with 'animal' very definitely including human beings. Essentially, Wiener was making an analogy between man as a self-regulating system, receiving sensory data and pursuing certain objectives, and mechanical or electrical servomechanisms.

Wiener spent his last years working on the applications of cybernetics to human thinking and pointing out its social dangers and the need to reconstruct society to allow for social evolution. Wiener's cybernetics was the forerunner of Artificial Intelligence. Norbert Wiener was one of twentieth century's true polymaths, and the breadth of his work remains apparent throughout mathematics and probability.

Reference

[1] Wiener, N. (1948). *Cybernetics*, Wiley, New York.

BRIAN S. EVERITT

Wilcoxon, Frank

David C. Howell

Volume 4, pp. 2117–2118

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Wilcoxon, Frank

Born: September 2, 1892, in County Cork, Ireland.
Died: November 18, 1965, in Tallahassee, Florida.

Frank Wilcoxon was born to a wealthy family in 1892 and grew up in the Hudson River Valley. He received his early education at home, and a BS degree from Pennsylvania Military College in 1917. He received his MS degree in chemistry from Rutgers in 1921 and a PhD degree in inorganic chemistry from Cornell in 1924.

Most of Wilcoxon's professional life was spent in industry, first with the Boyce Thompson Institute for Plant Research, and then with American Cyanamid. The major focus of his chemical research dealt with insecticides and fungicides, though later he headed up the statistics group at the Lederle Division of American Cyanamid.

Wilcoxon's interest in statistics began with a study group, of which W. J. Youden was also a part. The group studied **Fisher's** (1925) *Statistical Methods for Research Workers*. It was 20 years before Wilcoxon published anything in statistics, but he was always interested in the applications of statistical methods to chemical research.

Wilcoxon's most important paper [10] appeared in the first volume of what is now *Biometrics* in 1945, and concerned the application of ranking methods for testing differences in location. Both his matched-pairs **signed-ranks** test and his rank-sum test (*see* **Wilcoxon–Mann–Whitney Test**) were presented in that paper. This paper was important because it led to a growing interest in rank methods (*see* **Rank Based Inference**), and the development of similar methods for other designs.

Wilcoxon's approach relies on two statistical methods, the use of ranks and the use of permutation procedures. Ranked data had been used for a very long time, at least back to **Galton**. However, according to Kruskal and Wallis [7], the earliest treatment of them as a nonparametric statistical tool would appear to be a paper on rank correlation by Hotelling and Pabst [5] in 1936. Wilcoxon cited a paper by Friedman [4] on the use of ranks to avoid assumptions of normality but, interestingly, his own paper says very little about that issue, even though it is one of the major strengths of his approach. Permutation as a test procedure was considerably more recent and was first used by Fisher [3] and by Pitman [9] in the 1930s. The permutation tests (*see* **Permutation Based Inference**) produced by Fisher and Pitman (*see* **Pitman Test**) were unwieldy, requiring lengthy calculations on all possible (or all extreme) permutations. Wilcoxon, however, hit upon the idea of replacing observations with ranks and permuting the ranks. The first thing this did was to simplify the calculations, which his paper seems to emphasize as the goal. Since, for a given sample size, ranks are constant from one experiment to another, it was possible for Wilcoxon to establish tables of extreme results, thereby standardizing the process. More importantly, using rank substitution allowed statistics to move away from normality assumptions that had underlain nonpermutation tests to that time. Initially, Wilcoxon said very little about what assumptions remained.

Interestingly, Leon Festinger [2] independently developed the same test, retaining the possibility of unequal sample sizes, and published in *Psychometrika* the next year. Mann and Whitney [8] published a very similar idea the next year, and the two-sample test is now frequently referred to as the *Wilcoxon-Mann-Whitney test.* Over the next several years, Wilcoxon, later in conjunction with Roberta Wilcox, published extensive table for his tests [11].

Wilcoxon's tests went on to form the core of a whole set of rank-permutation tests and remain some of the most powerful nonparametric tests. Kruskal [6] provides historical notes on the development of the two-sample test prior to Wilcoxon's time. Bradley [1] provides a summary of his life.

Wilcoxon retired in 1957 but joined the faculty of Florida State University in 1960, where he remained until his death. His later work dealt with sequential ranking methods.

References

[1] Bradley, R.A. (1966). Obituary: Frank Wilcoxon, *Biometrics* **22**, 192–194.

[2] Festinger, L. (1946). The significance of differences between means without reference to the frequency distribution function, *Psychometrika* **11**, 97–105.

[3] Fisher, R.A. (1935). *The Design of Experiments*, Oliver & Boyd, Edinburgh.

[4] Friedman, M. (1937). The use of ranks to avoid the assumption of normality, *Journal of the American Statistical Association* **32**, 675–701.

[5] Hotelling, H. & Pabst, M.R. (1936). Rank correlation and tests of significance involving no assumption of normality, *Annals of Mathematical Statistics* **7**, 29–43.

[6] Kruskal, W.H. (1958). Historical notes on the Wilcoxon unpaired two-sample test, *Journal of the American Statistical Association* **52**, 356–360.

[7] Kruskal, W.H. & Wallis, W.A. (1952). Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association* **47**, 583–621.

[8] Mann, H.B. & Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics* **18**, 50–60.

[9] Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any populations, *Supplement to the Journal of the Royal Statistical Society* **4**, 119–130.

[10] Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics Bulletin (now Biometrics)* **1**, 80–83.

[11] Wilcoxon, F. (1949). *Some Rapid Approximate Statistical Procedures*, American Cyanamid Company, New York.

DAVID C. HOWELL

Wilcoxon–Mann–Whitney Test

Venita DePuy, Vance W. Berger and YanYan Zhou

Volume 4, pp. 2118–2121

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Wilcoxon–Mann–Whitney Test

The Wilcoxon-Mann-Whitney test, also known as the Mann-Whitney U-test or the Wilcoxon rank sum test, is used to test the null hypothesis that two populations have identical distribution functions against the alternative hypothesis that the two distribution functions differ only with respect to location (the median). The alternative may be either directional or nondirectional. For example, in one study [9] the investigators wished to determine whether men and women supported traditional gender roles in differing amounts. After evaluating each person's level of support, the scores were grouped together and ranked from smallest to largest. The sum of the ranks for each gender were then compared. If the two rank sums differed significantly, then the conclusion would be that there is a significant gender difference.

One common approach to testing this hypothesis is by means of a parametric analysis, which assumes normality; typically a two-sample Student's t Test (*see* **Catalogue of Parametric Tests**). A problem with this approach is that the normality assumption is rarely warranted, and when it is not, the normal-based P value will differ from the exact P value by an amount that cannot be determined without actually computing each one, and then comparing them [1, 2]. But if one were to go through the effort to compute the exact P value, then why would one use it to validate the approximation rather than as the P value to be reported? How can an approximation be preferred to the quantity it is trying to approximate, if that gold standard quantity is already in hand? The alternative nonparametric approach does not assume normality and hence offers more robust results.

There are three primary assumptions of the nonparametric Wilcoxon-Mann-Whitney test:

1. Each sample is randomly selected from the specific population and the observations within each sample are independent and identically distributed.
2. The two samples are independent of each other (otherwise, consider the Wilcoxon **signed rank test**).
3. The populations may differ in their location (mean or median), but not in their distributional shape or spread (if this assumption is questionable, then consider the **Smirnov test** [11]).

Let $x_{1,\dots,}x_m$ be a sample from population X and $y_{1,\dots,}y_n$ be a sample from population Y. Let $F_X(t)$ and $G_Y(t)$ be the cumulative distribution functions for the two populations. The location shift model is that $G_Y(t) = F_X(t + \Delta)$, for all values of t. The null hypothesis, that the X and Y variables have the same probability distribution, can then be stated as: H_0: $\Delta = 0$. This makes sense only if the distributions are continuous, but the Wilcoxon-Mann-Whitney test can be applied even to ordered categorical data.

Let $N = m + n$. To compute the Wilcoxon test statistic, W, first order the combined sample of N values from least to greatest and assign ranks, 1 through N, to the observations (average ranks can be used in case of ties). Next, let $S_1, S_2, \dots, S_n$ denote the ranks assigned to $y_1, \dots, y_n$, the sample from population Y. Then W is the sum of the ranks assigned to the Y sample [14],

$$W = \sum_{j=1}^{n} S_j. \tag{1}$$

Interpretation of the Wilcoxon test statistic depends on the substantive hypothesis about the two population medians; hence the alternative hypothesis for the location shift parameter Δ may be either one-sided or two-sided:

a. $[Mdn(Y) > Mdn(X)]$ and H_a: $\Delta > 0$. H_0 is rejected at the α level of significance if $W \geq W_{\alpha,}$ where W_α is chosen to make the type I error probability equal to (or no greater than) α.
b. $[Mdn(Y) < Mdn(X)]$ and H_a: $\Delta < 0$. H_0 is rejected at the α level of significance if $W \leq n(m + n + 1) - W_{\alpha,}$ where W_α is chosen to make the type I error probability equal to (or no greater than) α.
c. $[Mdn(Y) \neq Mdn(X)]$ and H_a: $\Delta \neq 0$. H_0 is rejected at the α level of significance if $W \geq W_{\alpha/2}$ or $W \leq n(m + n + 1) - W_{\alpha/2,}$ where $W_{\alpha/2}$ is chosen to make the overall type I error probability equal to (or no greater than) α.

The Wilcoxon test is a permutation test (*see* **Permutation Based Inference**); that is, under the null hypothesis, the set of ranks may be permuted, any m of them being assigned to the X population, with the remaining n being assigned to the Y population regardless of true group membership. Thus, the null distribution for W consists of the set of Wilcoxon statistics for all $[N!/(m! \times n!)]$ permutations of the ranks. W_α is the $(1-\alpha)100\%$ quantile of this reference distribution. Because ranks, rather than raw data, are used in the computation of W, it is not necessary to carry out the permutation test for each new data set. For sample sizes, n or m, up to 10, the critical value of W_α may be obtained from tables such as in [7]. It should be noted that the tabled values are exact only if there are no ties among the $(m+n)$ observations.

For larger sample sizes, an approximation based on the asymptotic normality of W is often used [7]. For this purpose, W is standardized:

$$W^* = \frac{W - \{n(m+n+1)/2\}}{\sqrt{mn(m+n+1)/12}}. \tag{2}$$

The null hypothesis is rejected in favor of $\mathrm{H_a}$: $\Delta > 0$ if $W^* \geq -z_\alpha$ where z_α is the $(\alpha \times 100)\%$ quantile of the standard normal distribution, for example, -1.96 is the 5% quantile. Where the alternative hypothesis is that $\Delta > 0$, reject $\mathrm{H_0}$ if $W^* \leq z_\alpha$ and where the alternative hypothesis is nondirectional, reject the null hypothesis if $|W^*| \geq -z_{\alpha/2}$.

Both the exact permutation test and approximate calculations are available from a variety of commercial statistical packages (*see* **Software for Statistical Analyses**).

Correction for Ties

The use of this normal approximation may be unreliable when the sampled distributions are sparse, skewed, or heavily tied. In fact, when ties are present in the sample, the test statistic given above may be too conservative. The following revision decreases the denominator slightly, rendering the outcome less conservative and yielding smaller P values.

$$W^* = \frac{W - \{n(m+n+1)/2\}}{\sqrt{\dfrac{mn}{12}\left[m+n+1\,\dfrac{\displaystyle\sum_{j=1}^{g}(t_j-1)t_j(t_j+1)}{(m+n)(m+n-1)}\right]}}, \tag{3}$$

where g is the number of tied groups and t_j is the size of the jth tied group.

Continuity Correction

When the large sample approximation is used, a continuity correction may be desirable to allow for the fact that the test statistic W has a discrete distribution, whereas the normal distribution is continuous. This correction decreases the numerator and renders the outcome more conservative [12].

The *U*-test Formulation

The Mann–Whitney U-test has a completely different derivation, and is based on the set of pair-wise comparisons of x values to y values. There are $n \times m$ such pair-wise comparisons. For each comparison, the Mann–Whitney U statistic is incremented by 0 if $x_i > y_j$, by 1/2 if $x_i = y_j$, and by 1 if $x_i < y_j$. The resulting sum is related to the Wilcoxon W,

$$W = U + \left(\tfrac{1}{2}\right)[n(n+1)], \tag{4}$$

with the result that tests based on U are equivalent to tests based on W [10]. The U-test formulation allows a more natural handling of data that are only partially ordered [4].

Computation

Except where the sample sizes are small, the Wilcoxon-Mann-Whitney test usually is evaluated within a statistical computing package. Comparisons of the Wilcoxon-Mann-Whitney test in 11 commercial statistical packages is presented in [5]. Many packages offer corrections for ties and continuity as well as exact computation, although only SAS was found to have all three options in this study.

The Wilcoxon Test as a Linear Rank Test

The Wilcoxon test is a linear rank test. That is, the W statistic is a weighted sum of ranks. The regular spacing of the ranks, as integers, contributes to tied values in the permutation distribution of the W statistic. So too does the use of a small number of mid ranks (averages of tied ranks) when calculating the statistic from ordered categorical data. This discreteness in the null distribution results in a conservative and less powerful test [8]. Eliminating or reducing the number of tied outcomes could result in a more powerful test [13]. Several approaches have been proposed for assigning different values to some permutations that would, otherwise, have equal W statistics [3, 8, 11]. One of these assures that even the exact permutation version of the test becomes uniformly more powerful [3].

Asymptotic Properties

The **asymptotic relative efficiency** of the Wilcoxon-Mann-Whitney test, against the t Test, makes it a strong candidate for testing for differences in location. Where the distributions sampled are, in fact, normal, the Wilcoxon-Mann-Whitney test has an asymptotic relative efficiency of 0.955. In no case is the asymptotic relative efficiency of the Wilcoxon test lower than 0.864 [6]. And, the Wilcoxon test can offer much greater efficiency than the t Test for some types of distributions. The worst-case potential loss of efficiency of 13% – 0.136 = (1.000 – 0.864) – might be regarded as a relatively small insurance premium to be paid in case one of these distributions arises, and renders the t Test inefficient.

Example The following excerpt from [9] (see Table 1) shows the ranks of mens' and womens' scores after being tested regarding the strength of their endorsement of traditional sex ascriptions.

To test the null hypothesis that there is no difference in endorsement between genders against the alternative hypothesis that a difference exists, we compute the sums of the ranks for each gender.

To compute the large-sample approximation we note that

The sampling mean is $17^*(17 + 17 + 1)/2 = 297.5$.

The sampling variance is $17^*17^*(17 + 17 + 1)/12$, or a standard error of 29.033.

These result in a standardized test statistic of $W^* = (378 - 297.5)/29.033 = 2.77$.

We reject the null hypothesis at $\alpha = 0.05$ as $W^* \geq z_{\alpha/2} = 1.96$, and we see that a significant difference exists between the two groups. The P value of the normal approximation can be calculated as 0.0056.

References

[1] Berger, V.W. (2000). Pros and cons of permutation tests in clinical trials, *Statistics in Medicine* **19**, 1319–1328.

[2] Berger, V.W., Lunneborg, C., Ernst, M.D. & Levine, J.G. (2002). Parametric analyses in randomized clinical trials, *Journal of Modern Applied Statistical Methods* **1**(1), 74–82.

[3] Berger, V. & Sackrowitz, H. (1997). Improving tests for superior treatment in contingency tables, *Journal of the American Statistical Association* **92**(438), 700–705.

[4] Berger, V.W., Zhou, Y.Y., Ivanova, A. & Tremmel, L. (2004). Adjusting for ordinal covariates by inducing a partial ordering, *Biometrical Journal* **46**, 48–55.

[5] Bergmann, R., Ludbrook, J. & Spooren, W. (2000). Different outcomes of the Wilcoxon-Mann-Whitney test from different statistics packages, *The American Statistician* **54**, 72–77.

[6] Hodges, J.L. & Lehmann, E.L. (1956). The efficiency of some non-parametric competitors of the t-test, *Annals of Mathematical Statistics* **27**, 324–335.

[7] Hollander, M. & Wolfe, D.A. (1999). *Nonparametric Statistical Methods*, 2nd Edition, John Wiley, New York.

[8] Ivanova, A. & Berger, V.W. (2001). Drawbacks to integer scoring for ordered categorical data, *Biometrics* **57**, 567–570.

[9] Kando, T.M. (1972). Role strain: a comparison of males, females, and transsexuals, *Journal of Marriage and the Family* **34**, 459–464.

[10] Mann, H.B. & Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics* **18**, 50–60.

[11] Permutt, T. & Berger, V.W. (2000). A new look at rank tests in ordered 2xk contingency tables, *Communications in Statistics – Theory and Methods* **29**(5 and 6), 989–1003.

Table 1 Ranks of scores for men and women

	Ranks	Sum
Men	1, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16.5, 18, 19, 27.5, 34	217
Women	2, 3, 13, 16.5, 20.5, 20.5, 22, 23, 24, 25, 26, 27.5, 29, 30, 31, 32, 33	378

[12] Siegel, S. & Castellan, N.J. (1988). *Nonparametric Statistics for the Behavioral Sciences*, 2nd Edition, McGraw-Hill, New York.

[13] Streitberg, B. & Roehmel, J. (1990). On tests that are uniformly more powerful than the Wilcoxon-Mann-Whitney test, *Biometrics* **46**, 481–484.

[14] Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics* **1**, 80–83.

(*See also* **Hodges–Lehman Estimator**; **Kruskal–Wallis Test**; **Randomized Block Design: Nonparametric Analyses**; **Wilcoxon, Frank**)

Venita DePuy, Vance W. Berger and
YanYan Zhou

Winsorized Robust Measures

Rand R. Wilcox

Volume 4, pp. 2121–2122

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Winsorized Robust Measures

Consider any n observations and let g be $0.1n$, rounded down to the nearest integer. Then, 10% trimming refers to removing the g smallest, as well as the g largest values (*see* **Trimmed Means**). Winsorizing the values means that the g smallest values are reset to the smallest value not trimmed, and the g largest are set to the largest value not trimmed. As an illustration, consider the eleven values 6, 2, 10, 14, 9, 8, 22, 15, 13, 82, and 11. Then $g = 1$ and Winsorizing these values by 10% refers to replacing the smallest value, 2, with the next smallest, 6. Simultaneously, the largest value, 82, is replaced by the next largest, 22. So the 10% Winsorized values are 6, 6, 10, 14, 9, 8, 22, 15, 13, 22, and 11. The 20% Winsorized values are obtained in a similar fashion; only, now g is $0.2n$, rounded down to the nearest integer. The average of the Winsorized values is called a *Winsorized mean*, and the variance of the Winsorized values is called a *Winsorized variance*.

Winsorized means can be used to compare groups; under nonnormality, a Winsorized mean can have a substantially lower standard error than the usual sample mean, which can result in higher **power**. But, usually other robust estimators are used. Instead, Winsorization is typically used to obtain a theoretically correct estimate of the standard error of a trimmed mean, which has certain practical advantages over comparing groups with a Winsorized mean. For details, see [1–5]. Winsorization also plays a role when searching for robust alternatives to Pearson's correlation (*see* **Pearson Product Moment Correlation**) [4, 5]. The so-called Winsorized correlation guards against outliers among the marginal distributions, which can help detect associations that would be missed when using Pearson's correlation. A criticism, however, is that the Winsorized correlation does not take into account the overall structure of the data when dealing with **outliers**. For example, only two unusual values can mask an association that would be detected by other correlation coefficients [5].

References

[1] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. (1986). *Robust Statistics*, Wiley, New York.

[2] Huber, P.J. (1981). *Robust Statistics*, Wiley, New York.

[3] Staudte, R.C. & Sheather, S.J. (1990). *Robust Estimation and Testing*, Wiley, New York.

[4] Wilcox, R.R. (2003). *Applying Conventional Statistical Techniques*, Academic Press, San Diego.

[5] Wilcox, R.R. (2004). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd Edition, Academic Press, San Diego.

RAND R. WILCOX

Within Case Designs: Distribution Free Methods

LISA M. LIX AND H.J. KESELMAN

Volume 4, pp. 2122–2126

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Within Case Designs: Distribution Free Methods

In a within-case or repeated measures (RM) design (*see* **Repeated Measures Analysis of Variance**), subjects provide data at K successive points in time or for each of K experimental conditions. Data collected from different subjects are assumed to be independent, while data from the same subject are correlated. Tests of within-case main or **interaction effects** may be conducted using univariate or multivariate parametric or nonparametric procedures; the valid use of any one of these approaches depends on the data conforming to its underlying derivational assumptions.

The **analysis of variance (ANOVA)** F test, the usual parametric test for RM designs, rests on the assumption of **sphericity**,

$$\mathbf{C\Sigma C}' = \sigma^2\mathbf{I}, \qquad (1)$$

where $\mathbf{C}$, of dimension $(K-1) \times K$, defines a set of orthonormalized contrasts on the repeated measurements, $\mathbf{\Sigma}$ is the covariance matrix, and $\mathbf{I}$ is an identity matrix of dimension $(K-1)$. The multivariate parametric approach makes no assumptions about the structure of the covariance matrix of the repeated measurements. Both univariate and multivariate parametric procedures assume a K-variate normal distribution.

Type I **error rates** of parametric procedures for testing within-case effects are relatively robust (i.e., insensitive) to departures from normality, although skewed distributions may be associated with inflated error rates [12]. However, nonnormality can result in a substantial loss of statistical **power** to detect the presence of within-case effects [20]. When the data are highly skewed or have heavy tails due to the presence of outliers, nonparametric procedures, which make no assumptions regarding the distribution of the data, may result in more powerful tests of within-case effects. In this section, we select a number of different nonparametric procedures that may be applied to RM data, and illustrate their application.

Procedures Based on Rank Scores

Rank tests for within-case designs include procedures that are applied to inter-case ranks, where all NK scores are ranked without regard to subject membership. They also include procedures that are applied to intra-case ranks, where the scores for each of the N subjects are arranged in increasing order of magnitude, and then ranked from 1 to K. (*see* **Rank Based Inference**.)

For the simplest within-case design that contains a single group of subjects and a single within-case factor, let y_{ik} represent the score for the ith subject $(i = 1, \ldots, N)$ for the kth treatment or time period $(k = 1, \ldots, K)$, and let r_{ik} represent the intra-case rank of y_{ik}. Midranks are assigned for ties. **Friedman's test** [7], a well-known procedure that has been used to test within-case effects with intra-case ranks, is defined as

$$FR = \frac{K(K+1)}{12N} \sum_{k=1}^{K} \left(\bar{r}_{.k} - \frac{K+1}{2} \right)^2, \qquad (2)$$

where $\bar{r}_{.k}$ is the kth treatment mean rank. FR is asymptotically distributed as $\chi^2[\alpha; K-1]$ (see [10] for approximations to the distribution of FR).

Friedman's procedure [7] tests the exchangeability hypothesis, $G_i(\mathbf{y}) = G(y_1 \ldots y_K)$ for $\mathbf{y}_i = [y_{i1} \ldots y_{iK}]$ and any permutation of $[1 \ldots K]$, where $G_i(\mathbf{y})$ denotes the distribution function of $\mathbf{y}_i$. In other words, under the null hypothesis, the joint distribution of the observations is assumed to be invariant for any permutation of the intra-subject ranks. Accordingly, Friedman's test assumes a common correlation between pairs of observations [1]. Thus, while Friedman's procedure may be insensitive to the shape of the underlying distribution, it is known to be sensitive to departures from sphericity.

For designs that contain both within-case and between-case (i.e., grouping) factors, procedures based on intra-case ranks include extensions of Friedman's test [4], Hollander and Sethuraman's [8] two-group test, and extensions of Hollander and Sethuraman's test to multi-group designs [4, 17]. These procedures are used to test within-case interactions, which can be expressed as tests of discordance or inconsistency of intra-case ranks across independent groups of subjects.

Procedures based on inter-case ranks include rank transform tests, in which standard parametric tests, such as the analysis of variance (ANOVA) F test or a multivariate test, are applied to ranked data. The ANOVA F, which tests the exchangeability hypothesis, is given by

$$F_{\mathrm{RT}} = \frac{MS_K}{MS_{S\times K}} = \frac{\displaystyle\sum_{k=1}^{K} (\bar{r}^*_{.k} - \bar{r}^*_{..})^2}{\displaystyle\sum_{k=1}^{K}\sum_{i=1}^{N} (r^*_{ik} - \bar{r}^*_{i.} - \bar{r}^*_{.k} + \bar{r}^*_{..})^2}, \quad (3)$$

where $r^*_{ik}, \bar{r}^*_{i.}, \bar{r}^*_{.k}$, and $\bar{r}^*_{..}$ respectively represent the rank for the ith subject at the kth level of the within-case factor, the mean rank for the ith subject, the mean rank for the kth within-case factor level, and the grand mean rank. F_{RT} is approximately distributed as $F[\alpha;\ K-1, (N-1)(K-1)]$. For one-group within-case designs, the multivariate rank transform test is Hotelling's $\boldsymbol{T}^2$ [9], $T_{\mathrm{RT}} = N(\mathbf{C\bar{R}})'(\mathbf{CS_rC})^{-1}(\mathbf{C\bar{R}})$, where $\mathbf{C}$ defines a set of $(K-1)$ contrasts for the repeated measurements, $\mathbf{\bar{R}} = [\bar{r}^*_{.1} \ldots \bar{r}^*_{.K}]'$, and $\mathbf{S}_{\mathrm{r}}$ is the covariance matrix of the ranks. The statistic $F_T = [(N/(K-1)]T_{\mathrm{RT}}/(N-1)$ is approximately distributed as $F[\alpha;\ K-1, N-K+1]$ [1]. Hotelling's T^2 for rank transform data is used to test the hypothesis of equality of the marginal distributions of the repeated measurements, that is, $G_1(\mathbf{y}) = G_2(\mathbf{y}) = \cdots = G_K(\mathbf{y})$.

Rank transform tests are appealing to researchers because they can be easily applied with standard statistical software package. One limitation is that they cannot be applied to tests of within-case interactions. The ranks are not a linear function of the original observations, therefore, ranking the data may introduce additional effects into the statistical model. Moreover, ranking may alter the pattern of the correlations among repeated measurements. Accordingly, rank transform tests, while insensitive to departures from normality, must be used with caution in multifactor designs [1, 2, 3, 6, 18].

Nonparametric Procedures Based on Resampling

When the assumption of multivariate normality is in doubt (*see* **Multivariate Normality Tests**), within-case effects may be tested using statistical procedures based on the resampling technique of bootstrapping (*see* **Bootstrap Inference**; **Permutation Based Inference**) [13, 19]. Under this approach, the usual univariate or multivariate parametric test statistic is computed on the original data, but statistical significance of within-case effects is assessed using the empirical distribution of the test statistic rather than the theoretical distribution.

To illustrate, let F denote the conventional ANOVA F statistic for testing the within-case effect in a single group design. An iterative process is used to obtain the empirical distribution of the test statistic as follows: A bootstrap data set is generated by randomly sampling with replacement the K-variate vectors of repeated measurements. Let $\mathbf{y}^*_i$ represent the ith resampled vector. Each $\mathbf{y}^*_i$ is centered on the sample mean vector, $\mathbf{\bar{y}} = [\bar{y}_{.1} \ldots \bar{y}_{.K}]^T$, where $\bar{y}_{.k} = \sum_{i=1}^{n} y_{ik}$, so that $\mathbf{y}^{*\mathrm{C}}_{\mathrm{i}} = \mathbf{y}^*_{\mathrm{i}} - \mathbf{\bar{y}}$. The test statistic, F^*, is computed on the centered bootstrapped data set. This process is repeated B times. Let $F^*_{(1)} \le F^*_{(2)} \le \cdots \le F^*_{(B)}$ denote the B bootstrapped test statistics arranged in ascending order, and let $m = (1-\alpha)B$. Then, F is referred to the critical value $\mathrm{F}^*_{(m)}$. The bootstrapped ANOVA F Test will control the rate of Type I errors to α under departures from both normality and sphericity [5]. The bootstrapped Hotelling's T^2 also performs well under departures from normality.

A Numeric Example

To illustrate these various nonparametric tests based on ranks and the bootstrap, we selected a data set ([14], p. 571) for an experiment in which the length of gaze (in seconds) at a stimuli was obtained for each of 14 infants (see Table 1). Four different stimuli were considered: face, concentric circles, newspaper, and unpatterned white circle. We modified the original data by adding a constant to each of the measurements for the first two subjects in order to produce a skewed distribution. The inter-case ranks for the modified data set are in Table 2. Table 3 contains $\mathbf{S}$ and $\mathbf{S}_{\mathrm{r}}$, the covariance matrix of the raw scores and the ranks, respectively. Both covariance matrices reveal the presence of increasing heterogeneity in the data across the four stimuli. Friedman's test gives $FR = 6.8$ with a P value of $p_{FR} = .08$. Applying the rank transform Hotelling's T^2 to the data gives $T_{\mathrm{RT}} = 17.0$ with $p_{\mathrm{RT}} < .0001$. The bootstrap P value for these

Table 1 Raw scores for one-group within-case design

	Stimuli			
Infant	Face	Circle	Newspaper	White
1	6.1	6.4	6.7	6.8
2	6.3	6.6	6.7	6.5
3	2.1	1.7	1.2	0.7
4	1.5	0.9	0.6	0.4
5	0.9	0.6	0.9	0.8
6	1.6	1.8	0.6	0.8
7	1.8	1.4	0.8	0.6
8	1.4	1.2	0.7	0.5
9	2.7	2.3	1.2	1.1
10	1.5	1.2	0.7	0.6
11	1.4	0.9	1.0	0.5
12	1.6	1.5	0.9	1.0
13	1.3	1.5	1.4	1.6
14	1.3	0.9	1.2	1.4

Table 2 Inter-case ranks for one-group within-case design

	Stimuli			
Infant	Face	Circle	Newspaper	White
1	49	51	54.5	56
2	50	53	54.5	52
3	46	43	26	10
4	37.5	17.5	6	1
5	17.5	6	17.5	13
6	41	44.5	6	13
7	44.5	33	13	6
8	33	26	10	2.5
9	48	47	26	23
10	37.5	26	10	6
11	33	17.5	21.5	2.5
12	41	37.5	17.5	21.5
13	29.5	37.5	33	41
14	29.5	17.5	26	33

Table 3 Variance-covariance matrix for raw scores and inter-case ranks

Raw scores			
2.98	3.30	3.52	3.53
	3.72	3.96	4.01
		4.44	4.49
			4.58
Inter-case ranks			
84.8	113.7	56.6	51.4
	210.7	128.0	155.7
		245.4	262.9
			346.5

data, p_B, which was based on 1000 replications (see [19]), is also <.0001. Applying the ANOVA F test to the original observations gives $F = 8.0$ with $p = .0003$. The bootstrap P value for this test is also <.0001.

Concluding Remarks

Behavioral scientists may be reluctant to bypass conventional parametric approaches for the analysis of within-case effects in favor of nonparametric tests based on ranking or resampling methods. This reluctance may stem, in part, from the belief that parametric procedures are robust to departures from normality. While Type I error rates of parametric procedures may be relatively robust to the presence of nonnormal distributions, power rates can be substantially affected, particularly when the data are skewed. Researchers may also be reluctant to adopt nonparametric procedures because they are unfamiliar with test for multi-factor and multivariate designs, or with methods for testing linear contrasts on the within-case effects. Recent research has focused on the development of nonparametric tests for a variety of univariate and multivariate repeated measures designs [15, 16]. Procedures based on the bootstrap can be readily applied to a variety of complex univariate and multivariate designs to test hypotheses on omnibus effects as well as linear contrasts of within-case effects [19].

Finally, we note that alternative parametric procedures have been proposed for testing within-case effects when the data are nonnormal. For designs that contain both within-case and between-case factors, Keselman, Kowalchuk, Algina, Lix, and Wilcox [11] examined approximate degrees of freedom procedures that assume neither equality (i.e., homogeneity) of group covariances nor sphericity of the common covariance of the repeated measurements. These procedures were extended to the case of nonnormality by substituting the usual (i.e., least-squares) estimators with robust estimators based on **trimmed means**. Trimmed means are obtained by removing the most extreme observations from the tails of the data distribution prior to computing the average score. These approximate degrees of freedom tests based on trimmed estimators were shown to be insensitive to the presence of both skewed and heavy-tailed distributions. The tests were also examined when critical values were generated via the bootstrap. As expected, the bootstrapped tests were also robust to nonnormality, although the Type I error rates of the two approaches were not appreciably different.

References

[1] Agresti, A. & Pendergast, J. (1986). Computing mean ranks for repeated measures data, *Communications in Statistics, Theory and Methods* **15**, 1417–1433.

[2] Akritas, M.G. (1991). Limitations of the rank transform procedure: A study of repeated measures designs, part I, *Journal of the American Statistical Association* **86**, 457–460.

[3] Akritas, M.G. (1993). Limitations of the rank transform procedure: a study of repeated measures designs, part II, *Statistics & Probability Letters* **17**, 149–156.

[4] Beasley, T.M. (2000). Non-parametric tests for analyzing interactions among intra-block ranks in multiple group repeated measures designs, *Journal of Educational and Behavioral Statistics* **25**, 20–59.

[5] Berkovits, I., Hancock, G.R. & Nevitt, J. (2000). Bootstrapping resampling approaches for repeated measure designs: Relative robustness to sphericity and normality violations, *Educational and Psychological Measurement* **60**, 877–892.

[6] Blair, R.C., Sawilowsky, S.S. & Higgins, J.J. (1987). Limitations of the rank transform statistic in test for interactions, *Communications in Statistics: Simulation and Computation* **16**, 1133–1145.

[7] Friedman, M. (1937). The used of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the American Statistical Association* **32**, 675–701.

[8] Hollander, M. & Sethuraman, J. (1978). Testing for agreement between two groups of judges, *Biometrika* **65**, 403–411.

[9] Hotelling, H. (1931). The generalization of student's ratio, *Annals of Mathematical Statistics* **2**, 360–378.

[10] Inman, R.L. & Davenpot, J.M. (1980). Approximations of the critical region of the Friedman statistic, *Communications in Statistics, Theory and Methods* **A9**, 571–595.

[11] Keselman, H.J., Kowalchuk, R.K., Algina, J., Lix, L.M. & Wilcox, R.R. (2000). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping, *British Journal of Mathematical and Statistical Psychology* **53**, 175–191.

[12] Keselman, J.C., Lix, L.M. & Keselman, H.J. (1996). The analysis of repeated measurements: A quantitative research synthesis, *British Journal of Mathematical and Statistical Psychology* **49**, 275–298.

[13] Lunnenborg, C.E. & Tousignant, J.P. (1985). Efron's bootstrapping with application to the repeated measures design, *Multivariate Behavioral Research* **20**, 161–178.

[14] Maxwell, S.E. & Delaney, H.D. (2004). *Designing Experiments and Analyzing data: A Model Comparison Perspective*, 2nd Edition, Lawrence Erlbaum, Mahwah.

[15] Munzel, U. & Brunner, E. (2000). Nonparametric methods in multivariate factorial designs, *Journal of Statistical Planning and Inference* **88**, 117–132.

[16] Rahman, M.M. & Islam, M.N. (1998). Nonparametric analysis of multifactor repeated measures experiments, *Journal of the Indian Society of Agricultural Statistics* **51**, 81–93.

[17] Rasmussen, J.L. (1989). Parametric and non-parametric analysis of groups by trials design under variance-covariance inhomogeneity, *British Journal of Mathematical and Statistical Psychology* **42**, 91–102.

[18] Sawilosky, S.S., Blair, R.C. & Higgins, J.J. (1989). An investigation of the Type I error and power properties of the rank transform procedure in factorial ANOVA, *Journal of Educational Statistics* **14**, 255–267.

[19] Wasserman, S. & Bockenholt, U. (1989). Bootstrapping: Applications to psychophysiology, *Psychophysiology* **26**, 208–221.

[20] Wilcox, R.R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size?, *Review of Educational Research* **65**, 51–77.

LISA M. LIX AND H.J. KESELMAN

Yates' Correction

Catalina Stefanescu, Vance W. Berger and Scott L. Hershberger

Volume 4, pp. 2127–2129

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Yates' Correction

Yates' correction [15] is used as an approximation in the analysis of 2×1 and 2×2 **contingency tables**. A 2×2 contingency table shows the frequencies of occurrence of all combinations of the levels of two dichotomous variables, in a sample of size N. A schematic form of such a table is shown in Table 1.

A research question of interest is often whether the variables summarized in a contingency table are independent of each other. The test to determine if this is so depends on which, if any, of the margins are fixed, either by design or for the purposes of the analysis. For example, in a randomized trial in which the number of subjects to be randomized to each treatment group has been specified, the row margins would be fixed but the column margins would not (it is customary to use rows for treatments and columns for outcomes). In a matched study, however, in which one might sample 100 cases (smokers, say) and 1000 controls (non–smokers), and then test each of these 1100 subjects for the presence or absence of some exposure that may have predicted their own smoking status (perhaps a parent who smoked), it would be the column margins that are fixed. In a random and unstratified sample, in which each subject sampled is then cross–classified by two attributes (say smoking status and gender), neither margin would be fixed. Finally, in **Fisher's** famous tea–tasting experiment [13], in which a lady was to guess whether the milk or the tea infusion was first added to the cup by dividing eight cups into two sets of four, both the row and the column margins would be fixed by the design. Yet, in the first case mentioned, that of a randomized trial with fixed row margins but not fixed column margins, the column margins may be treated as fixed for the purposes of the analysis, so as to ensure exactness [2].

When the row and column margins are fixed, either by design or for the analysis, independence can be tested using Fisher's exact test [4] (*see* **Exact Methods for Categorical Data**). This test is based on the hypergeometric distribution (*see* **Catalogue of Probability Density Functions**), and it is computationally intensive, especially in large samples. Therefore, Fisher advocated the use of Pearson's statistic,

$$X^2 = \frac{N(AD - BC)^2}{(A+B)(C+D)(A+C)(B+D)}, \tag{1}$$

which, under the null hypothesis, has a χ^2 distribution with one degree of freedom. Yates [15] argued that the χ_1^2 distribution gives only approximate estimates of the discrete probabilities associated with frequency data, and, thus, the P values based on Pearson's X^2 statistic will generally underestimate the true P values. In general, when a statistic takes discrete values $a < b < c$, the P value corresponding to b is estimated by the tail of the continuous function defined by the point $(a + b)/2$. Therefore, the tail of the continuous function computed at b will underestimate the P value. In this context, Yates suggested that X^2 should be corrected for continuity and proposed the corrected test statistic

$$\frac{N\left(|AD - BC| - \frac{1}{2}N\right)^2}{(A+B)(C+D)(A+C)(B+D)}.$$

Although Yates' correction is best known for its use in the analysis of 2×2 contingency tables, it is also applicable to the analysis of 2×1 contingency tables. A 2×1 contingency table displays the frequencies of occurrence of two categories in a random sample of size N, drawn from a population in which the proportions of cases within the two categories are p and $1 - p$. The research question is usually whether the observed numbers of cases x and $N - x$ in the two categories have been sampled from a population with some prespecified value of p. This can be tested using Pearson's statistic,

$$X^2 = \frac{(x - Np)^2}{Np(1-p)}, \tag{2}$$

which asymptotically has a χ_1^2 distribution under the null hypothesis. Yates showed that, in this case as well, the use of Pearson's X^2 results in P values that systematically underestimate the true P values based on the binomial distribution. Therefore, he suggested the corrected statistic

$$\frac{\left(|x - Np| - \frac{1}{2}\right)^2}{Np(1-p)}. \tag{3}$$

Table 1 A 2×2 contingency table

	Column variable		
Row variable	1	2	Totals
1	A	B	$A + B$
2	C	D	$C + D$
Totals	$A + C$	$B + D$	N

Kendall and Stuart [7] remarked that Yates' procedure is a special case of a general concept of a continuity correction, while **Pearson** [10] noted that Yates' correction derives naturally from the Euler-Maclaurin theorem used to approximate binomial and hypergeometric distributions. Subsequently, the use of Yates' correction to Pearson's X^2 has been widely emphasized for the analysis of contingency tables [14]. There are, however, several issues related to Yates' correction, and we shall discuss some of these in turn.

Firstly, in the analysis of 2×1 contingency tables, the P values associated with the corrected statistic (3) tend to overestimate the true P values in the tails of the distribution and to underestimate them towards the center. This is illustrated in Table 2, which displays the two-tailed P values in a contingency table with $N = 10$ and $p = 0.5$, obtained with Pearson's X^2 statistic and Yates' correction. The table reports as well the true binomial P values, which are the gold standard. It should also be noted [15] that the P values obtained with the continuity correction are much less accurate when the binomial probability p is substantially different from 0.5.

Secondly, Yates' correction is appropriate only for one-sided tests, as it is based on a comparison between the observed contingency and the next strongest contingency in the same direction ([6, 8]). For two-sided tests, the statistic involves an overcorrection. Along the same lines, it can be proven analytically that Yates' correction is systematically conservative when carrying out two-sided tests [9].

Thirdly, a more important issue related to Yates' correction is its applicability to the analysis of contingency tables arising from different research designs. Many researchers have argued that Yates's correction is based upon comparisons among contingency tables with fixed row and column marginal totals, particularly since Yates is specifically concerned with approximating the hypergeometric distribution from Fisher's exact test. However, Yates' method has also been recommended for the analysis of 2×2 contingency tables arising from sampling schemes, where one or both sets of marginal totals are free to vary, and are, thus, subject to sampling errors. It should be noted that such sampling schemes are the ones most frequently found in actual research context. While Yates [16] argues along the lines of Fisher's reasoning that the analysis of 2×2 contingency tables should always be performed conditional on the observed marginal totals, this approach is still subject to debate [12]. On the other hand, when the marginal totals are not fixed, Yates' procedure involves an additional overcorrection, and the test statistic is conservative. This has been investigated through Monte Carlo simulations ([5, 11]), and confirmed analytically ([3, 6]). In particular, Grizzle [5] notes that for contingency tables with nonfixed marginal totals, Yates's procedure 'produces a test that is so conservative as to be almost useless'.

Table 2 Binomial distribution for $N = 10$ and $p = 0.5$, and two-tailed P values. (Adapted from [12])

		P values		
x	$p(x)$	Pearson	Yates	Binomial
0, 10	0.0010	0.0016	0.0044	0.0020
1, 9	0.0098	0.0114	0.0268	0.0215
2, 8	0.0439	0.0580	0.1138	0.1094
3, 7	0.1172	0.2060	0.3428	0.3437
4, 6	0.2051	0.5270	0.7518	0.7539
5, 5	0.2461	1.0000	1.0000	1.0000

Finally, Yates's correction originated as a device of eliminating the discrepancies that arose when approximating the hypergeometric distribution in Fisher's exact test. The approximation using Pearson's X^2 was necessary 'for the comparative simplicity of the calculations' ([4], p. 99), because the exact analysis of 2×2 contingency tables with the limited computing power available at the time was prohibitive in many cases. This is no longer the case today. Indeed, Agresti [1] notes that Yates' correction is not necessary anymore since current software makes Fisher's exact test computationally feasible even when the sample sizes are large.

References

[1] Agresti, A. (2002). *Categorical Data Analysis*, Wiley, New York.

[2] Berger, V.W. (2000). Pros and cons of permutation tests in clinical trials, *Statistics in Medicine* **19**, 1319–1328.

[3] Conover, W.J. (1974). Some reasons for not using the Yates continuity correction on 2×2 contingency tables, *Journal of the American Statistical Association* **69**, 374–376.

[4] Fisher, R.A. (1934). *Statistical Methods for Research Workers*, 5th Edition, Oliver and Boyd, Edinburgh.

[5] Grizzle, J.E. (1967). Continuity correction in the χ^2 test for 2×2 tables, *The American Statistician* **21**, 28–32.

[6] Haber, M. (1982). The continuity correction and statistical testing, *International Statistical Review* **50**, 135–144.

[7] Kendall, M.G. & Stuart, A. (1967). *The Advanced Theory of Statistics*, Vol. 2, 2nd Edition, Griffin, London.

[8] Mantel, N. (1976). The continuity correction, *The American Statistician* **30**, 103–104.

[9] Maxwell, E.A. (1976). Analysis of contingency tables and further reasons for not using Yates correction in 2 × 2 tables, *Canadian Journal of Statistics* **4**, 277–290.

[10] Pearson, E.S. (1947). The choice of statistical tests illustrated on the interpretation of data classed in a 2 × 2 table, *Biometrika* **34**, 139–167.

[11] Richardson, J.T.E. (1990). Variants of chi-square for 2 × 2 contingency tables, *British Journal of Mathematical and Statistical Psychology* **43**, 309–326.

[12] Richardson, J.T.E. (1994). The analysis of 2 × 1 and 2 × 2 contingency tables: an historical review, *Statistical Methods in Medical Research* **3**, 107–133.

[13] Salsburg, D. (2002). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*, Owl Books, London.

[14] Siegel, S. & Castellan, N.J. (1988). *Nonparametric Statistics for the Behavioural Sciences*, 2nd Edition, McGraw-Hill, New York.

[15] Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test, *Journal of the Royal Statistical Society* **1**(Suppl. 1), 217–235.

[16] Yates, F. (1984). Tests of significance for 2 × 2 contingency tables, *Journal of the Royal Statistical Society Series A* **147**, 426–463.

Further Reading

Haber, M. (1980). A comparison of some continuity corrections for the chi-squared test on 2 × 2 tables, *Journal of the American Statistical Association* **75**, 510–515.

Plackett, R.L. (1964). The continuity correction in 2 × 2 tables, *Biometrika* **51**, 327–337.

Catalina Stefanescu, Vance W. Berger
and Scott L. Hershberger

Yates, Frank

Brian S. Everitt

Volume 4, pp. 2129–2130

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Yates, Frank

Born: 12 May, 1902 in Manchester, England.
Died: 17 June, 1994 in Harpenden, England.

Born in Manchester, Yates read mathematics at St. John's College, Cambridge and received a first-class honors degree in 1924. After a period working in the Gold Coast (now Ghana) on a geodetic survey as a mathematical advisor, he obtained a post as assistant statistician at Rothamsted Experimental Station in 1931, where he worked under **R.A. Fisher**. Two years later when Fisher left to take up a chair at University College, London, Yates became Head of Statistics at Rothamsted, a post he held until his retirement in 1968.

Yates continued the work on design of experiment, replication, **randomization** and blocking (*see* **Block Random Assignment**) (to reduce error), topics introduced to Rothamsted by Fisher (see [2] for a selection of work in this area). All these ideas were originally applied to agriculture but spread rapidly to many other disciplines. Yates extended and clarified the ideas of orthogonality, confounding, and balance, and suggested the use of split-plot designs. During World War II, Yates studied food supplies and applications of fertilizers to improve crops, and applied experimental design techniques to a wide range of problems such as control of pests. But despite all his important contributions to designing studies, Yates is most widely remembered for his continuity correction in **contingency tables** (**Yates' correction**); ironically this correction has been made almost obsolete by the development of software for applying exact methods [1].

Yates was quick to realize the possibilities for statistics and statisticians provided by the development of electronic computers in the 1950s. And in 1954, the first British computer equipped with effective magnetic storage, the Elliot 401, was installed at Rothamsted. Using only machine code, Yates and other members of the statistics department produced programs both for the **analysis of variance** and to analyze survey data. Yates helped establish the British Computer Society, of which he was made President in 1960–1961. In 1948, Yates was made a Fellow of the Royal Society, and in 1960 he was awarded the Royal Statistical Society's Guy Medal in Gold. In 1963, he was awarded the CBE.

Despite retiring from Rothamsted in 1968, Yates kept a room there and never lost touch with agriculture. Just before his death in Harpenden in 1994, Yates completed, in 1993, 60 years of work at Rothamsted.

References

[1] Mehta, C.R. & Patel, N.R. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables, *Journal of the American Statistical Society* **78**, 427–434.

[2] Yates, F. (1970). *Experimental Design: Selected Papers of Frank Yates*, Griffin Publishing, London.

BRIAN S. EVERITT

Yule, George Udny

Michael Cowles

Volume 4, pp. 2130–2131

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Yule, George Udny

Born: February 18, 1871, in Morham, Scotland.

Died: June 26, 1951, in Cambridge, England.

Scottish by birth, Yule was educated in England and spent all his working life there. After school, the young Yule studied engineering at University College, London. Engineering and physics never captured his interest though he was greatly influenced by his Professor of Applied Mathematics, **Karl Pearson**. He attended Pearson's Gresham Lectures in 1891 and corresponded with him. He spent a year in Bonn studying and researching under Hertz. His work there produced his first published papers. On his return to England, Pearson offered him the post of Assistant Professor. Though poorly paid, it provided Yule with the impetus and experience that led to his adopting the study and teaching of statistics as his lifelong career, apart from a spell as a civil servant during World War I.

Yule was Pearson's workhorse and he often lectured on his behalf when Pearson was indisposed. Their personal interactions were at that time cordial. They worked closely and took holidays together. The exigencies of his financial situation forced Yule to apply for other employment and in 1899 he became Secretary to the Examining Board of the Department of Technology of the London City and Guilds Institute. He married in the same year, a union that proved to be unhappy and led to a separation. Yule's tenure of the Newmarch Lectureship in Statistics at University College from 1902 to 1909 produced the material that became his famous textbook, *An Introduction to the Theory of Statistics*, the earliest of the major texts, running to 14 editions during Yule's lifetime, the final 4 editions being written in collaboration with **Maurice Kendall** [3].

In 1912, the University of Cambridge offered Yule a newly established lectureship in statistics and later he became Reader. He took up residence at St John's College for the rest of his life until his health forced him into a nursing home.

Yule's contributions to statistics consist largely of his clarification, interpretation, and expansion of Pearson's work and the laying down of the groundwork for the future contributions of others. He demonstrated the least squares approach to the method of correlation and regression and showed that Pearson's coefficient and the regression model can be derived without assuming a bivariate normal distribution of the variables of interest. This not only indicated that r could be used as a descriptive statistic but gave us a readily understandable theoretical approach that led to the same outcome as Pearson's original **maximum likelihood method** [1]. He worked on measures of association in 2 × 2 tables and devised mathematical procedures that reduced the sometimes dense Pearsonian algebra of partial correlation [2]. His approaches became accepted by those who were interested in the theoretical bases of the methods.

Yule was elected a Fellow of the Royal Statistical Society in 1895 and remained so for almost 60 years. He was its Honorary Secretary for 12 years and the Society awarded him its highest honor, the Guy Medal in gold. He became a Fellow of the Royal Society in 1922.

In later life, he turned his considerable powers to the study of literary vocabulary, showing how statistical techniques might be used to determine authorship and to compare authors.

Yule was a determined, cheerful, and kindly man, articulate and well-read. He had a gift, as his correspondence with his great friend Major Greenwood shows, for irony and parody. In later years he regretted the transition from what he termed '*Karlovingian*' (Pearson) statistics to the 'piscatorial' approach of Fisher and felt that he himself had little more to offer, though he was gratified to see his own contributions recognized. Even today, there are teachers and practitioners of statistics who use his text and benefit from it.

References

[1] Yule, G.U. (1897). On the theory of correlation, *Journal of the Royal Statistical Society* **60**, 812–854.

[2] Yule, G.U. (1900). On the association of attributes in statistics, *Philosophical Transactions, A* **194**, 257–319.

[3] Yule, G.U. (1911 to 1973. Later editions with Maurice Kendall). *An Introduction to the Theory of Statistics*, Griffin Publishing, London.

MICHAEL COWLES

z Scores

David Clark-Carter

Volume 4, pp. 2131–2132

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

z Scores

A z-score is a form of standardized score. That is, it is a linear transformation of a raw score using the mean and the standard deviation (SD) of the sample or, if known, the population mean and SD. It is defined as

$$z = \frac{\text{score} - \text{mean}}{\text{standard deviation}}. \qquad (1)$$

A set of scores that has been transformed to z-scores has a mean of 0 and an SD of 1. Note that a sample statistic, such as a mean, can also be transformed to a z-score using the mean and SD of the sampling distribution of the statistic.

Suppose that a sample of students had mathematics test scores with a mean of 100 and a SD of 10, then a person whose mathematics score was 120 would have a z-score of 2. Here, we would say that this person's score was two SDs above the mean. The advantage of standardizing scores is that it allows comparison across different tests. For example, if the same sample of students had also taken a language test with a mean of 50 and an SD of 4, and this same person had scored 51 on the test, then the standardized score would be 0.25. This tells us that, whereas the student was two SDs above the mean in mathematics, he or she was only quarter of an SD above the mean for language.

Furthermore, when a sample comes from a normal distribution with a known mean and SD, we can use standard normal tables to find the **percentile** point for a given score. Our hypothetical person is in the top 2.28% for mathematics but only in the top 40.1% for language. The z-statistic employed in common hypothesis tests, for example, about means from normal populations is based on the z-score transformation.

z-scores can also be used to identify scores that could be **outliers**. An often quoted value for what might constitute an outlier is when the absolute (unsigned) value of the z-score is greater than or equal to 3 [1].

Reference

[1] Lovie, P. (1986). Identifying outliers, in *New Developments in Statistics for Psychology and the Social Sciences*, A.D. Lovie, ed., The British Psychological Society and Methuen, London, (pp 44–69).

David Clark-Carter