SPRINGER BRIEFS IN STATISTICS

David J. Bartholomew

Unobserved Variables

Models and Misunderstandings

Springer

SpringerBriefs in Statistics

For further volumes:
http://www.springer.com/series/8921

David J. Bartholomew

Unobserved Variables

Models and Misunderstandings

Springer

David J. Bartholomew
London School of Economics
London
UK

ISSN 2191-544X ISSN 2191-5458 (electronic)
ISBN 978-3-642-39911-4 ISBN 978-3-642-39912-1 (eBook)
DOI 10.1007/978-3-642-39912-1
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013944889

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Chapter 1
Unobserved Variables

Abstract Although unobserved variables go under many names there is a common structure underlying the problems in which they occur. The purpose of this Brief is to lay bare that structure and to show that the adoption of a common viewpoint unifies and simplifies the presentation. Thus, we may acquire an understanding of many disparate problems within a common framework. The case of missing observations in a sample is, perhaps, the most obvious example, but the field of latent variables provides a wider field which also draws attention to the fact that unobserved variables may be hypothetical as well as 'real'. Other fields, like time series analysis, also fit into this framework even though the connection may not be immediately obvious. The use of these methods has given rise to many misunderstandings which, we shall argue, often arise because the need for a statistical, or probability, model is unrecognised or disregarded. A statistical model is the bridge between intuition and the analysis of data.

Keywords Categorical variables · Factor analysis · Latent variables · Measurement · Missing values · Mixtures · Prediction · Time series

1.1 Background

Unobserved variables are a characteristic of many statistical problems but the links between them are often obscured by, both terminology and notation. In sample surveys they may, for example, be missing from the sample because respondents refuse to respond or are unobtainable or their responses may be lost. In time series they may be unobservable because they lie in the future. In some applications, factor analysis, for example, they are purely hypothetical and cannot therefore unobservable, even in principle. Terminology likewise reflects their diverse origins and includes, for example the adjectives: hidden, latent, and missing. Notation also tends to be peculiar to particular applications, not to mention the disciplinary

D. J. Bartholomew, *Unobserved Variables*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-39912-1_1, © The Author(s) 2013

allegiances of their originators. As well as giving rise to duplication of research effort, these features have tended to foster and perpetuate misunderstandings.

This Brief aims to lay bare the common underlying structure of some of the problems involving unobserved variables and so to simplify our understanding of them. Its purpose is not, primarily, to provide new statistical methods but to give greater insight into the common characteristics of many problems hitherto regarded as distinct and to reduce the misunderstandings which have arisen.

A typical problem in statistical inference may be expressed as follows. We have a sample (often random) drawn from a population of known form which depends on a set of parameters which we denote, collectively, by the vector $\boldsymbol{\theta}$ of dimension k. We suppose that the observed variables $\boldsymbol{x}' = (x_1, x_2, \ldots x_n)$, are given and they are, of course, natural numbers obtained by some measurement process. The aim is to make some inference about $\boldsymbol{\theta}$ on the evidence of $\boldsymbol{x}$.

In the class of problems which we are discussing here this specification may be supplemented by a further set of unobserved variables denoted by $\boldsymbol{y}' = (y_1, y_2 \ldots, y_m)$. If, of course, the ys were to constitute a further independent random sample, there is no problem because we can simply ignore them. If on the other hand, they were a part of an original random sample of size $n + m$ we have a standard case of inference with missing observations. Any link between the xs and ys can be exploited by the methods described, for example, in Little and Rubin (2002) (see Chap. 11). However, most of the problems we shall meet are a little more subtle because the term 'missing' has many connotations.

1.2 Models: Parameters and Random Variables

The key idea, lying behind everything we shall do is a probability model. This specifies the joint distribution of the xs and, where appropriate, the ys. It is clear that the xs should be treated as random variables and the elements of $\boldsymbol{\theta}$ as parameters as in any standard inference problem. However, the role of the ys in any model is crucial. According to context they may be treated as, either, random variables or parameters. The failure to make this distinction has led to much confusion, especially in the field of educational testing. It also lies, as we shall see later, behind the appropriateness of the Rasch model and it is crucial to the resolution of the so-called factor scores problem in factor analysis.

1.3 Continuous and Categorical Variables

Conceptually there is very little difference between continuous and categorical variables. But in practice they involve mathematical and arithmetical operations which look very different. For this reason, methods appropriate to the two types of variable have tended to develop separately, with categorical methods generally

lagging behind their continuous counterparts. This difference has been particularly striking in the field of latent variable modelling where the essential unity of the many diverse methods has been overlooked and has led to the creation of apparently watertight compartments for different methods. This matter was highlighted and largely remedied in the unified treatment given in the recent book by Bartholomew et al. (2011). In this Brief we shall continue in that tradition by treating continuous and categorical variables within a common framework. At first sight this makes for a greater degree of abstraction, and hence difficulty, but first impressions are deceptive. What appears to be lost through abstraction is more than regained by the conceptual simplicity which results.

Wherever possible we shall ignore the difference between continuous and categorical variables in our notation and terminology. For the most part we shall use the terminology and notation normally reserved for continuous variables using, for example, integrals rather than sums. We shall use the term 'probability function' to refer both to the constituent terms of a discrete probability distribution and to the probability density function of a continuous random variable. Where ranges covered by variables, discrete or continuous, are self- evident, they will usually be omitted. This is to emphasise our overall aim of displaying the structure of the problems as clearly as possible. The style adopted is more akin to that of the lecture than to a treatise and mathematical rigour has been sacrificed to facilitate understanding.

1.4 Particular Cases

We shall consider seven problems which fall within the general area we have defined although, at first sight, the selection may appear somewhat eclectic. They are listed below along with the chapter number to indicate where they may be found. The seven topics are neither exhaustive nor shall we treat each in its most general form. But they share a common structure and illustrate the range of very familiar problems which can be viewed from the present standpoint. Starting with mixtures of distributions we move on to consider latent variable models but we approach them by following an unfamiliar route via the data matrix and the analysis of variance. This leads on to a number of individual topics which concern unobserved variables in a variety of senses but which are connected by their shared structure. Some chapters do not deal with specific topics but serve a linking role. By approaching the various topics from a common standpoint, many of the common misunderstandings will dissolve, or never appear. Finally, we consider Bayesian models which, in a sense, brings all of statistical analysis within a common framework. The seven topics are:

Simple Mixtures (see Chap. 3)

A simple mixture problem may be expressed in the preceding framework as follows. The observed and unobserved variables may be bracketed in pairs, thus:

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n).$$

The first term in each bracket is the observed value and the second, y, indexes the 'mixing' variable. This term requires further explanation. If the ys were all equal they would be redundant and the xs would all be sampled from the same distribution. However, the mixing model supposes that each x has been sampled from a different population indexed by y. One of the earliest practical examples to be considered relates to the distribution of accidents. If a set of persons were exposed to the same degree of risk one might expect the distribution of the number of accidents per unit time to follow a Poisson distribution. In practice the actual distribution often displays greater dispersion than the Poisson distribution predicts. One way of explaining this is to suppose that the degree of risk ('proneness') varies among people. If the proneness can be characterised by a varying quantity, this may be regarded as a random variable and the situation is then as described above. This is an example of what is sometimes called 'unobserved heterogeneity'. The distribution of y may be continuous, as in this example, but it may be discrete, taking on, perhaps, as few as two possible values.

The One-way Analysis of Variance (see Chap. 4)

The situation may be set out as in the following array

$$\begin{array}{cccc} y_1 & y_2 & \ldots & y_m \\ x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{array}$$

As before, there are m groups with n observations in each group. (The numbers in each group need not be equal, of course, but the point we wish to make does not require that degree of generality.) In the usual set-up the ys represent the unknown group means and the aim is to estimate those means or to test the hypothesis that they are equal. In the fixed effects version of the analysis of variance, the ys are treated as parameters and we require any inference to apply to that particular set of means. In the random effects version of the problem, the m groups are selected at random from some population of categories so that the ys are random variables. The usual purpose of the analysis is then to estimate the variance of the ys or to test that it is zero.

Time Series Prediction (see Chap. 9)

Initially, this problem appears to be quite different from the foregoing. Here we have a sequence of xs, x_1, x_2, x_3, x_4,…x_n observed at successive discrete points in time where time n represents the 'present'. The next and subsequent members of the sequence are unobserved because they lie in the 'future'. They are denoted by y_1, y_2,… as far as necessary. The xs do not constitute a random sample from any population in this case because if they did there would obviously be no possibility of predicting future members of the series. A common purpose of time series

analysis is to predict future members of the series given those which have already occurred. Traditional methods have typically approached the problem as one of curve-fitting or of modelling patterns of change in the short term. The more general approach followed here does not add to the armoury of techniques for time series analysis but it shows, rather, that all methods flow from the same distributional foundation.

Models for Human Ability (see Chaps. 4 and 5)

These models are often treated under the general heading of 'Latent Variable Models' but they have distinctive features which justify them being dealt with separately. They have a wider field of application, of course, but here we introduce them as they arise in the context of educational testing. Each individual in a random sample of subjects from some population is supposed to possess an ability which can be located on a scale. In particular cases this might be designated as 'arithmetical ability', 'verbal ability' or, even, general intelligence. Several test items are administered to each individual and the resulting score for that ability is recorded. The result is an array of data which may be set out as follows.

$$\begin{array}{ccccc} x_{11} & x_{12} & \ldots & x_{1m} & y_1 \\ x_{21} & x_{22} & \ldots & x_{2m} & y_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{nm} & y_n \end{array}$$

The score given to the ith individual on the jth test is x_{ij}. The ys, have now been placed in the last column, where they represent the abilities of the individuals. Whether or not the ys are to be regarded as parameters or random variables turns out to be a key question to which we shall return. The essential point is that they are unobservable either in practice or theory: that is they are genuinely latent. A point to which we shall return, concerns the manner in which the tests themselves are selected.

Latent Variable Models and Factor Analysis (see Chap. 5)

This is an umbrella title which has been used to cover all statistical models which include latent variables. For our purposes, the *factor analysis* part is redundant and henceforth, we shall often omit it. Some special cases have already been covered above and the same array of observed and latent variables will serve here. However, the distinctive feature of the models included here is that the latent variables are now vector-valued of *unknown* dimension. When fitting a model we are interested in discovering whether or not a model fits the data where the dimension of $\boldsymbol{y}$ is to be determined. A related, but non-statistical problem, is to identify the latent variables with some hypothetical variables of substantive interest. In the case of ability testing we knew in advance that there was one latent variable and that it represented ability. More generally, the question of how many latent variables there are and what they represent is left open.

Missing Values (see Chap. 10)

Most statistical analyses in practice are bedevilled by missing values. In many cases this threatens any analysis carried out which ignores them. This is because, in the social sciences especially, being missing is often related to the subject matter of the investigation. The ease with which missing values can be disregarded has often led to them being unjustifiably ignored as van Buuren (2012) has shown. We shall discuss the problem in the present context which enables us to see it as one more example of looking at the distribution of the unobserved variables conditional on those that are observed. However the major difference between this problem and those considered earlier is that we are seldom interested in the missing values themselves but rather in the effect which their loss has on the estimation of the parameters of the model.

Social Measurement (see Chap. 11)

The proposition that all measurement can be expressed, in statistical terms at least, as one of estimation or prediction, as proposed in Bartholomew (1996), is something of a counsel of perfection. It presupposes that there is already a model to hand. In some very important practical problems the confusion lies at an earlier point, where measures have been proposed without reference to any explicit model and without seeing the need for one. Two such examples are provided by labour turnover (or wastage) and heritability. Both play an important role in public debate and yet, although the concept lying behind each of them is central to many debates, they are potentially highly misleading. In Chap. 11 we shall therefore illustrate how poorly founded both concepts are. We shall do this by reference to models of the processes underlying them, showing that the commonly used measures are confounded with other factors which can easily obscure what one is really looking for these examples do not, in every case, correspond precisely with individual chapters and, as we have already noted, some chapters cover aspects which span several fields of application. Nevertheless the foregoing categories will serve to locate the subject of this Brief on the broader map of Statistics.

1.5 Models and Misunderstandings

The subtitle of this Brief is intended to act as a warning. We shall have occasion to draw attention to particular instances where this warning is particularly apposite in the course of our exposition but some preliminary remarks are in order.

Modern statistics is built on the idea of models—probability models in particular. The standard approach to any new problem is to identify the sources of variation, to describe those sources by probability distributions and then to use the model thus created to estimate, predict or test hypotheses about the undetermined parts of that model. It was not always thus. It is difficult to identify any point in time at which the transition to analyses based on probability models took place, but

in the middle of the last century it was becoming increasingly common to construct and use models as part of everyday statistical practice. It might be justly argued that models were often implicit long before they were formulated explicitly but the failure to be explicit, especially in applications outside statistics, has given rise to many misunderstandings, as we shall see.

A statistical model involves the identification of those elements of our problem which are subject to uncontrolled variation and a specification of that variation in terms of probability distributions. Therein lies the strength of the statistical approach and the source of many misunderstandings.

Paradoxically, misunderstandings arise both from the lack of an adequate model and from over reliance on a model. Perhaps the best example of that, within our present purview, is in the case of factor scores treated in Chap. 6 but this is not an isolated example. More serious, is the failure to recognise the limitations of the modelling approach. At one level is the failure to recognise that there are many aspects of a model which cannot be tested empirically. At a higher level is the failure is to recognise that any model is, necessarily, an assumption in itself. The model is not the real world itself but a representation of that world as perceived by ourselves. This point is emphasised when, as may easily happen, two or more models make exactly the same predictions about the data. Even worse, two models may make predictions which are so close that no data we are ever likely to have can ever distinguish between them. We shall emphasise this point in relation to linear structural equations and other models in Chap. 7 but it is an ever-present danger. ***All model-dependant inference is necessarily conditional on the model.*** This stricture needs, especially, to be borne in mind when using Bayesian methods. Such methods are totally model-dependent and thus all are vulnerable to this criticism. The problem can apparently be circumvented, of course, by embedding the model in a larger model in which any uncertainties are, themselves, expressed in probability distributions. However, in doing this we are embarking on a potentially infinite regress which quickly gets lost in a fog of uncertainty.

References

Bartholomew, D. J., & David, J. (1996). *The statistical approach to social measurement*. San Diego: Academic Press.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis* (3rd ed.). Chichester, UK: Wiley.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley. (1st ed. 1987).

van Buuren, S. (2012). *Flexible imputation of missing data*. London: Chapman and Hall/CRC Press.

Chapter 2
Measurement, Estimation and Prediction

Abstract Measurement is commonly taken for granted in statistical work but, in the fields where missing observations occur, it is often the main objective. This is because the quantities to be 'measured' turn out to be represented by the parameters or random variables of a statistical model. Measurement then becomes a matter of predicting the values of random variables or of estimating the parameters of a distribution. When the unobserved variables are latent and, possibly indeterminate in number, the key idea is to determine their conditional distribution given what has been observed. This is essentially a routine matter involving the manipulation of probability functions. However, it is necessary to make clear what has to be defined and what are the constraints imposed by the logic of probability theory. This is important because much controversy, for example in relation to factor scores, has resulted from a failure to appreciate this point. We also introduce the one-parameter exponential family of distributions. This achieves a substantial simplification without incurring a serious loss of generality. In fact, it permits a considerable degree of unification of existing models and the development of new ones.

Keywords Conditional distributions · Estimation · Exponential family · Factor scores · Measurement · Prediction · Missing values

2.1 Measurement

In psychometrics and related branches of Science there is much discussion of measurement. In psychometrics, for example, there is the classical measurement model which supposes that what we observe differs from what we seek to measure by an 'error'. There is no comparable theory of measurement in Statistics where the term measurement is used in less specific ways. It is important, therefore, to be clear about how the general term 'measurement' is linked to the standard statistical procedures.

D. J. Bartholomew, *Unobserved Variables*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-39912-1_2,

Measurement is commonly defined as the assignment of numbers to objects in such a way that the numbers are related in ways which reflect the relationship between the objects. In one of the simplest cases, the length of objects, rods say, is reflected in the numbers which measure length. So if two rods of the same length are put end to end, the measure length of the combination will be twice that of each individual rod. It is not immediately obvious how this relates to statistical theory. The objects with which we deal in a statistical model are either parameters or random variables. The former are treated as fixed and the latter as varying in a way that can be described by a probability distribution. In Statistics the process of assigning numbers to parameters is known as *estimation* and the corresponding procedure for random variables is *prediction*. In statistical language, then, measurement is achieved by estimating unknown parameters or by providing predictors for random variables.

In the last chapter we saw that the unobserved variables in our models, the ys, could be regarded either as parameters or as random variables. We shall therefore need to consider the estimation and prediction problems to which these give rise.

2.2 Estimation

With one exception, the estimation problems posed by our models for unobserved variables are standard and straightforward and therefore require no special discussion. Thus, in the notation introduced in Sect. 1.1, if the ys are to be regarded as parameters they are no different from the θs and can, in principle at least, be estimated by standard methods. The important exception occurs with latent variable models where the number of ys may be proportional to the sample size. The asymptotic theory which is used to support the method of maximum likelihood in such cases, for example, requires the sample size to go to infinity with the number of parameters remaining fixed. In particular, this difficulty arises with the Rasch model which we shall look at in more detail in Chap. 4.

2.3 Prediction

All that we can know about the random variables in a statistical model is contained in their distribution conditional on all else that is known at the time the prediction has to be made. Any prediction for a random variable, based on a single number, will then be some measure of location of that distribution—often the mean. The key step, which lies behind all subsequent analysis, is then the determination of the relevant conditional distribution. In the remainder of this chapter we shall therefore set out the theory which is common to all of the models mentioned in Chap. 1 and which will be worked out in more detail in the following chapters.

2.4 Some Basic Distributional Results

All of the diverse procedures we shall meet share the same basic structure. There are two classes of variable to be distinguished: the observed, denoted by $\boldsymbol{x}$ and the unobserved variables, denoted by $\boldsymbol{y}$. The model, whatever the particular application, specifies the joint probability distribution of $\boldsymbol{x}$ and $\boldsymbol{y}$ but any inference has to be based on $\boldsymbol{x}$ alone since that is all that we can observe. The relationship between the two joint distributions is

$$f(\boldsymbol{x}) = \int f(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} \tag{2.1}$$

where the integral is over the range space of $\boldsymbol{y}$ and which, for reasons stated in Chapter 1, we have assumed $\boldsymbol{y}$ to be continuous. For the moment, any unknown parameters on which the distributions depend are to be understood, even though they are not made explicit. It is clear that further progress depends upon being able to specify the link between $\boldsymbol{x}$ and $\boldsymbol{y}$ and then this must be added to the specification. Equation (2.1) may place some restrictions on what models are possible. If, for example, we factorise the joint distribution as $f(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) f(\boldsymbol{y}|\boldsymbol{x})$, the factor $f(\boldsymbol{x})$ can be taken outside the integral where it cancels with the same factor on the left hand side. This produces the trivial and otherwise obvious result that the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$ must integrate to one. A more interesting case arises if we make the alternative factorisation $f(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{y}) f(\boldsymbol{x}|\boldsymbol{y})$, for then we have

$$f(\boldsymbol{x}) = \int f(\boldsymbol{y}) f(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{y} \tag{2.2}$$

It is clear from this equation that, though it does place some restrictions on the choice of the two distributions within the integral, the latter are not uniquely determined by Eq. (2.2). Once one member of the pair $\{f(\boldsymbol{y}), f(\boldsymbol{x}|\boldsymbol{y})\}$ is specified the other is determined by Eq. (2.2). Thus, in general, there will be infinitely many such pairs satisfying Eq. (2.2). This representation, and the associated equations, will form the starting point of almost every chapter. We shall illustrate the indeterminacy by a simple example in Chap. 3.

There is one important example of the situation we have described which is of considerable generality and widespread application, especially to latent variable models. This arises when the xs are assumed to be mutually independent, given $\boldsymbol{y}$. That is, we suppose that

$$f(\boldsymbol{x}|\boldsymbol{y}) = \prod_i f(x_i|\boldsymbol{y}) \tag{2.3}$$

and we let

$$f(x_i|\boldsymbol{y}) = F(x_i) G(\alpha_i) \exp(\alpha_i x_i) \tag{2.4}$$

with

$$\alpha_i = \alpha_i(0) + \alpha_i(1) y_1 + \alpha_i(2) y_2 + \ldots + \alpha_i(m) y_m \tag{2.5}$$

The probability function in Eq. (2.4) is known as the one-parameter exponential family. The family includes both continuous and discrete distributions—among which are the normal, Poisson, gamma distributions and many others. The parameter α_i is known as the canonical parameter and we have supposed in Eq. (2.5) that it is a linear function of the unobserved variables. First, under these assumptions, we start from the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$, given by

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{x}) &= \frac{f(\boldsymbol{x},\boldsymbol{y})}{\boldsymbol{f}(\boldsymbol{x})}, \\ &= \frac{f(\boldsymbol{y})f(\boldsymbol{x}|\boldsymbol{y})}{\int f(\boldsymbol{y})f(\boldsymbol{x}|\boldsymbol{y})d\boldsymbol{y}}. \end{aligned} \tag{2.6}$$

Next we substitute from Eq. (2.4) into Eq. (2.3) and then use the expression given by Eq. (2.6). If we look first at the parts which depend on the xs we note that the factor $\prod \psi(x_i)$ occurs in both numerator and denominator of Eq. (2.6) and thus cancels. In the remainder, xs only occur in the sums $\sum \alpha_i x_i$. So if we substitute the expression for α_i from Eq. (2.5) the sum becomes $\sum_j y_j X_j$ where $X_j = \sum_i x_i \alpha_j(i)$. It is clear, therefore, that the distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$ depends on the $\boldsymbol{x}$s only through the m linear functions $\{X_j\}$.

As we shall see later, this result has important practical implications. It supports the widespread empirical practice of choosing linear functions of the variables as indicators of an underlying latent variable. Furthermore, it delineates the circumstances under which such a practice may be justified. A fuller account of these manipulations will be found in Bartholomew et al. (2011),

Reference

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis* (3rd ed.). Chichester, UK: Wiley.

Chapter 3
Simple Mixtures

Abstract Mixtures of distributions play a fundamental role in the study of unobserved variables as Eq. (2.2) shows. The present chapter serves a double purpose in that it both prepares the ground for later chapters and treats a subject which has an intrinsic interest of its own. The two important questions which arise in the analysis of mixtures concern how to identify whether or not a given distribution could be a mixture and, if so, to estimate the components. We define finite and continuous mixtures and show, by examples, that it is very often extremely difficult to distinguish between them. Thus even if it is theoretically possible to make the distinction, it may be very difficult to do so in practice. Mixtures of normal and exponential distributions are both common and important and the mathematical simplicity of the latter makes them an ideal vehicle for exploring some of the fundamental issues.

Keywords Exponential distribution · Mixtures · Negative binomial distribution · Normal distribution · Poisson distribution

3.1 Introduction

All of the models discussed in this Brief can be regarded as mixtures and many of the topics covered in this chapter will also occur elsewhere in various guises. Here we shall treat mixing as an important topic in its own right and in its simplest form. This will also prepare the ground for the more subtle applications which occur in other types of problem.

Mixtures arise in practice because of failure to recognise that samples are drawn from several populations. If, for example, we measure the heights of men and women without distinction the overall distribution will be a mixture. It is relevant to know this because women tend to be shorter than men. The analysis of mixtures has two closely related objectives. Firstly, to identify whether a given sample could have arisen by mixing and then to estimate the components. The unobserved

D. J. Bartholomew, *Unobserved Variables*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-39912-1_3,

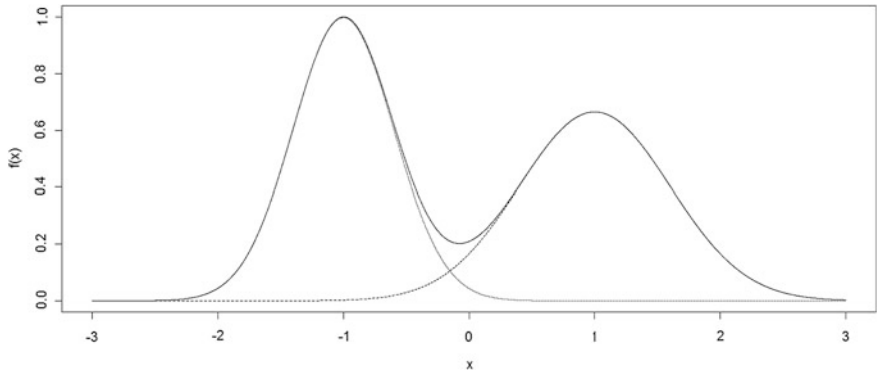

Fig. 3.1 A mixture of two normal distributions (*solid curve*) where the mixing is obvious

variable in this case indexes the member of the family of distributions from which the components of the mixture have come.

Sometimes a histogram will strongly suggest that mixing has occurred as Fig. 3.1 illustrates. The two humps strongly suggest that what we see is the result of mixing two distributions, each with a single mode.

This example is exceptional in that the shape of the solid curve strongly suggests that the distribution is actually a mixture of the two normal components which are shown as dotted curves on the figure. It is often not at all obvious whether a given distribution could be a mixture and this situation is illustrated in Fig. 3.2 The distributions shown in Figs. 3.1 and 3.2 may both be described as 'two-component mixtures'.

In general the probability distribution of a two-component normal mixture may be written

$$f(x) = pN\left(\mu_1, \sigma_1^2\right) + (1-p)N\left(\mu_2, \sigma_2^2\right) \tag{3.1}$$

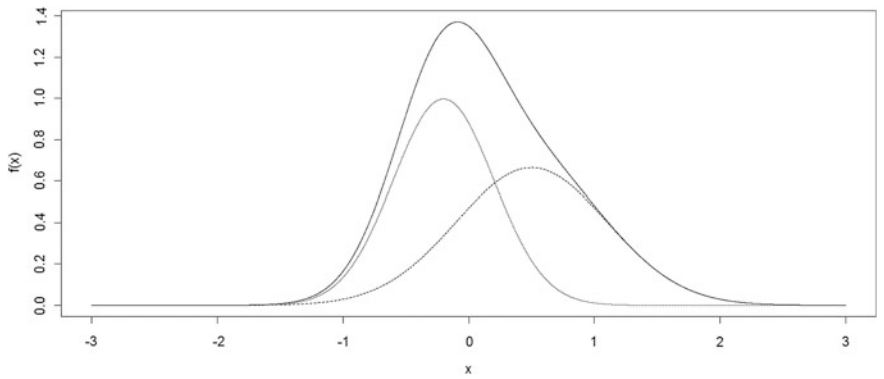

Fig. 3.2 A mixture of two normal distributions where the mixing is not obvious

where $N(\mu_i, \sigma_i^2)$ is the probability function of a normal random variable with mean μ_i and variance σ_i^2. It will be evident that, even a two-component mixture of normals, has 5 unknown parameters. As further components are added the estimation problems become formidable. If there are many components, separation may be difficult or impossible. Many other examples are given in Titterington et al. (1985).

The distribution represented by Eq. (3.1) is a simple example of a *finite* mixture. It can easily be extended to include more components. It is then a natural model to use when it is possible that the sample has come from a distribution formed by mixing a small number of simpler distributions—which do not have to be of the same form, of course. In other contexts it may be more natural to consider *infinite*, or *continuous*, mixtures. Such a mixture may be written

$$f(x) = \int_0^\infty f(x|\lambda)h(\lambda)d\lambda \tag{3.2}$$

where λ indexes the members of the family of distributions being mixed and $h(\lambda)$ is the probability function of λ. Such a model is often used when it is desired to build into the model the possibility that λ varies continuously in some, possibly unknown, manner. The unobserved variable here is λ and its variability is often described as unobserved heterogeneity.

3.2 The Negative Binomial Distribution

The example of unobserved heterogeneity which arises in the study of accident distributions has already been mentioned in Chap. 1. For any given individual we supposed that accidents occurred randomly and hence that the distribution of the number per unit time would have a Poisson distribution with mean λ, say. But if the risk of having an accident varies from one individual to another, what we observe will be a continuous mixture whose probability distribution will have the form,

$$f(x) = \int_0^\infty \frac{\lambda^x}{x!} e^{-\lambda} h(\lambda)\, d\lambda \tag{3.3}$$

The two questions which we raised above about mixtures in general can now be expressed by asking what can be deduced about $h(\lambda)$ and about whether the conditional distribution of x can be inferred from the observed distribution, $f(x)$, of x.

One approach to the first question is to select a parametric form for $h(\lambda)$ which is sufficiently flexible to describe wide patterns of variation. One such distribution, which also has the useful property that the integral of Eq. (3.3) can be evaluated in closed form, is

$$h(\lambda) = \frac{c^q}{\Gamma(q)} \lambda^{q-1} e^{-c\lambda}. \tag{3.4}$$

In this case it is easy to show that

$$f(x) = \binom{x+}{x} q^{q-1} \left(\frac{c}{c+1}\right)^q \left(\frac{1}{c+1}\right)^x. \tag{3.5}$$

This is a discrete distribution called the negative binomial distribution because the probabilities are such as would be obtained from the expansion of a binomial expression with negative index. It is more highly dispersed than a Poisson distribution having the same mean. If we fit the distribution of Eq. (3.5) we may obtain estimates of the parameters q and c which determine the mixing distribution. $h(\lambda)$.

3.3 Determination of the Mixing Distribution

The last example may have suggested that the mixing distribution could always be determined. This is not necessarily true either exactly or, even approximately. We illustrate the situation by two examples.

Suppose with x has a conditional distribution which, given μ is $N(\mu, \sigma^2)$ and that μ is a random variable has a distribution which is $N(0, \tau^2)$. It is then straightforward to show that the unconditional distribution of x is $N(0, \sigma^2 + \tau^2)$. (This is also a special case of a basic result in factor analysis—see Bartholomew et al. (2011) equations (1.11 and 1.12). This is what we described earlier as a continuous mixture; in this case of normal distributions. The first important point to note is that the *form* of the distribution is unaffected by the mixing. Thus there is no way that we can recognise that mixing has taken place by inspecting the form of the resulting distribution alone. Any given normal distribution could have arisen naturally or be the result of normal mixing.

The second point to notice is that the variance of the mixture is greater than that of the original distribution. In the case of the negative binomial distribution we also noted that its spread would be greater than that of the Poisson distribution, having the same mean, from which it was generated. It is generally the case that mixing will increase the spread of a distribution, as measured by the variance, and later examples will also show this.

The fitting of mixtures of distributions to random samples is not always easy but can sometimes be achieved using the E-M algorithm discussed in Chap. 10.

3.4 The Mixed Exponential Distribution

The exponential distribution is, perhaps, second in importance only to the normal distribution and mixtures of exponential distributions have found many applications both because of their practical relevance and of their mathematical

tractability. Many of the general questions which arise in the study of mixtures can be answered explicitly for the mixed exponential and thus shed some light on more general issues. Here, it will be useful to consider the family in its finite and continuous forms.

The probability function of the finite exponential mixture may be written

$$f(x) = \sum_{i=1}^{k} p_i \lambda_i e^{-\lambda_i x} \tag{3.6}$$

Where $p_i > 0$ and $\sum_{i=1}^{k} p_i = 1$.

The probability function of the corresponding continuous mixture is

$$f(x) = \int_0^\infty \lambda e^{-\lambda x} h(\boldsymbol{\lambda}) d\lambda \tag{3.7}$$

where $h(\lambda)$ is the probability function of λ.

A particularly convenient feature of Eq. (3.7) is that it has the form of a Laplace transform. In fact $f(x)$ is the Laplace transform of $\lambda h(\lambda)$ and this fact makes it possible, in principle, to find the mixing distribution in any particular case.

3.5 The Sensitivity of the Mixing Distribution

The foregoing results suggest approaches to ways of determining the mixing distribution once $f(x)$ is known. However, a much more important practical question is how to *estimate* the mixing distribution from an *estimate* of $f(x)$. This question is not one which lends itself to an immediate answer by the proof of general mathematical theorems but we can obtain a few indications from mathematical analysis about show how the land lies. In particular we shall see that there may be little information in our estimate of $f(x)$ about the form of the mixing distribution. We have already seen that, if $f(x)$ is normal, there is no way of knowing whether it is the result of mixing and hence, if it is, what the mixing distribution might be. Some further light will be shed on the matter by looking at a finite mixture of exponentials and comparing the finite and continuous mixture of exponentials in particular cases.

Suppose we have a continuous mixture of exponentials defined as follows. Let $h(\lambda)$ in Eq. (3.7) have the probability function

$$h(\lambda) = \frac{c^q}{\Gamma(q)} \lambda^{q-1} e^{-c\lambda} \tag{3.8}$$

then it follows that

$$f(x) = \frac{q}{c} \frac{1}{(1 + \frac{x}{c})^{q+1}}. \tag{3.9}$$

It is obvious that this distribution, like the exponential, is monotonic decreasing over its whole range and that it starts from a point at $t = 0$ at which it is greater than the exponential having the same mean. Some more information about its shape can be deduced by looking at its behaviour when q is large (meaning that the mixing distribution shows little variation). For the distribution of Eq. (3.9)the mean, μ, is $c/(q-1)$ and so we may re-parameterise the distribution in terms of μ and a 'shape' parameter q as follows

$$f(x) = \frac{q}{(q-1)\mu} \frac{1}{\left(1 + \frac{x}{(q-1)\mu}\right)^{q+1}} \tag{3.10}$$

Using the approximation

$$\left(1 + \frac{a}{n}\right)^n \sim e^x \left(1 - \frac{a^2}{2n}\right) \tag{3.11}$$

we find

$$f(x) \sim \frac{1}{\mu} e^{-x/\mu} \left[1 + \frac{1}{q}\left\{1 - 2\frac{x}{\mu} + \frac{1}{2}\frac{x^2}{\mu^2}\right\}\right] \tag{3.12}$$

For large q it is clear that $f(x)$exceeds the exponential near the origin and in the upper tail but is below it in the neighbourhood of the mean.

A good deal can be learnt about sensitivity by a study of the finite mixture given by

$$f(x) = \sum_{i=1}^{k} p_i \lambda_i e^{-\lambda_i x} \tag{3.13}$$

If $k = 2$ for example, the mixing distribution consists of two discrete probabilities of magnitude p and $1 - p$. This appears at first sight to be radically different from the continuous distribution given by (3.7). Yet in practice it has proved very difficult to distinguish the two types of mixture. Qualitatively, both distributions are monotonic decreasing over their whole range with excesses of frequency, compared with the exponential, near the origin and in the upper tail. When the mean is fixed, the continuous distribution of Eq. (3.9) has one free parameter which determines the shape of the distribution. The corresponding two-term exponential has three parameters altogether so when the mean is fixed there are, effectively, two free parameters remaining to determine the shape. One would therefore expect to be able to bring the two distributions close together by appropriate choice of parameter values though this is not immediately obvious because the parameters do not have unrestricted range. However, it is easy to verify numerically that pairs of members of the two families can be very close. In Fig. 3.3 we have illustrated the position by taking particular examples. We fix the means to be unity in each case. For the continuous family of Eq. (3.9) we take $q = 4.25$. For the two term exponential the distribution having parameters

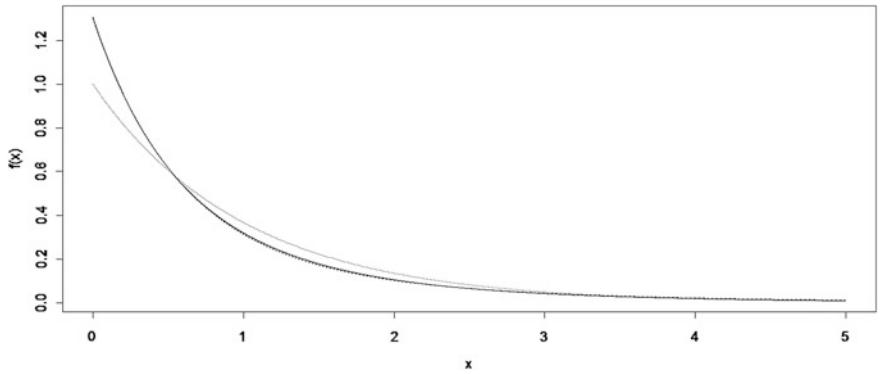

Fig. 3.3 Comparison of a continuous and a discrete mixed exponential distributions with an exponential distribution having the same mean

$p = 0.391$, $\lambda_1 = 0.6$ and $\lambda_2 = 1.75$ has the same mean and upper and lower quartiles.

The dotted curve is the exponential; the solid curve the continuous mixed exponential and the dashed line (hardly distinguishable from the solid curve) the two-term mixture. It is obvious that any information in the last two cases about the mixing distribution has to be gleaned from the minute difference between the two plotted curves which would be undetectable in practice. For all practical purposes it is thus clear that although the effect of mixing is to make the distribution more skewed, the form of the mixing distribution has virtually no detectable influence. These somewhat fragmentary numerical results suggest that, for exponential mixtures at least, the information contained in the distribution $f(x)$ about the form of the mixing distribution is negligible. This simple fact will have far-reaching implications when we come to consider latent variable models in later chapters.

References

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis* (3rd ed.). Chichester, UK: Wiley.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Chichester, UK: Wiley.

Chapter 4
Models for Ability

Abstract Here we begin our approach to a general class of latent variable models by studying a simple example set within the context of a particular practical problem. This offers the readily intelligible vocabulary of ability testing and it firmly anchors what is sometimes seen as a somewhat esoteric topic in the mainstream of statistics by drawing on the ideas of the analysis of variance. First we show that the Rasch model fits neatly into that framework and that, by generalising it in a number of directions, the link with a special case of factor analysis can be made. In particular, it is the random effects version of the analysis of variance which provides that link and it also brings into the picture what psychometricians call generalizability theory. Maximum likelihood estimation and notions of sufficiency also appear in a central role which is developed in later chapters. A good deal of misunderstanding and controversy has surrounded the Rasch model and we hope some of this may be dispelled by the present approach.

Keywords Analysis of variance · Factor analysis · Generalizability theory · Maximum likelihood · Rasch model · Psychometrics · Latent variable models · Random effects

4.1 The Problem and the Models

Although the class of models we are about to consider is quite general, we shall discuss it in the context of ability testing. It is this application which provides the motivation and the context for the models and the misunderstandings to which they have given rise. We begin with the data matrix having the form set out in Chap. 2, namely

D. J. Bartholomew, *Unobserved Variables*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-39912-1_4, © The Author(s) 2013

$$\begin{array}{ccccc} x_{11} & x_{12} & \ldots & x_{1m} & y_1 \\ x_{21} & x_{22} & \ldots & x_{2m} & y_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{nm} & y_n \end{array}$$

In the testing context, the xs would represent the scores obtained by each of n individuals on m tests, or items. The ys in the original table in Chap. 2 were the unobserved variables, which in the present context, will be ability scores; for the moment we ignore these and begin farther back.

Any individual score will depend on both the individual providing it and the test item being undertaken. Since the score will vary and depend only on the row and column in which it appears, it needs to be modelled by a random variable whose distribution depends likewise on both row and column. The form which this model takes will depend on what kind of a variable x is. Commonly x is a binary variable which, conventionally, only takes the values 0 and 1. This specification has given rise to several models; these include the Rasch model and item response models to which we come later. (It should be noted that the interpretation of the Rasch model has been extremely controversial in educational testing circles and this makes it doubly important to see it arise naturally in the present setting). There is an advantage, conceptually at least, in starting with the case where the xs are assumed to be continuous and normally distributed.

In the case of the so-called *fixed effects* model of the analysis of variance we have something which is equivalent to the two-way standard analysis of variance set up with one observation per cell. The latter feature is important and it limits what can be learnt from the data.

The standard linear model would express all of this in the linear equation,

$$x_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \tag{4.1}$$

where $\sum_i \alpha_i = \sum_j \beta_j = 0$ and the errors, e_{ij}, are independently and normally distributed with means zero and common variance σ^2. The parameters α_i and $\beta_j$are the 'row' and 'column' effects respectively. Under these assumptions the maximum likelihood (or least squares) estimators of the parameters and the unknown variance can easily be obtained. If the xs are continuous but not normal, it may be possible to transform them to normality and thus make the model applicable. Note that in this model there is no replication within the cells and so it is not possible to test for any interaction between person and item effects. Any such interaction must therefore be assumed to be zero since there is no possibility of replication within cells because this would mean obtaining several independent test scores on the same item for the same individual and this is clearly impossible.

If, instead of being fixed, the subjects are sampled at random from some population, we would have a random effects model with the ys in the array above being random variables. On the assumption that the 'person' effects were normally distributed, we could estimate the variance of their distribution from the appropriate mean squares of the analysis of variance table. The whole analysis is

therefore covered by the standard theory of the analysis of variance though the assumptions underlying the analysis must be emphasised; in particular, the normality of all the random variables involved and, especially, the non-existence of an interaction. An interaction might easily arise in practice if some of the persons had particular familiarity with test items of a particular character.

So far we have said nothing, about the items. If the item effects are treated as fixed, then our analysis will apply only to those particular items. If they are regarded as sampled at random from a population of items a random effects model with two sets of unobserved variables would apply. However, in practice it is rare for the test items to be selected at random from any population.

The foregoing model will serve if the test scores are normal, or can be transformed to normality. But one of the commonest situations, which has attracted the greatest attention in this field, is the one where the scores are binary. This arises if the items are either 'right' or 'wrong'—usually scored 1 and 0. In this case the linear Eq. (4.1) cannot be appropriate and an alternative must be sought. The uncertainty about the responses can be captured by the probability that person i gives a correct answer to item j. Problems which involve this specification are often referred to as item response tests (IRT) or item response models (IRM) and they have generated an enormous literature. A good survey of this field is provided by Thissen and Steinberg (2009) who also provide a wide-ranging list of references including the relevant papers by Rasch.

Let the probability that person i makes a positive response to item j be denoted by π_{ij}. We need to express this as a function of the parameters μ, α_i and β_j in such a manner as to make the probability lie between 0 and 1. Since the linear combination of Eq. (4.1) is unbounded, a natural way to achieve this to choose

$$\pi_{ij} = F(\mu + \alpha_i + \beta_j) \tag{4.2}$$

where F(.) is a cumulative distribution function. However, there is a natural constraint to be placed on the choice of this function which arises from the fact that the labelling chosen for the binary response is entirely arbitrary. We labelled the positive response 1 but we could equally have labelled it 0. We would not wish this choice to change the form of the response probability so we also require that $F(x) = 1 - F(-x)$ which then implies that $\pi_{ij} = 1 - (1 - \pi_{ij})$, as it obviously does. F(.) is therefore the cumulative distribution function of a symmetrical distribution centred at 0.

There are two distributions commonly used for this purpose, namely the normal and the logistic. We shall see that these distributions have particular interpretations when we come to consider what we shall call 'underlying variable' models but, at this stage, they are adopted simply as convenient functions which have the right form. The logistic form is chosen here because it matches the conditions required by the general theory set out in Chap. 2. It may be written

$$\text{logit}\pi_{ij} = \log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = (\mu + \alpha_i + \beta_j). \tag{4.3}$$

4.2 Maximum Likelihood Estimation

The likelihood function for the model defined above is

$$\begin{aligned} l &= \prod_{i,j} \pi_{ij}^{x_{ij}} \left(1 - \pi_{ij}\right)^{1-x_{ij}}, \\ &= \prod_{i,j} \left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right)^{x_{ij}} (1 - \pi_{ij}) \\ &= \prod_{i,j} (1 - \pi_{ij}) \exp\{x_{ij}(\mu + \alpha_i + \beta_j)\} \end{aligned}$$

hence

$$\mathrm{L} = \log l = \text{constant} + \mu X + \sum_i \alpha_i X_i + \sum_j \beta_j X_j \tag{4.4}$$

where

$$X = \sum_{i,j} x_{ij},\ X_i = \sum_j x_{ij} \text{ and } X_j = \sum_i x_{ij}.$$

This is known as the Rasch model after its originator who developed it from first principles, apparently without realising how closely it was connected to the two-way analysis of variance. It is usually parameterised in a slightly different way without the αs and βs being referred to an arbitrary origin. The advantage of doing it in our way is that it makes the link with the familiar analysis of variance and emphasises that the origin of the row and column parameters is arbitrary. One advantage often claimed for the Rasch model is that once the column parameters have been estimated using one selection of test items, the estimates of the row parameters should be unchanged even if a different set of persons is tested using those same items. This property is automatically ensured by the standard formulation of the analysis of variance model. Although the Rasch model has many attractive properties it must be remembered that it is a 'fixed affects' model and thus only enables us to test its fit with the particular set of items and persons for which we happen to have data. This is rarely what we are actually interested in. The persons will have been selected from some population to which we may wish to generalise our conclusions. It is straightforward to write down a model to cover this case but this does not appear to have been studied. Further, we might wish to generalise about the whole set of items though, since they are often specially constructed there may not be an extant population to which we can refer. This latter point is one to which we shall return shortly.

4.3 Continuous Variables

If we return to the case where the xs are continuous variables we are back with the analysis of variance set-up. If we then make the additional assumption that the variables are normally distributed, the standard theory will apply. The model will then be that of Eq. (4.1) where the αs represent the row(person) effects and the βs the column(item) effects.

This means that for any model for which we specify in terms of the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$, that model will not be unique. We cannot make inferences about the interaction effects in the two-way analysis of variance with only one observation per cell because all the degrees of freedom have been used up by the main effects. Or, put another way, the interaction and residual effects are confounded. However, it is possible to make some progress in this direction in the case of the random effects version of the model. This is evident if we write it in the forms

$$x_{ij} = \mu + \alpha_i + \beta y_j + e_{ij} \tag{4.5}$$

where $\mu + \alpha_i$ continues to represent the item effects and βy_j is now the contribution of the randomly selected jth person. In analysis of variance terminology this is a 'mixed' effects model with one fixed effect and one random effect. When expressed in this form it is, as we shall see later, also essentially the same as a special case of the 'one-factor' model of factor analysis. To make the correspondence complete we add a subscript i to β. By this addition we have allowed the contribution made by the 'person' effect to depend on the item whose score is being determined. This is a rather special kind of interaction which makes particular sense in ability testing. The parameter β_i is referred to in this context as the *discriminating power* of the item. Thus if β_i is large the person with y makes a bigger contribution than someone with a smaller discriminating power. The only difference between Eq. (4.5) and the usual factor analysis model is that we have expressed the item effect as the sum of a 'grand' mean, μ, and a deviation, α_i, instead of the more usual μ_i.

With equal discriminating power we have a continuous analogue of the Rasch model. When we are dealing with continuous test scores, we find we have enough further information in the data to estimate the discriminating effect of the items as well. Maximum likelihood estimation could be used in this case for the fixed effect but, as the model we have just described is a special case of a more general class to be discussed in the next chapter, we shall return to it there.

4.4 The Selection of Items

In the foregoing discussion we have supposed, in turn, that the 'person' effect was fixed or random. The 'item' effect has been fixed in both cases. From a theoretical perspective it would seem natural to allow the item effect to be random also, but

there is rarely any physical justification for doing so. It is certainly possible to imagine a population of items from which those actually used have been drawn and one can envisage this happening sometimes in practice. For example, if the test items consist in adding up a fixed number of 2-digit numbers there is a finite number of such 'sums' and hence those used are a sample from such a population. However, it would be hard to find many cases where this had been done. Alternatively, the test items may be constructed to span the range of ability which the item is intended to cover. It is not usually possible to relate this in a meaningful way to any formal process of sampling. Nevertheless, psychologists and others have recognised the need (which statisticians have often failed to do) to generalise the results obtained for a particular set of items to the larger population from which the items have been drawn. This has become a major field of research activity know as *generalizability theory*, or, somewhat more narrowly, as Psychometric inference.

Although we are dealing here with unobserved variables they are certainly not random variables. But, even if nothing is known about how the test items were selected, or constructed, it may be possible to draw some conclusions, however limited. We can certainly learn something about the variability of item difficulty by inspecting the 'between items' sum of squares. If this is very small then, however the items have been selected, we would be more confident in generalising the conclusions from the analysis of 'all items' than if they were widely dispersed. However, we can go a little farther if we consider the two-way analysis of variance for the data matrix. If we set out the table as follows, using SS and MS to denote 'sum of squares' and 'mean square' respectively and DF the degrees of freedom, then the analysis of variance table is as follows

Between persons	SS_P	DF_P	MS_P
Between tests	SS_T	DF_T	MS_T
Residual	SS_R	DF_R	MS_R

The residual mean square, which is the same as the interaction mean square for this data matrix, is a measure of the extent to which variations in persons is associated with variation in tests. If this is small, it means that there is very little association between tests and persons and hence MS_P is largely unaffected by which of those tests happen to have been selected. Conversely, a large residual mean square tells us that the choice of tests does matter very much. Hence a suitable coefficient for measuring the extent to which the variation is unaffected by test differences is:

$$\frac{MS_P - MS_R}{MS_P}$$

This is equivalent to one version of a coefficient often known as 'coefficient alpha' due to Cronbach (1951). It is also closely related to measures of test reliability.

There is, of course an assumption implicit in this argument. We have assumed that the interaction mean square would be much the same for any other set of test items used. This would be assured if the items were sampled randomly from some population of items but otherwise it would be difficult to justify.

References

Cronbach, L. J. (1951). Coefficient alpha and the structure of tests. *Psychometrika, 48*, 171–192.
Thissen, D., & Steinberg, L. (2009). Chapter 7. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology*. London: Sage Publications Ltd.

Chapter 5
A General Latent Variable Model

Abstract This chapter shows that, starting from the two-way analysis of variance with random effects, it is possible to arrive at a general latent variable model. It does this, first by enlarging the family of conditional distributions considered and secondly, and more fundamentally, by allowing the random effect associated with the rows of the two-way table to be linear in a set of (unobserved) latent variables. In the case when the observed variables are conditionally normal, the standard model for factor analysis emerges but the framework adopted includes a great many other possibilities, including non-linear models. One advantage of adopting this general framework is that it makes the essential arbitrariness of the distribution of the latent variables transparent. It also paves the way for the following chapter in which we turn to clarifying what can be said about the prediction of the latent variables.

Keywords Analysis of variance · Factor analysis · Random effects · Latent variables · Latent variable models · Two-way classification

5.1 An Extended Model

The two-way table which has been the basis of the models for ability in the last chapter can be extended to become the basis of a more general class of models as we now show. The extra generality is achieved partly by enlarging the class of distributions from which the *x*s are supposed to have come. Instead of considering only two conditional distributions—the binary and the normal—we now suppose that the distribution may any member of the exponential family. This leads, quite naturally, to the classical factor analysis model although it is not usually approached in this way.

There are two particular aspects of inference with which we shall be concerned. One, which we deal with in this chapter, is that of estimating the parameters of the model. The second is measuring latent variables, to which we come in the next chapter.

D. J. Bartholomew, *Unobserved Variables*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-39912-1_5, © The Author(s) 2013

We start with the data matrix as set out at the beginning of the last chapter. The only observable quantities are the xs and estimation must therefore start with their joint distribution. We have already met the cases where the xs are normal or binary. Here we shall suppose that their distribution is a member of the one-parameter exponential family. To be specific, we suppose that the probability function, $f(x)$, has the form

$$f(x) = F(x)G(\theta)\exp\theta x \tag{5.1}$$

This includes the binary and normal distributions as special cases and we note again that x can be categorical or continuous. In the two-way table x will be indexed by i and j. We adopt the model

$$\theta_{ij} = (\mu + \alpha_i + \beta_j) \tag{5.2}$$

where θ_{ij} is the value of theta for the distribution of the observation in cell (i, j) of the table. Then, following in the same steps as in Chap. 4, we find that the likelihood function has exactly the same form as in Eq. (4.4) of that chapter. Hence, we deduce that the row and column totals are jointly sufficient for their corresponding row and column parameters. This is what intuition might lead us to expect and it shows that the well-known results for the Rasch model apply to a much wider class of models.

If we are dealing with the random effects version of the model, β_j in Eq. (5.2) will be a random variable and the model is as expressed in the same fashion as in Eq. (4.5) of the last chapter. In the notation of the present chapter, the values which it takes for each column of the table may be supposed to be drawn from a population with probability function $\psi(\beta)$ centred at zero. The joint probability function for the xs appearing in the jth column of the table may then be written

$$f_j(x_1, x_2, \ldots\ldots x_n) = \prod_i F(x_{ij}) \int \prod_i G(\mu + \alpha_i + \beta) \left\{ \exp \sum_i (\mu + \alpha_i + \beta) x_{ij} \right\} \psi(\beta) d\beta \tag{5.3}$$

One could construct a likelihood function from this joint probability function if the form of $\psi(\beta)$ were known and then if we could estimate the α_is and any unknown parameters in $\psi(\beta)$. We now look at this matter in more detail.

5.2 More Than One Latent Variable (Factor Analysis)

In the last section we referred to two generalisations that were to be made. The first was to use the exponential family in place of the normal or binary. The second, with which we deal now, is to allow the 'random effects' to be multi-dimensional. In the two-way analysis of variance situation it is natural to think of the 'random

effect' as one-dimensional. But we now move on to consider situations where more than one random variable may be involved and where the various 'effects' combine to produce the overall effect. In the analysis of variance problem the 'effect', if there is one, is a clearly identifiable source of variation in which we are interested though it is often called a factor—a particularly confusing use of the term in this context! In what is called 'factor analysis', the nature of the 'effect' is less well-defined and a prime purpose of the analysis is then to establish its existence and elucidate its character.

We begin by supposing that there may be a set of random variables, of unknown number, which contribute to the value of the xs in each row of the table. Clearly this contribution may be specified in a variety of ways but the obvious way to do this is by supposing that it is through some function of the latent variables. We shall formalise this by supposing that the single random variable envisaged in the random effects analysis of variance is replaced by some function of these new latent variables. Almost all the work is based on linear functions though other possibilities, including polynomial functions, have been envisaged. Here we consider only linear functions, supposing that the random effect is so represented and thus we have the model

$$x_{ij} = \mu + \alpha_i + \lambda_{1i}y_1 + \lambda_{2i}y_2 + \cdots + \lambda_{qi}y_q + e_{ij} \tag{5.4}$$

where q is unknown and e_{ij} is the usual error term, which is assumed to be independent of the ys. There are good reasons, which we shall come to in the following chapter, for treating the ys as mutually independent standard normal variables. In that case the xs turn out to have a joint multivariate normal distribution with covariance matrix

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}'\boldsymbol{\Lambda} + \boldsymbol{\Psi} \tag{5.5}$$

where $\boldsymbol{\Lambda}$ is a $n \times q$ matrix of the loadings (λs) as they are known in this context, $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix whose elements are the variances of the e_{ij}s. The problem in the case of this model is to make inferences about the number and interpretation of the ys.

Other latent variable models arise when, in effect, we make different distributional assumptions about the e_{ij}s or choose a different function of the ys in Eq. (5.4).

An alternative and more usual way of introducing the general factor analysis model, which does not start from the two-way-table, is that given, for example, in Bartholomew et al. (2011) and most standard works. It is interesting to see how the two are related. The usual approach starts along the lines set out in Chap. 1 (Sect. 1.1) where we considered two types of random variable, the xs which were collected in a p-dimensional vector $\boldsymbol{x}$ and the latent variables in a q-dimensional vector, $\boldsymbol{y}$. Only the vector $\boldsymbol{x}$ was observed and so the only probability distribution about which we could make inferences was $f(\boldsymbol{x})$. Any model tells us how the xs are related to the unobserved latent variables. This may be expressed, as we saw in Chap. 2, by the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y}$ which we denoted there by $h(\boldsymbol{x}|\boldsymbol{y})$. The prime purpose of any latent variable model was then to learn

something about $\boldsymbol{y}$ when $\boldsymbol{x}$ is given and this information is conveyed by $h(\boldsymbol{y}|\boldsymbol{x})$. By Bayes' theorem we deduced that

$$h(\boldsymbol{y}|\boldsymbol{x}) = h(\boldsymbol{y})h(\boldsymbol{x}|\boldsymbol{y})/f(\boldsymbol{x}). \tag{5.6}$$

(The fact that we have used 'h' to denote probability functions here rather than the 'f' of Eq. (2.2) has no significance but is to conform with usage in the context of this chapter). This result contains two important messages which are often overlooked and are worth repeating because they are a common source of misunderstanding. First, the conditional distribution, in which we are interested, depends on $h(\boldsymbol{y})$ which will usually be unknown. We return to this point in the following chapter which is about the prediction of $\boldsymbol{y}$. The second point, which is essentially the first in a different guise, is that the distribution of $f(\boldsymbol{x})$, from which any unknown parameters must be estimated, also depends on this unknown distribution. The reason that this rather important element of the model is often overlooked is that there is a common but unjustified tendency among modellers to regard any unknown distribution as standard normal. The present situation is quite different and, in general, no such assumption is justified.

There is no necessity, of course, to be restricted to the linear model of Eq. (5.4). The possibilities are endless but it turns out that the linear model includes virtually all models in current use and provides a foundation from which others could be developed. There are models in use which do not belong to this family but, in practice, they are hardly distinguishable from those which are included.

Reference

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis* (3rd ed.). Chichester, UK: Wiley.

Chapter 6
Prediction of Latent Variables

Abstract The second main object in finding the conditional distribution of a latent variable is to make a prediction. In the factor analysis literature this is known as 'the problem of factor scores'. This has been a major source of controversy and debate because of the failure to recognise the distinction between a parameter and a random variable. The debate has concerned only the case of the normal model, for good historical reasons. For all other models, which have only come into prominence recently, the approach advocated here has been followed without controversy. For the normal case there are two competing solutions, known by the names of their originators, Thomson and Bartlett. They sometimes coincide numerically, but otherwise differ because they provide solutions to different problems. Thomson's scores aim to predict the values of future observations of the latent variable whereas Bartlett's scores are estimators of the values taken by the latent variables for particular individuals. This feature means that they must be regarded as parameters.

Keywords Bartlett · Bayes theorem · Factor analysis · Factor scores · Latent variables · Prediction · Thomson

6.1 Prediction and Factor Scores

Perhaps the greatest misunderstanding in this field surrounds what is called the factor scores problem. This confusion can be largely dispelled by recognising that the problem centres upon latent variables and their conditional distributions and hence that the problem is essentially one of prediction. Since latent variables are elements in a model, the whole issue resolves itself into one of appropriately specifying the model in which they appear.

We shall begin by considering the problem in general terms because this makes it easier to see how to handle the problem at that level. Because it has been at the root of so much misunderstanding, we shall repeat much about the model that has

D. J. Bartholomew, *Unobserved Variables*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-39912-1_6,

already been covered, especially in Chap. 4. This, more general approach also serves to clear up the long-standing, but only apparent, confusion surrounding what have become known as Thomson's regression scores and Bartlett's scores.

In any latent variable problem, as we have mentioned several times before, there are two kinds of variable—the manifest, or observable, variables and the latent, or unobservable variables. As before, the former will be denoted by x and the latter by y and individual values will be distinguished by subscripts. Since we observe the xs alone, all inferences must ultimately depend on their joint distribution. It is important to emphasise again that, at the stage our problem arises, the xs will already have been observed and hence all that we can know about the latent variables is contained in their joint distribution conditional on the xs. This, as we know, is given immediately by Bayes theorem and can be expressed as

$$h(\boldsymbol{y}|\boldsymbol{x}) = \frac{h(\boldsymbol{y})f(\boldsymbol{x}|\boldsymbol{y})}{f(\boldsymbol{x})} \tag{6.1}$$

Of the quantities appearing on the right hand side, $f(\boldsymbol{x})$ is known, or can be estimated, $f(\boldsymbol{x}|\boldsymbol{y})$ is specified by the model but $h(\boldsymbol{y})$ is unknown. Since $f(\boldsymbol{y}|\boldsymbol{x})$ is a probability function its integral, or sum over the whole sample space must be 1, so the denominator of Eq. (6.1) is a constant factor determined by that fact. There is no means of determining this distribution from the data and the 'prior' distribution is completely arbitrary. This means that that we cannot determine $h(\boldsymbol{y}|\boldsymbol{x})$, or any summary measure derived from it. The general problem, as we expressed it above is therefore insoluble. However, this also means that any distribution we may choose to use must be a matter of convention only and cannot claim any support from the data. This elementary point, which we have already made, has often been overlooked and papers have been published purporting to 'estimate' the distribution. It is evident that in order to do this, something else must be assumed.

The normal model expressed as a linear equation in Eq. (6.4) of the last chapter can be cast into this form by writing the conditional distribution of x as

$$x_{ij}|\boldsymbol{y} \frown N(\mu + \alpha_i + \lambda_{1i}y_1 + \lambda_{2i}y_2 + \cdots \lambda_{qi}y_q, \sigma_i^2. \tag{6.2}$$

There are good practical reasons in favour of using a standard normal distribution for the prior distribution for each of the elements of $\boldsymbol{y}$ and assuming them to be mutually independent. This is because the normal distribution commonly arises in measurement work, partly since many naturally occurring quantities used as measures have distributions close to normal and partly because it spaces individuals along the scale of measurement in a way which accords with our intuition. Furthermore, the normal distribution is well-known and its properties are readily available. However, we emphasise again, that its adoption is a matter of convenience and not an empirical fact.

If we insert this conditional distribution into Eq. (6.1) and choose the elements of $\boldsymbol{y}$ to be independent standard normal, it is readily shown that

$$\boldsymbol{y}|\boldsymbol{x} \frown N(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}(x - E(\mathbf{x})),\ \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} + \boldsymbol{I})). \tag{6.3}$$

This distribution expresses our uncertainty about $\boldsymbol{y}$ when $\boldsymbol{x}$ is given. In this most general form the ys are not conditionally independent and hence what we say about any particular y will depend on the values of the other. An important special case arises when $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ is diagonal in which **case** the ys are conditionally independent. (In practice one can ensure that this condition is satisfied by an appropriate rotation.) If we need a single summary measure for y there is still a choice of what is the most appropriate location measure but, in the case of the normal distribution, all the usual measures, mean, median and mode coincide. We may therefore take the mean as our summary measure, or factor score. We then have

$$E(\boldsymbol{y}|\boldsymbol{x}) = \boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}x \tag{6.4}$$

Where here, and subsequently, $\boldsymbol{x}$ is assumed to be standardised with mean zero. An equivalent and more convenient, version of these factor scores is given by

$$E(\boldsymbol{y}|\boldsymbol{x}) = (\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} + \boldsymbol{I})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{x} \tag{6.5}$$

Bearing in mind the diagonality of the first matrix on the right hand side of (6.5), these scores are proportional to the elements of the vector $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{x}$. Shortly we shall discover that there are many other problems of a similar kind which produce scores which are linear in the x s like this and we shall refer to them generically as components. This is by analogy with principal components, which are also linear functions and serve a similar purpose, but the two are quite distinct.

When dealing with the 'fixed effects' version of the data matrix in the last chapter, we found that the row and column totals were jointly sufficient for the respective row and column parameters. A somewhat similar result holds for components if we suppose that, instead of being restricted to normal distributions, the xs have the more general conditional distributions of the exponential form.

Thus suppose that x_{ij} has a distribution in the exponential family of the form

$$f\left(x_{ij}|\theta_i\right) = F_i\left(x_{ij}\right)G_i(\theta_i)exp\theta_i x_{ij} \tag{6.6}$$

where each θ_i is a function of $\boldsymbol{y}$. The ys are, of course, hypothetical variables so we suppose that they are defined so as to exert their influence through a linear function. That is we suppose that

$$\theta_i = \alpha_{i0} + \alpha_{i1}y_1 + \alpha_{i2}y_2 + \cdots + \alpha_{iq}y_q. \tag{6.7}$$

The notation used in (6.7) expresses the fact that the distribution of x_{ij} depends on which row it belongs to but not on the column. The conditional distribution of $\boldsymbol{y}$ given the xs is thus, according to Eq. (6.1),

$$h(\boldsymbol{y}|\boldsymbol{x}) = h(\boldsymbol{y})\frac{\prod_1^p \prod_1^n f\left(x_{ij}|\theta_i\right)}{f(\boldsymbol{x})}. \tag{6.8}$$

Substituting from Eq. (6.6) we then find that

$$h(\boldsymbol{y}|\boldsymbol{x}) \propto h(\boldsymbol{y}) \prod_1^n G_i(\theta_i) \exp \sum_i \sum_j x_{ij}\theta_i \tag{6.9}$$

Since θ_i is assumed to be a linear function of the ys, we may substitute the expression given in Eq. (6.7) for θ_i, obtaining

$$\sum_i x_{ij}\theta_i = \alpha_{i0} \sum_i x_{ij} + y_1 \sum_i x_{ij}\alpha_{i1} + y_2 \sum_i x_{ij}\alpha_{i2} + \cdots + y_q \sum_i x_{ij}\alpha_{iq}. \tag{6.10}$$

If we write

$$X_l = \sum_j \sum_i x_{ij}\alpha_{il} \quad (l = 1, 2, \ldots q). \tag{6.11}$$

We note that $h(\boldsymbol{y}|\boldsymbol{x})$ as given by Eqs. (6.8) or (6.9) depends on $\boldsymbol{x}$ **only** through the linear functions X_l given by (6.11). It does not follow, of course, that the expected value of $\boldsymbol{y}$ is a linear function of the Xs (except in the normal case) but that it will not depend on any other function of the xs.

6.2 Thomson's Scores and Bartlett's Scores

It is common when defining factor scores for the normal model, to say that there are two families of scores without giving any insight into the reasons why they are often not the same. The two families are known as Thomson's scores and Bartlett's scores. The issue is further confused by the fact that the two are usually close and, sometimes, equivalent. In reality they provide solutions to rather different problems as we shall now show.

The scores arrived at using the approach of this chapter, as given by Eq. (6.5), are in fact Thomson's scores. The conditional expectations which we have derived are also known as the regression scores. Although Thomson derived them in a fairly straightforward way using the standard theory of least squares regression, Thomson, himself, would not have understood the significance of the regression terminology. Nevertheless, his derivation and ours are essentially equivalent. Thomson had considerable correspondence with Bartlett about factor scores but neither seems to have recognised, what is evident to us, albeit with the benefit of hindsight. Bartlett was essentially showing how to estimate a set of parameters whereas Thomson was aiming to predict the value of a random variable, as we now show.

The crucial difference between Thomson's approach and Bartlett's is that Bartlett treats the unknown values of the factors as parameters. Bartlett's proposal was to find the best linear unbiased estimators of those parameters. These are the same as the maximum likelihood estimators which are obtained by maximising the log likelihood and this amounts to minimising

$$(\boldsymbol{x} - \boldsymbol{\Lambda y})' \boldsymbol{\psi}^{-1} (\boldsymbol{x} - \boldsymbol{\Lambda y})$$

with respect to $\boldsymbol{y}$. Equating the derivative with respect to $\boldsymbol{y}$ to zero we obtain

$$-2\boldsymbol{\Lambda}'\boldsymbol{\psi}^{-1}(\boldsymbol{x}-\boldsymbol{\Lambda}\boldsymbol{y}) = 2(\boldsymbol{\Gamma}\boldsymbol{y}-\boldsymbol{\Lambda}'\boldsymbol{\psi}^{-1}\boldsymbol{x})$$

where $\boldsymbol{\Gamma} = (\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}+\boldsymbol{I})^{-1}$ and hence

$$\boldsymbol{y} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}'\boldsymbol{\psi}^{-1}\boldsymbol{x} \tag{6.12}$$

Equation (6.12) differs from Eq. (6.5) only by the fact that the factor $(\boldsymbol{I}+\boldsymbol{\Gamma})^{-1}$ in Eq. (6.5) is replaced by $\boldsymbol{\Gamma}^{-1}$ in Eq. (6.12). If $\boldsymbol{\Gamma}$ is diagonal the two sets of scores differ only by a scale factor and in that sense they are equivalent.

The misunderstandings which have arisen over these two approaches to factor scores centre on the distinction between parameters and random variables. The failure to make this distinction has led to the scores being evaluated in wholly inappropriate ways. For example, Lawley and Maxwell (1973, p. 113) say that "No general preference can therefore be given" on the grounds that the choice depends on a trade-off between bias and precision. But the unbiassedness of the Bartlett scores is a consequence of treating $\boldsymbol{y}$ as fixed and $\boldsymbol{x}$ as a random variable and the greater precision of the regression scores by lies in reversing the roles of $\boldsymbol{x}$ and $\boldsymbol{y}$. The real question is whether we want to *estimate* the value of $\boldsymbol{y}$ for a given set of individuals or to *predict* the values of $\boldsymbol{y}$ for a random sample of individuals drawn from some population. These are two distinct questions and it is fortunate that for practical reasons that their answers are often so similar. A fuller treatment of the relationship between the two sets of scores will be found in Bartholomew et al. (2009).

This confusion reflects a wider misunderstanding which is prevalent throughout the extensive literature on factor scores which is to be found mainly in the psychological publications. It results from a failure to notice that the concept of a random variable as used in probability and statistics, is not the same as what we may term a 'mathematical' variable such as occurs in everyday algebraic expressions. No doubt, this is partly because it is cumbersome to distinguish between them notationally. In fact, we have failed to make that notational distinction completely clear here, in the interests of greater simplicity, but we have compensated for this omission by using forms of words intended to make the distinction clear.

References

Bartholomew, D. J., Deary, I. J., & Lawn, M. (2009). The origin of factor scores: Spearman, Thomson and Bartlett. *British Journal of Mathematical and Statistical Psychology, 62*, 569–582.

Lawley, D. N., & Maxwell, A. E. (1973). *Factor analysis as a statistical method* (2nd ed.). London: Butterworths.

Chapter 7
Identifiability

Abstract Even if there is close agreement between a model and the data it does not follow that the model provides a true account of how the data arose. It may be that several models explain the data equally well. When this happens there is said to be a lack of identifiability. Failure to take full account of this fact, especially in the social sciences, has led to many over-confident claims about the nature of social reality. Lack of identifiability within a class of models may arise because different values of their parameters provide equally good fits. Or, more seriously, models with quite different characteristics may make identical predictions. Both kinds of lack of identifiability are common where the observations are incomplete, or latent, and several examples are given in this chapter. One arises in the field of intelligence measurement where our analysis shows that the common assumption that Spearman's *g* corresponds to a physical reality in the brain is not necessarily true. A second concerns the widespread use of latent structure models where, since it may be exceedingly difficult to determine a satisfactory metric for the latent variables, little confidence can be placed in any results of the analysis.

Keywords Bonds · Brain · Endogenous variables · Exogenous variables · Factor model · Social science · Intelligence · Measurement · Spearman's *g* · Latent structure models (LISREL)

7.1 The Meaning of Identifiability

We are using *identifiability* as an umbrella term to cover several different concepts but they all have a common principle underlying them. This is that they involve an asymmetry between a model and the data. If we start with a model we can predict, albeit uncertainly, what data it should generate. But if we are given a set of data we cannot necessarily infer that it was generated by a particular model. In some cases it may, of course, be possible to achieve identifiability by increasing the sample size but there are cases in which, no matter how large the sample size, no

D. J. Bartholomew, *Unobserved Variables*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-39912-1_7,

separation is possible. It is that situation with which we are particularly concerned in this chapter.

Identifiability matters can be considered under three headings. First there is lack of *parameter identifiability* which is the most common use of the term. This refers to the situation where there is more than one value of a parameter in a given model each of which gives an equally good account of the data. For example, if we are aiming to estimate a parameter by the method of maximum likelihood, it may happen that the likelihood function has exactly the same value for two different values of the parameter—or, more likely, that the likelihood function is flat in some region of the parameter space.

Secondly there is what we shall call lack of *model identifiability* which occurs when two or more models make exactly the same data predictions. This has serious practical implications because, if we find that there are at least two explanations for our data, the whole basis of induction by statistical modelling is compromised. We shall give an important example of model unidentifiablity later in this chapter.

The third type of identifiability is actually the combination of the foregoing types.

Mathematical statistics is not well-equipped to cope with situations where models are practically, but not precisely, indistinguishable because it typically deals with things which can only be expressed in unambiguously stated theorems. Of necessity, these make clear-cut distinctions which do not always correspond with practical realities. For example, there are theorems concerning such things as *sufficiency* and *admissibility.* According to such theorems, for example, a proposed statistic is either sufficient or not sufficient for some parameter. If it is sufficient it contains all the information, in a precisely defined sense, about that parameter. But in practice we may be much more interested in what we might call 'near sufficiency' in some more vaguely defined sense. Because we cannot give a precise mathematical definition to what we mean by this, the practical importance of the notion is easily overlooked. The same kind of fuzziness arises with what are called structural eqation models (or structural relations models) which have played a very important role in the social sciences. In Sect. 7.4 of the present chapter we shall argue that structural equation models are almost always unidentifiable in the broader sense of which we are speaking here. This has far-reaching implications for social research.

7.2 The Factor Model

A well-known example of non-identification of parameters arises in the factor model if there are at least two factors. This can easily be demonstrated using the standard equation given in Eq. (5.4). We start with the model in the form

$$\boldsymbol{x} = \boldsymbol{\Lambda y} + \boldsymbol{e} \tag{7.1}$$

If we introduce an orthogonal matrix $\boldsymbol{M}$ (implying that $\boldsymbol{M}'\boldsymbol{M} = \boldsymbol{I}$) and write

$$\boldsymbol{x} = \boldsymbol{\Lambda}\boldsymbol{M}'\boldsymbol{M}\boldsymbol{y} + \boldsymbol{e} \tag{7.2}$$

This model is equivalent to that of Eq. (7.1) but its loading matrix is now $\boldsymbol{\Lambda}\boldsymbol{M}'$ and the factors have become $\boldsymbol{M}\boldsymbol{y}$. This is called a rotation because that is what it is if we look at the geometry of the transformation. The set of loadings generated by the new model has exactly the same likelihood and so cannot be distinguished; this means the model is not identifiable. In effect we have an infinite set of equivalent solutions. If a choice has to be made among them it has to be based on substantive and not statistical considerations (unless, of course, one can propose some objective criterion to make the rotation unique, but this is extremely rare in practice). There is a substantial literature on how non-stistical considerations might be brought to bear on this matter. For future reference it should be noted that rotation does not depend on whether or not the number of factors is given or whether it has to be estimated from the data.

7.3 The g-Factor and Bonds

Ability testing and intelligence testing in particular have generated an enormous literature and considerable controversy. It is therefore highly desirable to have a satisfactory statistical account of the situation so that inferences, about heritability, for example, are securely based. The somewhat disconcerting fact is that there are two statistical models which both give an equally good description of the situation and, therefore, provide a perfect illustration of, model unidentifiability.

We begin by giving a brief description of the essentials of the two models associated with the names of Charles Spearman and Godfrey Thomson, the latter of whom was mentioned in connection with factor scores in the last chapter.

At the root of the whole matter is the empirical fact that individuals who perform well in one mental test tend to perform well in other similar tests. Put another way, the correlations among test scores tend to be positive—the so-called positive manifold. Why does this happen?

Spearman attempted to explain this fact by supposing that a person's test score was the sum of two parts. The first part reflected their general ability and the second part was a contribution specific to that particular test. Scores on any two tests would therefore be correlated because the person's general ability was common to the scores obtained on both tests.

Using modern notation and terminology Spearman proposed a one-factor model which, in the notation we have used hitherto, may be written

$$x_i = \mu + \lambda_i y + e_i \tag{7.3}$$

where x_i is the score on the ith test, y is the position of the individual on the scale of ability, $\lambda_i y$ is the contribution which that individual makes to the test score and

e_i is the 'error'. In Spearman's terminology, $\mu + \lambda_i y$ represents the general ability of the individual and e_i the 'specific factor'. The use of the term 'factor' here is out of line with much modern usage, but it accounts for the fact that Spearman called his model the 'two-factor model' whereas it would now be called a 'one-factor model'. Also it is obviously a special case of the general factor model considered in Chap. 5. Spearman used the symbol g instead of y in Eq. (7.3) to designate what he called the general factor. Thus was because he did not wish to pre-judge any issues by referring to it as general intelligence; he therefore preferred a less specific description. For this reason the quantity is referred to as the 'g-factor' or Spearman's g.

Thomson proposed an alternative, and very different, explanation. He supposed that a person's brain contained a number of what he called 'bonds'. When a person attempted a test, a random selection of bonds was activated. The resulting test score was supposed to be the sum of the contributions from the selected bonds. On attempting a second test the selection of bonds would include some of those used for the first test. The correlation was supposed to arise because those common bonds would make the same contribution to the score as in the two tests.

A comparison of the two models was made in Bartholomew et al. (Bartholomew 2009a) where Thomson's model was also expressed in modern notation. This showed that the two models made exactly the same statistical predictions and therefore they could not be distinguished on statistical grounds. It was also concluded that the biological evidence did not unequivocally favour either model. The question, still unresolved, is: which of the two models more accurately represents what goes on in the brain?

One statistical possibility for distinguishing between the two models is to see what happens when we look beyond the first factor. Although the positive manifold is practically universal in ability testing, it is not true that the one factor model provides a complete description of the data. The g-factor usually accounts for much of the correlation structure but not all of it. The fit can often be improved by introducing a second factor (and possibly a third and so on). This would prove to be an advantage over Thomson's model if the latter could not be extended in a natural way to accommodate such deviations. There is such an extension, which we describe briefly below which suggests that the lack of identifiability which we shall uncover is quite general.

The generalisation of the bonds model needed for this purpose was given in Bartholomew et al. (2013) and it depends of the idea of what is there called a 'pass'. It is supposed that, when attempting a test item, the brain makes a number of passes through the items focussing on a different aspect at each pass. Thus, for example, it may focus on the quantitative aspects of the items at the first pass and the number of bonds required for each item will therefore reflect the quantitative aspects of that item. At the second pass the brain may be focussing on, say, the verbal aspects of each item and this will require more or fewer bonds according to the amount of verbal content of the items. The neural basis for the idea of successive 'passes' is, on current evidence, neither stronger nor weaker than that for

the existence further dimensions in Spearman's model. Both models are therefore on an equal footing on that score.

In order to show that the bonds model with several passes produces a correlation structure which is identical with that of the factor model it is only necessary to consider a single pass, because results for a multi-pass model can be obtained from it by a process of aggregation. The score obtained from a single pass may be written

$$x_i = a_{1i}e_{1i} + a_{2i}e_{2i} + \cdots a_{Ni}e_{Ni} \tag{7.4}$$

where the as are indicator variable each indicating whether or not the ith bond is active. N is the number of bonds in the brain (this need not be the same for each person) and e_{ji} is the contribution which the jth bond makes to the score of that individual. It is supposed that p_{ij} is the probability that the jth bond is used when attempting the ith item. Starting from these simple assumptions it is possible to determine the covariance and hence the correlations between any pair of xs and the variance of any individual x. If there are several independent passes the corresponding covariances and variances are obtained by adding up the results for the constituent passes. As a result it is easy to show that, for k passes

$$\text{Corr}(x_i,\ x_j) = \frac{\sum_r p_{ri}p_{rj}}{\sqrt{\sum_r p_{ri}\sum_r p_{rj}}} = \sum_r \left(\frac{p_{ri}}{\sqrt{\sum_r p_{ri}}}\frac{p_{ri}}{\sqrt{\sum_r p_{ri}}}\right). \tag{7.5}$$

This has exactly the same form as the off-diagonal elements of the corresponding matrix for Spearman's model. This immediately recognisable if we write

$$\lambda_{ri} = \frac{p_{ri}}{\sqrt{\sum_r p_{ri}}}. \tag{7.6}$$

This analysis shows that for every bonds model with k passes there is a factor model with exactly the same correlation structure. The converse is not necessarily true which means that there is not a bonds model corresponding to every possible factor model. This is not surprising because factor models have a much wider range of applicability, but this correspondence has been shown empirically to exist for all ability matrices so far examined.

7.4 Linear Structural Equation Models

Linear structural relations models are a generalisation of the factor model which involve linear relations among the latent variables. They are sometimes referred to as covariance structure models because they specify the structure of the covariance matrix; the parameters of the model are estimated by fitting the theoretical matrix to that which is observed. Although the models are a generalisation of the linear factor model, they are a rather special kind of generalisation. In the first place, the

number and identity of the latent variable is assumed to be known in advance, so there is no need to estimate their number or to use rotations to select the most meaningful. The models imply a fairly well-developed theory in which the variables have been already specified. In this sense they are more naturally thought of as developments of models for ability, where we knew what the single latent variable was at the outset. Secondly, the latent variables are of two kinds, known as endogenous and exogenous. These terms are well-known in econometrics. They derive their meaning from the notion of an 'inside' and an 'outside' to the system being studied. The structural part of the model concerns the relationship between the endogenous (internal) and the exogenous (external) variables which relationships, in the standard model, are assumed to be linear. Before specifying the model we need to pause to make an observation on notation. Any attempt to present statistical methods from a more general and unified framework runs into the problem of using a notation which is both internally consistent and also consistent with the published literature. This is impossible in the present instance. Structural relations modelling has developed as a fairly self-contained subject using a notation which is firmly established but at variance with common statistical usage. In Bartholomew et al. (2008) the balance of advantage lay with adopting the notation usual in the field, but in Bartholomew et al. (2011) we judged that the advantage lay in the other direction. Here, because we wish to emphasise the link with factor analysis and other latent variable models, I have followed the usual statistical conventions, that is, Roman letters are used to denote variables and Greek letters to denote constants—that is, parameters.

Let y denote an endogenous latent variable and z an exogenous variable; as before, subscripts will be used to distinguish one variable from another and bold type to denote vectors. The core of a structural relations model expresses the relationship between the ys and the zs as follows,

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{y} + \boldsymbol{C}\boldsymbol{z} + \boldsymbol{f} \tag{7.7}$$

where $\boldsymbol{f}$ is now a vector of error terms. We do not need to specify the dimensions of the various vector and matrices for our very limited purpose. A slightly curious feature of (7.7) is the appearance of $\boldsymbol{y}$on both sides of the equation. This reflects the difference between endogenous and exogenous variables. Variables external to the system do not affect one another but those inside may do so.

The second part of the model, often called the measurement model, links all the latent variables to observable variables, or indicators. A special feature of the standard model is that the endogenous and exogenous latent variables have distinct indicators. This assumption imposes a special structure on the usual factor model which covers both sets of variable. If all the indicators are collected together in a single vector $\boldsymbol{x}$ the model has the same form as in (7.1); but if we designate the separate parts by adding to the various matrices the subscript y or z, the measurement model may be written

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{\Lambda}_y & 0 \\ 0 & \boldsymbol{\Lambda}_z \end{pmatrix} \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix} + \begin{pmatrix} \boldsymbol{e}_y \\ \boldsymbol{e}_z \end{pmatrix}. \tag{7.8}$$

From Eqs. (7.7) and (7.8) we can easily determine the covariance matrix of $\boldsymbol{x}$ and so proceed to fit the model by minimising some measure of the distance between the observed and predicted covariance matrices. It is at this stage that questions of identifiability arise.

The identifiability issue is more subtle than at first sight appears, especially if we break it up, as we shall and deal with parameter and then model and identification in turn.

Being a generalisation of the factor model the lack of identification evident there is also present in the case of the structural relations model. Thus in Eq. (7.7) we can replace C by CM' and $\boldsymbol{z}$by $\boldsymbol{Mz}$ without changing the model (doing the same with $\boldsymbol{y}$ is less straightforward, but does not affect the point being made). However, making such a change contradicts the starting assumption that we already know what the latent variables are.

The question of parameter identifiability is much more important. The introduction of the relationship between the latent variables introduces many additional variables and we clearly run the risk of not being to estimate them all. This topic has been the subject of intensive investigation and a full discussion, with various, incomplete, tests being given, for example in Bollen (1989), especially pp. 88–104). Perhaps the simplest of these arises directly from the fact that, if the estimates are obtained by solving equations, there must be at least as many equations as there are parameters to be estimated.

Model identifiability has hardly been looked at, and deserves much more attention. There is one example in Bartholomew et al. (2011, Sect. 8.10). This is a simple example specially constructed to demonstrate non- identifiability. It shows that a structural model with two categorical latent variables is statistically indistinguishable from one with two continuous latent variables. It is not known how easy it would be to construct other examples but the ease with which this one was found suggests that it may not be too difficult. It may be objected that one would not use a structural model unless one knew what the latent variables were and, in particular, whether they were continuous or discrete. However the force of this argument is much diminished by the following remark I wish to make about identifiability.

This further point, which is not easy to express in precise mathematical terms, could be particularly damaging to the whole enterprise. Briefly stated it is that prior distributions (the assumed forms for the latent variables) are poorly determined by the data. The reason for this is illustrated by what we learned in Chap. 3 about the information in the data about single latent variable. Essentially, we are dealing here with a mixture model of the kind we met in Chap. 3. This is because a latent variable model is a mixture model with the mixing distribution being equivalent, in the present context, to the prior distribution of the latent variable. The most extreme example of the phenomenon we wish to illustrate is provided by the result on a normal mixture of normals. We showed in Chap. 3 that if we are presented with a normal distribution it is impossible to know whether or not it has

been generated as a mixture and if so, what the mixing distribution was. A similar situation arises at the empirical level with mixtures of exponentials. We also noted in Chap. 3 that the distribution resulting from a mixture of exponentials had very similar characteristics, whether the mixing distribution was a continuous gamma distribution or a two-point discrete distribution. A striking example of essentially the same point in latent variable modelling is provided by an example in which a latent class model, with two classes, was fitted to the same empirical distribution (the Law School Admission Test data), see Bartholomew et al. (2011, especially Table 6.1, p. 160 for an alternative comparison making the same point) as a latent trait model with a normal prior distribution for the latent variable. The two fits were hardly distinguishable, which means that there could be no empirical evidence favouring either prior even though they are so radically different. If it is not even possible to know whether the prior is continuous (i.e. normal) or two-point discrete, it seems over-ambitious to formulate and estimate models which, it is assumed, involve linear relationships between variables whose own status is so poorly defined.

When all these results are brought together they constitute a formidable argument against the careless use of structural relations models. If some parameters are non-identifiable, the variables with which they are associated will probably be dropped from the model but it is not easy to reconcile this with the original judgement that they should be included. Even if we are sure that a variable should be included we need a good deal of prior knowledge to ensure that it is indeed continuous and normal. for this is something the data cannot tell us. In brief, the valid use of a structural equations model requires us to lean very heavily upon assumptions about which we may not be very sure. It is undoubtedly true that if such a model provides a good fit to the data, then it provides a *possible* account of how the data might have arisen. It says nothing about what other models might provide an equally good, or even better fit. As a tool of inductive inference designed to tell us something about the social world, linear structural relations modelling has very little to offer.

References

Bartholomew, D. J., Steele, F., Moustaki, I., & Galbraith, J. I. (2008). *Analysis of multivariate social science data* (2nd ed.). London: Chapman and Hall/CRC Press.

Bartholomew, D. J., Deary, I. J., & Lawn, M. (2009a). A new lease of life for Thomson's bonds model. *Psychological Review, 116*, 567–579.

Bartholomew, D. J., Deary, I. J., & Lawn, M. (2009b). The origin of factor scores: Spearman, Thomson and Bartlett'. *British Journal of Mathematical and Statistical Psychology, 62*, 569–582.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis* (3rd ed.). Chichester, UK: Wiley.

Bartholomew, D. J., Deary, I. J., & Allerhand, M. (2013). Measuring mental capacity: Thomson's bonds model and Spearman's g-model compared. *Intelligence, 41*(4), 222–233.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Chapter 8
Categorical Variables

Abstract At the conceptual level continuous variables and categorical variables need not be distinguished but, at the practical level, the form which the analysis of latter takes needs to be spelt out. This is done in the present chapter where categorical variables appear in the standard data matrix. The key step is the replacement of a continuous variable by an indicator vector showing into which of a number of categories a sample member falls. In some cases more information is available in the shape of an ordering of the categories. This can be accommodated by introducing a further kind of unobserved hypothetical variable which is assumed to induce the ordering of the categories. The analysis can then be carried out as if these hypothetical variables had actually been observed. The same idea can be extended to other situations and the chapter concludes with one such example where it is assumed that there is an underlying model in continuous variables for which only categorical observed variables are available. This also provides another example of the lack of identifiability discussed in Chap. 7.

Keywords Binary data · Identifiability · Ordered categories · Indicator variables (vectors) · Random effects · Purchasing behaviour

8.1 The Role of Categorical Variables

It is a curious feature of the development of statistical methods that the analysis of categorical data has lagged behind that for continuous data. This may owe more to the accidents of history than the inner logic of the subject. Biological data of one kind or another were the raw material of much of the early work of Karl Pearson and R A Fisher on correlation and the analysis of variance was largely, but not entirely, continuous and this fact dictated the direction that developments took. Yet in many ways categorical data is simpler in the sense that it makes weaker assumptions about measurement. It was relatively late in the development of statistical methods that books began to appear on the analysis of categorical data,

D. J. Bartholomew, *Unobserved Variables*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-39912-1_8, © The Author(s) 2013

as though that were a primary classification. The point of view of this Brief is quite opposed to this view of the subject. As far as the conceptual framework of the subject is concerned, whether or not variables are continuous or discrete is a secondary matter. Accordingly, our main purpose in devoting a special chapter to this topic, is to show how it fits naturally into the modelling framework which underlies our treatment. But, in the course of doing this we shall need to introduce another kind of unobserved variable. The most comprehensive treatment of categorical variables from the traditional viewpoint will be found in Agresti (2013).

It is already clear from Chap. 5, where distributions belonging to the exponential family were at the centre, and that the character of the variables was not an issue. The exponential family includes both discrete and continuous distributions as we have already noted and the results obtained apply equally to both kinds of variable. We have already met the simplest kind of categorical variable in the shape of binary data and in this chapter we shall go beyond this to cases where there are more than two categories.

8.2 Unordered Categories

If we start with the problem of Chap. 4 we can return to the standard data matrix which we supposed to have arisen if n persons each take m tests. The entries in the table, x_{ij}, were supposed to be the scores obtained by the ith person on the jth test. In that chapter we started with the usual analysis of variance assumption that the scores had normal distributions with means that depended only on the row and column. We next went on to suppose that the x_{ij}s had a Bernoulli distribution. In this chapter we suppose that each x_{ij} is a categorical variable which records into which of c_j categories the ith member falls. On that supposition each entry may be replaced by a vector and the parameter of the exponential family distribution also becomes vector valued. Once this is done, the results of Chap. 5 carry over with relatively little change.

A typical score records into which category the member falls and this may be done by replacing x_{ij} by an indicator vector which has a 1 in the position corresponding to the category number with zeroes elsewhere. Thus, for example, the indicator vector would be $\boldsymbol{x}_{ij}(r) = (0, 0, \ldots 1 \ldots 0)'$,with 1 in the rth position, if the individual in cell (i,j) of the data matrix fell into the rth category. We further define $\pi_{ij}(r)$ as the probability that the indicator variable takes this value. It is important to note that binary data is included as a special case when there are only two possible categories. In that case $\pi_{ij}(2) = 1 - \pi_{ij}(1)$ because the two probabilities must sum to 1. This makes the designation of the category number redundant and we then simplify the notation by writing $\pi_{ij}(1) = \pi_{ij}$.

In the general case one of the $\pi_{ij}(r)$'s is redundant and so we define $\pi_{ij}(c) = 1 - \pi_{ij}(1) - \pi_{ij}(2) - \ldots - \pi_{ij}(c-1)$. The frequencies can be arranged in a $c \times m$

contingency table. The number in the (r, j)th cell is $\sum_i x_{ij}(r)(r)$; the rth row total is $\sum_j \sum_i x_{ij}(r)$ and the jth column total is $\sum_r \sum_i x_{ij}(r)$.

Armed with these definitions, we can write down the likelihood function for the data in the table as follows.

$$l = \prod_i \prod_j \pi_{ij}(1)^{x_{ij}(1)} \pi_{ij}(2)^{x_{ij}(2)} \ldots \pi_{ij}(c-1)^{x_{ij}(c-1)} (1 - \pi_{ij}(1) - \pi_{ij}(2) - \ldots \pi_{ij}(c-1)^{x_{ij}(c)} \tag{8.1}$$

Recalling that, by definition, $x_{ij}(c) = 1 - x_{ij}(1) - x_{ij}(2) - \ldots x_{ij}(c-1)$ we may write the log likelihood as

$$L = \log l = \text{constant} + \sum_i \sum_j \sum_r x_{ij}(r) \log \pi_{ij}(r)/\pi_{ij}(c). \tag{8.2}$$

If we now assume an additive model we shall have that

$$\log \pi_{ij}(r)/\pi_{ij}(c) = \mu + \alpha_{ij} + \beta_{ir}. \tag{8.3}$$

In this case, the likelihood depends on the data only through the expressions

$$X_j(r) \text{ and } X_i(r) \tag{8.4}$$

where $X_i(r) = \sum_j x_{ij}(r)$ and $X_j(r) = \sum_i x_{ij}(r)$. The first of these quantities is the total number of times the ith individual falls into the rth category and the second is the total number of individuals who fall in the rth category taken across all variables. If we were to think of the data as summarised in a $c \times m$ contingency table (persons against items) the sums $\{X_j(r)\}$ would be the item totals which this analysis shows are jointly sufficient for the item effects. This is the same result as we found for binary data where, in effect, we were dealing with a $2 \times m$ contingency table. The set of totals over items, $\{X_i(r)\}$ is likewise jointly sufficient for the person effects.

8.3 Random Effects (Items)

When we moved from the fixed effects to the random effects model in Chap. 4 we utilised the fact that a simple model which allowed the item effect to depend linearly on q latent variables was readily available if the distribution if the typical cell entry had a distribution belonging to the one-parameter exponential family. This result is easily extended to the case where the variable in question and the parameter θ is vector valued. This is all that we need because the multinomial probability distribution is a member of this extended family. To see this we only need to observe that the multinomial probability

$$\pi_{ij}(1)^{x_{ij}(1)}\pi_{ij}(2)^{x_{ij}(2)}\ldots\pi_{ij}(c-1)^{x_{ij}(c-1)}\ (1-\pi_{ij}(1)-\pi_{ij}(2)-\ldots-\pi_{ij}(c-1)^{x_{ij}(c)} \tag{8.5}$$

is proportional to

$$\exp\left[\sum\nolimits_r x_{ij}(r)\mathrm{og}\{\pi_{ij}(r)/\pi_{ij}(c)\}\right]$$

which, in turn, may be written

$$\exp\left\{\boldsymbol{x}'_{ij}\,\boldsymbol{\theta_{ij}}\right\} \tag{8.6}$$

where $\boldsymbol{x_{ij}}$ is the c-vector with elements $x_{ij}(r)$ and $\boldsymbol{\theta_{ij}}$ has elements given by

$$\theta_{ij}(r) = \log\{\pi_{ij}(r)/\pi_{ij}(c)\} \tag{8.7}$$

The model now has exactly the same form as we found in the unidimensional case except that the scalars are now replaced by vectors. Everything goes through as before if we make the appropriate transformations from scalar quantities to vectors and matrices. Thus we write

$$\boldsymbol{\theta}(r) = \boldsymbol{A}(r)\boldsymbol{y} \tag{8.8}$$

where $\boldsymbol{A}(r)$ is a matrix of coefficients $\{\alpha_{ij}(r)\}$. By going back to Eq. (8.7) we see that, together with Eq. (8.8), this equation specifies a linear model for the probabilities $\{\pi_{ij}(r)\}$.

The posterior distribution of $\boldsymbol{y}$ is obtained by multiplying the prior distribution by the likelihood whose logarithm is given by Eq. (8.2). On substituting for $\boldsymbol{\theta}(r)$ it is evident that this posterior distribution depends on the data only through the quantities $\boldsymbol{x}'_{ij}\boldsymbol{A}(r)$ where we recall that $\boldsymbol{x_{ij}}$ is a vector containing c elements $\{x_{ij}(r)\}$. These quantities are therefore sufficient for $\boldsymbol{y}$ in the sense of Chap. 4.

In arriving at this solution we encounter another example of non-identifiability, or rotational invariance. If we insert into Eq. (8.8) the identity matrix resulting from multiplying an orthogonal matrix $\boldsymbol{M}$ by its transpose $\boldsymbol{M}'$ we see that $\boldsymbol{\theta}(r)$is unchanged if $\boldsymbol{A}(r)$ is replaced by $\boldsymbol{A}(r)\boldsymbol{M}$ and $\boldsymbol{y}$ by $\boldsymbol{M}'\boldsymbol{y}$. There is therefore no unique value of the random effect $\boldsymbol{y}$ which explains the row differences.

8.4 Ordered Categorical Data

In our treatment of categorical data no assumption was made about the ordering of the categories, but often, in practice, there is additional information about how the categories stand in relation to one another. If such information is available it ought to be used to give the methods greater efficiency. The question does not arise with

binary data because two categories can always be thought of as ordered, but the more categories there are, the more there is to be lost by ignoring the ordering.

One inefficient method is to reduce all categorical variables to binary form by amalgamating categories. In practice this strategy may lose less efficiency than might appear at first sight. This is because it is quite common in practice to have several sparsely populated categories so little is lost by merging them with larger categories. Nevertheless, this does not apply universally and methods are needed to deal with the general case.

The commonest strategy for dealing with ordered categories introduces us to another type of unobserved variable. This arises by imagining that the categories have been formed by grouping values of a continuous variable. In some cases this may be exactly what has happened but, usually, the imagined underlying variable is hypothetical. In a real sense it is a latent variable but it would be confusing to use the same term as has become established in latent variable modelling. We shall therefore speak of *underlying variables* in this context.

The information about any underlying variable which the categorization yields is, of course, very crude. We may have several hundred individuals allocated to only three or four categories, so what we have is a very crude ranking with extensive ties. The prospect of progress is offered by the fact that it may be possible to estimate the correlation coefficients between the underlying variables from the grouped data. The simplest case arises when all the classifications consist of only two categories, that is simple dichotomies. We do not need any new theory here because this case is already covered by the model of the early part of this chapter relating to binary data. The point of mentioning it is that the method we are about to propose coincides with the earlier method in this case. This is a remarkable fact which suggests that the two approaches are not so far removed from each other as the different formulations of the models might suggest.

If two continuous variables have a normal bivariate distribution, they may be reduced to two binary variables by recording only whether each variable is above or below some threshold value. The result is a 2×2 contingency table. Given only this table, it is possible to estimate what the correlation coefficient of the underlying bivariate normal distribution must have been. This estimate is known as the *tetrachoric correlation* (because there are four categories). There are tables and computer programs which enable this to be calculated. If there are more than two categories for one or both variables the corresponding coefficient is known as the *polychoric correlation*. If we go on to treat these coefficients as if they were product moment correlations, we can proceed exactly as we did with the continuous normal model. This treatment supposes that all of the variables are categorical. Sometimes we have a mixture of continuous and categorical variables. Similarly, the product moment correlation between a continuous variable and a categorical variable can be estimated by a coefficient known as a polyserial coefficient.

All of the foregoing methods appear to rest on the assumption that underlying the categories there is a multivariate normal distribution. In fact what we are really doing is to carry out the analysis *as if* there really were underlying variables

measured in such a way as to render their joint distribution multivariate normal. This is not quite the same thing. If the underlying variables were treated as if they had some other distribution, the correlation coefficients would no longer be the appropriate summarisations of the distributions. The full implications of all this have not been worked through, but the introduction of underlying variables in this fashion enables information about ordering to be taken into account and practical experience suggests that they are useful. In the present state of knowledge these methods are best regarded us useful exploratory techniques.

In order to fit the model one can maximise the likelihood but there are various approximations, given in Bartholomew et al. (2011) which are of more interest, perhaps, for the light they throw on the relationships between this model and principal components analysis and correspondence analysis.

8.5 An Alternative Underlying Variable Model for Ordered Categorical Data

The more complicated the structure of our data, the greater the variety of models which become possible. The fact that the foregoing models have proved useful does not exclude the possibility that there might be other models which gives an equally good account of the data. Here we shall describe one such model which, in one special case, coincides with one of the models we have just considered. This serves to identify yet another example of lack of model identifiability.

This model is likely to be relevant when the categories are, in a certain sense, in competition with one another. Thus if a candidate in an examination faces a multiple choice question the possible answers may be thought of as in competition for selection as the correct one. Similarly, a shopper in a supermarket may be faced with a variety of brands of a commodity which are in competition for selection or purchase. Categorical data arises in such applications by giving the numbers of answers, or purchases, falling into each category. The pattern of responses may be explained by a latent variable model of the following kind. More information about this model may be found in Bartholomew et al. (2011, Sect. 8.9).

Suppose each item has an 'attractiveness' which relates to those qualities which contribute to its overall attractiveness according to the following model.

$$\boldsymbol{z_i} = \boldsymbol{\mu_i} + \boldsymbol{\Lambda_i y} + \boldsymbol{e_i}. \tag{8.9}$$

Individuals are indexed by i and $\boldsymbol{z_i}$ is the vector each of whose elements is the attractiveness of the items for the ith person. That person then selects the category which is the most attractive. One can make the usual assumptions about the random variables appearing in Eq. (8.9) and fit the model to categorical data arising from an examination or purchase records. There are circumstances under which this model is indistinguishable from the first model described previously. These occur when the error terms in Eq. (8.9) are assumed to have an 'extreme

value' distribution. This rather unusual characteristic is less surprising when we notice that in selecting the most attractive choice the subject is choosing an extreme value.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken: Wiley.
Bartholomew, D. J., Knott, M., & Irini, M. (2011). *Latent variable models and factor analysis* (3rd ed.). Chichester: Wiley.

Chapter 9
Models for Time Series

Abstract In a discrete time series the unobserved variables are latent only in the sense that they lie in the future and are therefore unknown. Any model may thus reasonably begin with the joint distribution of all variables—past and future—from which the relevant conditional distribution may be determined. Two types of model will be described. The first specifies the mean values of the joint distribution and the second, its covariance structure. The former will be described as 'regression-type' models and the latter as 'autoregressive' models. A regression-type model assumes that what we observe is the sum of a systematic part and an 'error'. The systematic part specifies the mean value at successive points in time and the errors are assumed independent. Over sufficiently short periods of time one may be willing to assume that the systematic part is a simple function of time, possibly linear or cyclical, but whatever form is chosen, it is part of the input. An autoregressive model involves an assumption about the covariance structure of the data and, in particular, about the serial correlations of members of the time series. We illustrate this by supposing that any member of the series is correlated with one or two immediate predecessors. The results correspond, as they should, with standard results.

Keywords Autoregressive models · Covariance structure · Multivariate normal distribution · Prediction · Regression models · Serial correlation

9.1 The Scope of Time Series Analysis

In Chap. 1 we showed why time series modelling, and forecasting problems generally, could be viewed as problems involving unobserved variables. Here we expand on that view not so much in order to provide new models but to provide a new perspective on old models. One traditional way of viewing time series is to suppose them to be made up of a systematic part and a random error. This is often appropriate because there may be substantive reasons for supposing that this is the way that the series has actually arisen. For example, any series which depends on

D. J. Bartholomew, *Unobserved Variables*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-39912-1_9, © The Author(s) 2013

seasonal variation may be expected to vary in a cyclical way as the seasons change. Thus consumption of gas by households will be strongly affected by the ambient temperature and this varies in a systematic way throughout the year. As an approximation, therefore, we might suppose that there is an underlying cyclical pattern on which unpredictable fluctuations are superimposed. Over relatively short periods one might expect many series to show a monotonic trend which, if the interval is short enough, might be approximately linear. All of these considerations would encourage us to specify a regression type of model in which the observed values are supposed to be the sum of a specified function and a random error. Within such a framework, the familiar techniques of regression analysis can be brought to bear. There are two other techniques of time series analysis in common use. One, known as spectral analysis, seeks to represent a series as a sum of harmonic components thus enabling us to identify any pronounced periodicity in the data. The other, known as autoregressive modelling, applies regression ideas to the relationships between successive members of the series. For example we might estimate the regression of any observation on its immediate predecessors and use that regression equation for the prediction of later members of the series.

Whatever the technique, one of the prime objectives of time series analysis is to predict future observations, particularly the next member of the series. The general approach used here, as anticipated in Chap. 1, is to regard future observations as unobserved variables and so to predict their values using the appropriate conditional distribution. Thus if we have observed the series $x_1, x_2, \ldots, x_n$ and wish to predict the next member of the series, denoted by y, the relevant distribution will be $f(y|x_1, x_2, \ldots, x_n)$ where n is the length of the observed series, or that part of it which we wish to use for prediction. An awkward, but inevitable, feature of our general approach seems to be absent when we come to time series. This is the unknown, and in general, arbitrary character of the prior distribution of y. This difficulty does not arise with time series because the assumption of a joint distribution of all the variables embraces both those that are observed and those that are not. Thus, for example, if we assume that the joint distribution of $x_1, x_2, \ldots, x_n, y$ has a $(n+1)$-variate normal distribution there is nothing left to say, in general, about y beyond what is implied by that statement. If, of course, y is, itself vector-valued the only change is in the dimension of the distribution. This fact serves to emphasise that more 'work' is being done by the initial distributional assumption, not that we are avoiding the arbitrariness of the distribution of y.

9.2 A General Treatment

Our starting point is the joint distribution of the whole series, including the value or values to be predicted. In practice, there may be no particular reason for selecting any a particular distribution but the main value in starting from here is that we can see how the form of the prediction function is related to the joint

distribution. For example, when the overall joint distribution is multivariate normal the regression of the last member of the series on its predecessors w is linear showing that multivariate normality overall is linked to the linearity of the predictor. In this chapter we shall concentrate on the case when the overall distribution is multivariate normal because this covers most known models. However the general framework we are adopting may, in principle, be used for any joint distribution whatsoever.

It is instructive to begin by looking at the form of this regression function when we wish to predict m future observations. Suppose that the time series is $(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m)$ and that we are interested in predicting the ys when the xs are given. For this we need the distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$, if we suppose that the variables are standardised we then have a well-known result in distribution theory (see, for example, Kendall and Stuart 1999, Vol 2A, p. 512) that may be expressed as

$$\boldsymbol{y}|\boldsymbol{x} \frown N\left(\sum_{\boldsymbol{xy}} \sum_{\boldsymbol{xx}}^{-1} \boldsymbol{x}, \quad \sum_{\boldsymbol{yy}} - \sum_{\boldsymbol{yx}} \sum_{\boldsymbol{xx}}^{-1} \sum_{yx}\right) \tag{9.1}$$

where $\sum_{\boldsymbol{xx}}$ is the covariance matrix of the xs and $\sum_{\boldsymbol{xy}}$and $\sum_{\boldsymbol{yx}}$ are the covariance matrices of $\boldsymbol{x}$ and $\boldsymbol{y}$ and $\boldsymbol{y}$ and $\boldsymbol{x}$ respectively. It is clear immediately that the best predictor of $\boldsymbol{y}$, whether we use the mean or the mode, is a linear function of the observed variables. It is also evident that the uncertainty of prediction, expressed by the covariance matrix, does not depend on $\boldsymbol{x}$.

If the variables are not standardised, we can express Eq. (9.1) in a slightly more general form which makes it easier to bring out the link with traditional time series analysis. If the vector $\boldsymbol{x}$ has mean $\boldsymbol{\mu}_x$ and $\boldsymbol{y}$ has mean $\boldsymbol{\mu}_y$ the result corresponding to Eq. (9.1) is

$$\boldsymbol{y}|\boldsymbol{x} \frown N\left(\sum_{\boldsymbol{yx}} \sum_{\boldsymbol{xx}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_x), \quad \sum_{\boldsymbol{yy}} - \sum_{\boldsymbol{yx}} \sum_{\boldsymbol{xx}}^{-1} \sum_{xy}\right) \tag{9.2}$$

If the successive members of the time series are independent, it is obvious that a very simple result follows. In that case the elements of $\sum_{\boldsymbol{xy}}$ are zero and we then have

$$y|\boldsymbol{x} \frown N(\mu_y, \sigma^2) \tag{9.3}$$

where σ^2 is the variance of y and is the same as $\sum_{\boldsymbol{yy}}$, which in this case is a scalar. The mean, μ_y, is the expected value of y which is the first unobserved variable.

9.3 Regression-Type Models

To see how the general approach relates to classical time series modelling we simply compare Eq. (9.3) with what we may call a *regression-type model.* This is commonly written

$$x_t = \psi(t) + e_t. \tag{9.4}$$

We have used the subscript t instead of i, as above, because observations are usually made at equal intervals of *time*. The systematic part, $\psi(t)$, tells us how x changes with time and this is the expected value of x at time t. The random part, or error, is represented by the last term. It is assumed that successive errors are independent and (usually) that they have the same variance. It is this feature which ensures that, apart from its mean, the distribution of x_t does not depend on t.

9.4 Autoregressive Models

The assumption that the successive observations are independent is very strong and is only appropriate when the function $\psi(t)$ is the main 'driver' of the series—the error terms may then be errors of measurement or observation, which obscure the underlying pattern. We proceed to investigate some of the consequences of allowing dependence between the observations. In anticipation of what follows, we shall refer to all models in which the observations are not independent as *autoregressive models*. This is because the predicted values are linear functions of the observed values. A fuller justification must await the investigation of special cases. The precise form of the model depends on the nature of the assumed dependence between successive observations. Before investigating this in detail we make two observations. According to our approach, any model is specified by the covariance matrix of a set of consecutive observations. In general, we would expect the dependence to be stronger between observations which are close together than between those that are farther apart. In particular, it often seems reasonable to require that the strength of dependence between two observations should depend only on their distance apart. If the strength of the dependence is measured by the correlation coefficient, the resulting coefficients are designated *serial correlations.* The correlation between an observation and its immediate predecessor is said to be of *first orde*r; that between an observation and one k observations earlier is the kth order serial correlation. The entries in the covariance matrix $\sum_{\boldsymbol{xx}}$ are therefore all serial correlations. There is an obvious problem if the serial correlations have to be estimated from an existing time series. The sample correlation may be estimated by pairing each observation with the relevant predecessor. But the first member of the series has no predecessor and the second has no observation two positions earlier—and so on. Various devices are available for handling this situation but they are of no concern to us here as we are dealing only with the modelling aspect.

We begin with the simplest possible case where the prediction is to be based only on the immediately preceding observed variable. In this case the vector $\boldsymbol{x}$ has dimension one, hence $\sum_{xx}$ is a scalar which we denote by v. The matrix $\sum_{xy}$ is also one-dimensional with single element c, which is the covariance of the next

(unobserved) variable and the previous (observed) variable. The predicted value of the next unobserved variable is given, from Eq. (9.1), by

$$\left(\frac{c}{v}\right)x = \rho x \tag{9.5}$$

where ρ is the first order serial correlation of the series. This is a first order autoregressive model which in the conventional time series treatment would be written

$$y = \rho x + e \tag{9.6}$$

Here x is the last available observation, y is the value to be predicted and e, is the error term which is assumed to be standard normal. The conditional variance of the prediction follows by making the appropriate substitutions in Eq. (9.1) and it turns out to be

$$v - \rho c = v\left(1 - \rho^2\right). \tag{9.7}$$

The variance of the predicted value, given x, is thus equal to the original variance reduced by a factor, which is obviously less than one, and which decreases as the first order serial correlation increases. This accords with what intuition would have led us to expect.

The second order autoregressive model will be investigated similarly but the conventional treatment starts by specifying the coefficients in the conditional distribution of Eq. (9.1) and then deduces the properties of the model. Our approach starts with the correlation structure of the data leading to the matrices $\sum_{yx}$ and $\sum_{xx}^{-1}$ and then proceeds to determine the coefficients. The comparison of the two slightly different approaches is instructive.

In the present case there are three variables involved; the variable to be predicted,y, and the two predictor variables $x_1$and x_2. All have the same variance which we denote by v as before. Let us further suppose that adjacent observed variables have covariance c_1 and those a distance two apart (y and x_2 in this case) have covariance c_2. The two matrices required for the conditional distribution are given as follows:

$$\sum_{xx}^{-1} = \frac{1}{v^2 - c_1^2}\begin{bmatrix} v & -c_1 \\ -c_1 & v \end{bmatrix} \tag{9.8}$$

and

$$\sum_{xy}\sum_{xx}^{-1} = \frac{1}{v^2 - c_1^2}\left[vc_1 - c_1c_2,\ vc_2 - c_1^2\right]. \tag{9.9}$$

Expressed in terms of correlation coefficients the right hand side of Eq. (9.9) may be written

$$\frac{1}{1-\rho_1^2}\left[\rho_1-\rho_1\rho_2, \rho_2-\rho_1^2\right] \tag{9.10}$$

where ρ_1 and ρ_2 are the first and second order serial correlation coefficients, respectively. The expression for the conditional variance of prediction is, from Eq. (9.1)

$$\begin{aligned} v-\frac{v}{1-\rho_1^2}\left[\rho_1^2-\rho_1^2\rho_2+\rho_2^2-\rho_1^2\rho_2\right] \\ =v\left[\frac{1-2\rho_1^2+2\rho_1^2\rho_2-\rho_2^2}{1-\rho_1^2}\right]. \end{aligned} \tag{9.11}$$

Equations (9.8) and (9.9) give the coefficients which must be applied to $x_1$and x_2 respectively in order to predict y and Eq. (9.11) gives the variance of the prediction.

One interesting special case occurs when $\rho_2=\rho_1^2$ because then the predictive variance then reduces to Eq. (9.7) and nothing has been gained. This is because the first order dependence arising from a first order serial correlation of $\rho_1$implies a second order serial correlation of ρ_1^2 so nothing is added to the predictive value.

9.5 Concluding Remarks

The approach outlined here unifies the two rather *ad hoc* methods of time series modelling by emphasising that traditional regression-type methods are essentially concerned with changes in the mean levels over time whereas autoregressive methods are more concerned with serial correlation in the series. It also shows that the two aspects can easily be combined in a more comprehensive model. Two other remarks are in order before we leave this topic. As already noted we do not have to assume multivariate normality though, without it, the linearity is lost. As in routine time series analysis, transformations of the data may sometimes be made to induce normality and thus validate the procedures. Secondly, the choice of elements in a covariance matrix is not entirely arbitrary as our treatment might have suggested. They are constrained by the necessity that the matrix shall be positive definite, for example. Nevertheless, such complications need not concern us here because our concern is solely with the modelling framework.

Reference

Kendall, M. G., & Alan, S. (1999). *The advanced theory of statistics* (6th ed., Vol. 2A). London: Arnold.

Chapter 10
Missing Data

Abstract It is very common for data to be missing and this introduces a risk of bias if inferences are drawn from incomplete samples. However, we are not usually interested in the missing data themselves but in the population characteristics to whose estimation those values were intended to contribute. Learning something about the data that are missing is thus only the first step on the way to inference. One approach is to use a direct method, such as maximum likelihood but the price to be paid is usually much greater complexity in the estimation process. Methods such as the E-M algorithm sometimes make this easier by requiring us to solve a much simpler problem many times as the estimates converge to the desired values. Sometimes it is actually advantageous to introduce hypothetical variables. Which are then treated as unobserved and an example is provided concerning a mixture of exponential distributions. A different kind of approach is to impute values to replace those that are missing. This yields a complete sample which can then be analysed in the usual way. Imputed values can be derived from the conditional distribution of the missing values given those that are observed. This possibility depends upon being able to say something about why some sample members are missing and this may be done by specifying a probabilistic loss mechanism.

Keywords E-M algorithm · Imputation · Maximum likelihood · Missing at random · Missing completely at random · Mixed exponential distribution · Mixtures

10.1 The Problem

Missing data are very common in statistics, especially in social applications, and this topic provides, perhaps, the most obvious example of unobserved variables. However, as we first encounter it, the problem differs in one important respect from those we have discussed earlier because we are no longer directly interested

D. J. Bartholomew, *Unobserved Variables*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-39912-1_10, © The Author(s) 2013

in the values of the variables that are missing but in the analysis to which they were intended to contribute. Nevertheless, we shall note shortly that, since latent variables may also be regarded as 'missing', essentially the same methods can be used as for some of the problems we have met in earlier chapters. For this reason we shall pay particular attention to applications of that kind.

To take a very simple example, suppose that we have a random sample of males and another of females from a human population and ask each member whether or not they smoke. Not all members will respond and suppose we know that men are more likely to refuse than women. If we estimate the proportion of smokers in the population by the proportion in the combined sample the result is likely to be biased because women smokers will be over-represented. Most actual examples are much more complicated than this, involving many attributes which are all potential sources of bias but the problem is essentially the same. There are different ways of handling bias; in this example we would be likely to know how many men and women there were in the population and so we could correct the bias by an appropriate weighting. In general, this may not be possible. We return to this example later.

10.2 The E-M Algorithm

This is an iterative method of obtaining maximum likelihood estimates which is sometimes much easier to handle than the direct method of maximising the likelihood of the observed variables. The algorithm has been used in many guises but it was given its name and firmly established by Dempster et al. (1977). It depends on the fact that although we may not be able to easily maximise the loglikelihood itself, because some of the observations are missing, we can maximise its expectation iteratively. First, we state the algorithm in its general form and then illustrate its application when the missing values are, in fact, unobserved variables of the kind we met in earlier chapters.

Sometimes it is easier to estimate the parameters if we artificially introduce additional variables which are then treated as if they were unobserved. Paradoxically, this happens because it would be easier to solve the problem thus created than the original problem. In other words, the solution of the problem with 'missing' values may have a much simpler form in spite of the fact that some variables are not observed.

The loglikelihood for the complete sample may be written

$$\ln f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) \tag{10.1}$$

where $\boldsymbol{y}$ is a vector of unobserved random variables and $\boldsymbol{\theta}$ is a vector of parameters whose values we wish to estimate. The likelihood itself is a random variable, because of the presence of $\boldsymbol{y}$. When $\boldsymbol{x}$ is given, the expectation, with respect to the unobserved variables, $\boldsymbol{y}$, is

$$\int f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\ln f(\boldsymbol{x},\boldsymbol{y}|\theta)d\boldsymbol{y}. \tag{10.2}$$

Of course, we do not know the value of the parameters in (10.2) so we cannot calculate the expectation in this form. We therefore proceed iteratively by first assigning arbitrary values to the unknown parameters in $f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})$, and then maximising the expected loglikelihood with respect to $\boldsymbol{\theta}$. The expectation we actually calculate at the first stage is therefore,

$$\int f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}_g)\ln f(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{y} \tag{10.3}$$

where $\boldsymbol{\theta}_g$ contains the starting, or guessed, values of the unknown parameters. In the next round of the iteration this expression is then maximised with respect to $\boldsymbol{\theta}$ and these maximising values then replace $\boldsymbol{\theta}_g$ in Eq. (10.3). Before continuing with the cycle we must update $f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}_g)$ by replacing $\boldsymbol{\theta}_g$ by the value of $\boldsymbol{\theta}$ which maximises Eq. (10.3).We denote this by $\boldsymbol{\theta}_1$and at the j th iteration by $\boldsymbol{\theta}_j$. This cycle of expectation and maximisation continues until convergence is attained.

Finally, the updating of $f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})$ is achieved using Bayes' theorem as follows.

$$f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}_{j+1}) = f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}_j)/f(\boldsymbol{x},\boldsymbol{\theta}_j) = f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}_j)/\int f(\boldsymbol{y})f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}_j)d\boldsymbol{y}. \tag{10.4}$$

In practice it is clear that the benefit of the method depends on how easy it is to carry out the steps and this, in turn, depends on what we know about the reason for the data being missing. An important special case is where the complete sample has a distribution which belongs to the exponential family. This leads to a situation where the function to be maximised at each step has essentially the same form.

10.3 An Example with Hypothetical Variables

The E-M algorithm is particularly useful for estimating the parameters of mixtures of distributions. In this case we introduce a hypothetical latent variable and treat it just like any other variable which we might have observed but did not. This establishes a link with the treatment of mixture distributions treated in Chap. 3 and the latent variable models of the following chapters. We now show, in detail, how this works for the case of the two-class mixed exponential distribution and then indicate the generalisation to any number of classes.

We imagine that a random sample from a two term exponential distribution is generated as follows. Suppose there are available two simple exponentials with parameters λ_1 and λ_2, respectively and that we draw from the distribution having parameter λ_1 with probability p and from the other distribution with probability $1 - p$. Which distribution we have sampled from is unknown but may be identified by an unobserved random variable y which takes the value 1 if the first distribution is selected and 0 otherwise. In the long run the probability distribution of the sampled values will therefore be

$$f(x) = p\lambda_1 \exp - \lambda_1 x + (1-p)\lambda_2 \exp - \lambda_2 x. \tag{10.5}$$

It does not matter whether the real distribution with which we are working was actually generated in this way.

Because of the way we have supposed the distribution to have been generated, the likelihood can be expressed in a particularly simple form as follows. At each sampling we shall be drawing either from the simple exponential with parameter λ_1 or from one with parameter λ_2. In the former case the contribution from the observation $x_i$to the likelihood will be a factor $\lambda_1 \exp - \lambda_1 x_i$ and in the second it will be $\lambda_2 \exp - \lambda_2 x_i$. Let $a_i = 1$ if the first distribution is sampled and $a_i = 0$ otherwise The likelihood, conditional on observing this particular set of $a_i s$, will then be

$$l = \prod\nolimits_i \left(\lambda_1 e^{\lambda_1 x_i}\right)^{a_i} \left(\lambda_2 e^{\lambda_2 x_i}\right)^{1-a_i} \tag{10.6}$$

The loglikelihood is therefore.

$$L = \log l = \sum\nolimits_i \{a_i(\log\lambda_1 - \lambda_1 x_i) + (1-a_i)(\log\lambda_2 - \lambda_2 x_i)\} \tag{10.7}$$

$$= r\log\lambda_1 - \lambda_1 \sum\nolimits_i a_i x_i + (n-r)\log\lambda_2 - \lambda_2 \sum_i (1-a_i)x_i. \tag{10.8}$$

where $r = \sum a_i$ are contributions of the first kind and therefore, $n - r$ are of the second. We cannot maximise the loglikelihood as it stands because it involves random variables. We can, however, maximise its expectation as the E-M algorithm requires. If we take the expectation with respect to the as we obtain

$$E(L) = np\log\lambda_1 - \lambda_1 \sum\nolimits_i p_{i0} x_i + (n-np)\log\lambda_2 - \lambda_2 \sum\nolimits_i (1-p_{i0})x_i \tag{10.9}$$

Where p_{i0} is the starting (guessed) value of the probability that the ith sample member is drawn from the population with parameter λ_1 and $p = \sum_i p_{i0}/n$.

The expression in Eq. (10.9) may be maximised with respect to λ_1 and λ_2 giving a maximum which occurs at

$$\lambda_{11} = \sum\nolimits_i p_{i0} \Big/ \sum\nolimits_i p_{i0} x_i \text{ and } \lambda_{21} = \sum\nolimits_i (1-p_{i0}) \Big/ \sum\nolimits_i (1-p_{i0}) x_i \tag{10.10}$$

where the second subscript on λ denotes he iteration number. Next we must now update p_{i0} using Bayes theorem. This gives

$$p_{i1} = \frac{p_{i0}\lambda_{10} \exp - \lambda_{10} x_i}{p_{i0}\lambda_{10} \exp - \lambda_{10} x_i + (1-p_{i0})\lambda_{20} \exp - \lambda_{20} x_i}. \tag{10.11}$$

Using these new weights we return to Eq. (10.10) and compute λ_{12} and $\lambda_{22.}$ These, in turn, lead to new values of p_{i2}, and so on until convergence is reached. The final iterations for λ_1 and λ_2 are the maximum likelihood estimates and the

estimate of p is the arithmetic mean of the inclusion probabilities $\{p_{ik}\}$ where k is the number of iterations to convergence.

Although the method has been illustrated on a very simple example with a hypothetical missing observation, it is straightforward to extend it in two directions. The simplicity of the method depended on the fact that it was easy to obtain maximum likelihood estimators for the component distributions so two-component mixtures of other similar distributions can be handled in the same way. The extension to more than two components can also be handled by using vector-valued indicator variables instead of the binary indicators $\{a_i\}$. Programs for fitting mixtures of exponentials using the E-M algorithm are available in the R-library. One such is named "Renext".

10.4 Imputation

A very longstanding way of dealing with missing data is to fill in the gaps by some means or other and then carry out the standard analysis on the completed data set. This procedure is known as *imputation.* If we view the problem from the general perspective of this Brief it is essentially one of how best to use the information about the missing values obtained from the appropriate conditional distribution. This information is supplied by the probability function $f(\boldsymbol{y}|\boldsymbol{x})$ where y now represents the missing values and $\boldsymbol{x}$ the remainder that are observed. The problem of imputation is thus one of selecting values which are representative of this distribution, $f(\boldsymbol{y}|\boldsymbol{x})$. In its simplest form, each missing data point is replaced by a single value. Because there is, inevitably, uncertainty about what the imputed values should be, one can do better by substituting a range of plausible values and comparing the results in each case. This is known as *multiple imputation*. This was hardly feasible in the pre-computer era but now that that obstacle has been removed, imputation and multiple imputation is much to be preferred, especially as it is usually sufficient to repeat the analysis a small number of times, five say.

The E-M method of fitting a mixture distribution, as described above, involved imputation, in a sense, because we did not know which component the sampled member came from and so we began by guessing (i.e. imputing) a value. But the situation we have in mind here is much more general and, often, less well-defined. There is an enormous literature on this topic and, here, we shall pick out only a few topics to locate the subject within the framework of this Brief. Much of the pioneering work in this field is set out in Little and Rubin (2002, first edition 1987) but a recent and practically orientated treatment is provided by van Buuren (2012).

10.5 Probability Specifications of Missing Data

For our purposes the problem can best be set in the context of the data matrix, sometimes referred to as a rectangular array which we have already met in Chaps. 2 and 4. In the present instance missing values may appear anywhere in the body of the table so that, for example, the first row might be

$$x_{11}, x_{12}, \ldots, y_{1j}, \ldots, x_{1n}$$

where y_{1j} indicates a value of a single missing observation. Inpractice, of course, missing values may occur anywhere and in any number. They may occur haphazardly or in some pattern. In the latter case, the pattern may provide a clue to the mechanism underlying the loss of data and so suggest a method for dealing with it.

The conditional distribution which we have supposed might be the basis of imputation depends, of course, on the mechanism behind the loss of data. From a practical point of view the detailed information necessary to determine this may not be readily obtainable or, even, necessary. Nevertheless, it is useful to clarify some of the issues by introducing the idea of a probability mechanism governing the loss of data. This will enable us to classify the problems which would have to be faced in a more comprehensive treatment.

The simplest, if least realistic approach, is to assume that the chance of being missing is the same for all elements of the data matrix. In that case, we can, in effect, ignore the missing values and all that is lost is the information which those missing values would have contributed. In the smoking example used at the beginning of the chapter this would amount to saying that men and women were equally likely to refuse to answer. Such situations are designated as MCAR which is an acronym for Missing Completely at Random. We may express this assumption formally by saying that

$$Pr\{M|\boldsymbol{x},\boldsymbol{y}\} = Pr\{M\} \tag{10.12}$$

where M specifies the mechanism governing the loss of observations.

In the smoking example we have supposed that men are more likely to refuse than women. If we go further and assume that there are no other biasing factors we are, in effect, assuming that 'missingness' is completely at random for men and women, separately. This would be an example of what is known as Missing at Random(MAR). In terms of the standard data matrix layout this supposes that data are missing at random *within* columns of the table. This requirement may be specified probabilistically by requiring that

$$P\{M|\boldsymbol{x},\boldsymbol{y}\} = P\{M|\boldsymbol{x}\} \tag{10.13}$$

which means that the missing mechanism depends on the observed variables but not on those that are missing.

The final category is Missing Not at Random (MNAR) which is a residual category covering all other possibilities. This is difficult to deal with in practice unless one has an unusually complete knowledge of the missing mechanism.

Another term used in the theory of missing data is that of *ignorability*. The conditional distribution of y given x will, in general, depend on any parameters of the distribution of M yet these are unlikely to be of any practical interest. It would be convenient if this distribution could be ignored for the purposes of inference about the parameters of the distribution of $\boldsymbol{x}$. If this is the case the mechanism of loss is said to be ignorable. In practice it is acceptable to assume that the concept of ignorability is equivalent to that of MAR.

References

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

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley. (1st edn. 1987).

van Buuren, S. (2012). *Flexible imputation of missing data*. London: Chapman and Hall/CRC Press.

Chapter 11
Social Measurement

Abstract Latent variables are often necessary to provide an adequate model of a social or physical situation. Inferences about such variables can be made using the methods sketched earlier in this Brief, especially in Chaps. 4–6. However, there are situations which are not well enough defined to permit the construction of a comprehensive model. Two such examples are considered in this Chapter. The first is labour wastage and the second is heritability. Both are practically very important and have been the source of much misunderstanding and controversy arising through lack of an adequate model. However, it is possible to construct partial models which elucidate the complexity of the situation making clear where caution must be exercised and delineating the circumstances under which the simple measures in common use may be used legitimately. Our treatment of the former requires the introduction of the hazard function of a probability function for a positive random variable and the renewal rate, both of which are defined.

Keywords Environnmental effects · Heritability · Hazard function · Genetical effects · Latent variable models · Renewal function · Turnover · Wastage

11.1 The Problem

It may not be immediately obvious that social measurement falls within the purview of this Brief. However, it is common for public discussion to involve what appear to be quantitative arguments concerning measures which either cannot be observed directly or at all. For example, a recent lecture was entitled "Injustice: the cause of rising inequality". This implies that there exists a quantity appropriately described as injustice and that it is monotonically related to another quantity called inequality. Yet the measurement of injustice and inequality are both highly problematical. Both centre on variables which are essentially unobservable, or, latent, as they might be, alternatively, described. Indeed if we adopt the viewpoint of Bartholomew (1996), the only sound approach to social

D. J. Bartholomew, *Unobserved Variables*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-39912-1_11, © The Author(s) 2013

measurement is to identify the quantities to be measured with latent variables or parameters in a probability model. This is not always the case with social measures and certainly not the two to be discussed in this chapter.

We shall not attempt to cover the whole field of social measurement but focus, instead, on two particular instances where intuition has proved a very poor guide. They are both cases where important practical issues turn on the measure used. The first is the propensity of individuals to leave the organisation for which they work, This is held to be an important indicator of industrial health and there have been many attempts to construct appropriate measures. This is most commonly done by constructing a very obvious but potentially misleading index of what is variously called *wastage* or *turnover*. Thus, for example, the proportion of those who leave an organisation in a year, say, is used as a measure of attachment and numerical differences between such proportions have been interpreted as indicative of substantive differences. The importance of our second example stems from the fact that a great deal hangs on the extent to which intelligence is heritable. This is at the heart of the 'nature/nurture' debate which has rumbled on for decades. To give these arguments substance it is essential to be able to measure intelligence adequately and then to determine the relationship, if any, which exists between the intelligence of parent and offspring.

11.2 Propensity to Leave an Organisation

Labour wastage is only one example of the phenomenon we wish to discuss. There are also other phenomena, only loosely related to employment, such as the propensity for patients to be discharged from hospital, or for residents to move house, or to cease membership of a society or movement. But labour wastage has been studied in greater depth and it is on this that we shall concentrate here. All such phenomena, however, are extremely complicated and many factors exert an influence. At times of economic growth, for example, interest tends to focus on leaving because firms are anxious to retain employees. In times of recession, on the other hand, the interest is on propensity to leave the pool of the unemployed and to return to employment. In order to concentrate on essentials we shall divide all these influences into two classes; the extrinsic and the intrinsic. The extrinsic are all those factors which are external to the system and which could, in principle, at least, be controlled and, if necessary, held constant. The intrinsic influence consists of only one factor; namely the length of time an individual has been in the system.

If propensity to leave did not depend on length of stay, the intrinsic factor would be irrelevant and the problem we face would vanish. In reality, propensity to leave depends very strongly on length of stay and it is this empirical fact which creates our problem. Propensity to leave for any individual is conveniently measured by their *hazard function*.

11.3 The Hazard Function

Figure 11.1 gives three examples of hazard functions.

The hazard function, $\lambda(t)$, is formally defined as follows:

$$Pr\{\text{loss occurs in } (t, t+\delta t)\} = \lambda(t)\,\delta t + o(\delta t) \tag{11.1}$$

The distribution of the length of time that an individual stays with an organisation is often lognormal in form. The solid curve in Fig. 11.1 is the hazard function for a lognormal distribution with parameter values typical of those found in this field. Initially it rises steeply to a peak near the origin (not shown because it occurs very close indeed to the origin) and then decreases monotonically to zero. The dotted line is for a continuous mixed exponential as discussed in Chap. 3 with the typical parameter value of $q = 2$ and it is hardly distinguishable from the lognormal hazard. The horizontal line is the hazard for an exponential distribution. All three distributions have been chosen to have a mean of 1.

Compared with the exponential, the other curves show a much higher propensity to leave for those with short service. For longer lengths of service the position is reversed. It is immediately clear that the number of leavers in any time interval will depend strongly on the current lengths of service of those in the system. Just how strong this dependence is we must now investigate. When interpreting Fig. 11.1 it should not be forgotten that the average length of stay, with all the distributions illustrated is 1 so the very low propensity to leave in the right hand part of the diagram relates to lengths of stay much greater than the average.

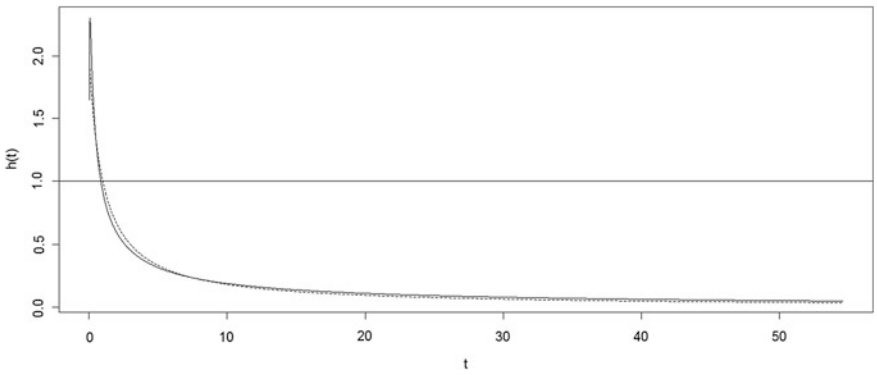

Fig. 11.1 Examples of hazard functions

11.4 The Renewal Rate

We now show why the crude rates may be so misleading. Consider a cohort of individuals who all join at the same time and let us trace their leaving history as time passes. For this we require what is called the renewal rate or density. The rate is denoted by $r(T)$ and is defined by the fact that $Nr(T)\delta T$ is the expected number, from an initial cohort of size N, who leave in the interval $(T, T + \delta T)$. Renewal theory shows this to be given by

$$Nr(T) = Nf(T) + N\int_0^T r(T-t)f(t)dt \tag{11.2}$$

Using this equation we can, in principle at least, find the expected number who will leave in any given interval.

The first relevant thing which may be deduced from (11.2) is that, as $T \to \infty$, $r(tT) \to \mu^{-1}$ where μ is the mean length of stay. This establishes the link between the renewal process and the asymptotic leaving rate. Furthermore it suggests that the expectation of length of stay would be a suitable way of summarising the hazard function. The use of expectation of service was, in fact, first suggested by Lane and Andrew (1955), although it suffers from a practical disadvantage. It depends strongly on the largest lengths of service. About which there is often the greatest uncertainty; an alternative measure such as the median may therefore be more practical in many circumstances.

Asymptotically, at least, the expected number leaving in the interval $(T_1.T_2)$ is $N(T_2 - T_1)\mu^{-1}$ which is a constant, not depending on where the interval is located, provided that T_1 and T_2 are both large. A key question is whether this limiting value is an adequate approximation in the early stages of the development of the process. In fact, it turns out to be a very poor approximation, as we shall see, and it is this fact which provides the motivation for this section.

The second deduction which can be made from (11.2) is that the only distribution for which the renewal rate is equal to the asymptotic value for all times is the exponential given by

$$f(t) = \lambda \exp -\lambda t \tag{11.3}$$

In this case, $r(T) = \lambda = \mu^{-1}$. If length of stay distributions did have this form, the use of wastage rates would be fully justified. In numerical terms, for example, this would mean that if the average length of stay were 4 years then the limiting wastage rate would be 25 % per annum. When this is not the case, organisations having a significant proportion of 'young' members will have wastage rates different from the long term values.

We now give some calculations which illustrate just how serious this discrepancy can be in practice. The problem of determining the renewal rate is that it is the solution of an integral equation which may not be easy to solve either

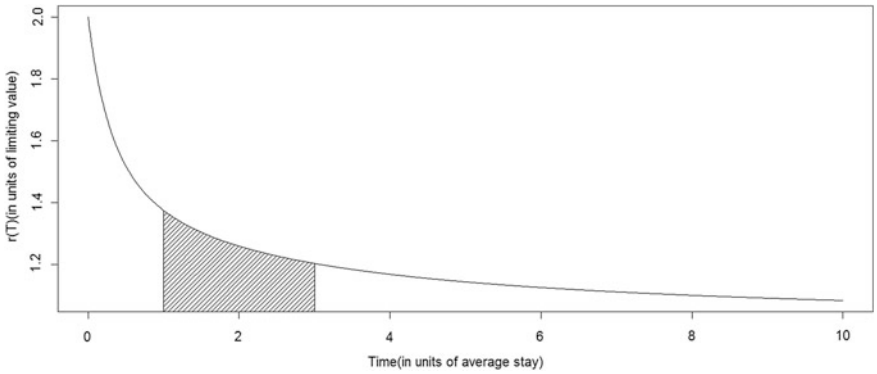

Fig. 11.2 The renewal rate for the mixed exponential distribution with $q = 2$ and unit mean compared with its asymptotic level

analytically or numerically. Fortunately, there is a simple approximation available which takes a very simple form for the continuous mixed exponential discussed above. It was given in Bartholomew (1963) and yields

$$r(T) \cong f(T) + \frac{F^2(T)}{\int_0^T G(t)dt} \tag{11.4}$$

In the case of the distribution of Eqs. (3.9) and (3.10) in Chap. 3 this becomes, when $q = 2$,

$$r(T) \cong \frac{2 + (T+1)^2}{(T+1)^3} \tag{11.5}$$

This approximation is plotted on Fig. 11.2. It shows that the renewal rate is always larger than its limiting value and that, in the early stages especially, it is very much larger.

Any organisation will be made up of many cohorts each recruited at different times and each will contribute something to the total wastage. But all will contribute at more than the average level. The global wastage thus depends on the particular mix of cohorts present but the important thing to notice is that it depends, not only on the intrinsic propensity to leave, but also on historic recruitment levels. It is not, therefore a pure measure of what it is often intended to measure. Furthermore the contamination can be quite serious as Fig. 11.2 illustrates.

There are less obvious ways in which this misunderstanding may lead to unrealised expectations. For example, it is common, especially in times of recession, to wish to contract the size of an organisation. It is then common to argue that the desired contraction can be achieved by 'natural wastage' without any need for compulsory redundancies. This claim is typically based, for example, on statements that the organisation already has a natural wastage of, 10 %, say, which suggests that a reduction at this rate can be achieved simply by cutting off

recruitment. This argument is fallacious because much of the contribution to overall wastage comes from those most recently recruited. Once these people have left, the loss from those remaining is necessarily less, because there are relatively fewer people with short service. Hence, the expectations prove to be over-optimistic.

11.5 Heritability

The second problem of social measurement which we shall discuss is that of heritability. This is a measure of the extent to which any biological or mental characteristic is inherited. It had its origins in the 1920s in the work of R A Fisher in the fields of plant and animal breeding. Latterly it has become prominent in a social context, particularly with respect to the inheritance of such things as intelligence, where it has become a key element in the nature versus nurture debate. It is usually expressed by a number in the range (0,1) or, equivalently, as a percentage. Such numbers have become part of public debate without a clear understanding of what they mean or what their limitations are. For example, in the Milburn report on social mobility published in the UK under the title *Unleashing Aspiration*, it was stated "first of all that, the evidence on whether intelligence is itself inherited is far from clear: estimates varying from zero to 80 %. Recent work suggests that genetics and environment interact in quite complex ways". For the source of these figures. the reader was referred to the Nobel Prize web site but the actual source appears to be in the proceedings of a conference on the Nature/Nurture debate published in Acta Paediatrica of 1997. Consultation of Wikipedia produces the range of estimates for heritability between 50 and 80 %. Given the enormous importance of these figures for social policy, it is clear that one needs to know what these figures actually mean and why there is so much variation in the ranges given in the sources.

The subject is an immensely complicated part of quantitative genetics and there is neither the need nor the possibility of going into the details here. A good introduction to the problem will be found in Daniels et al. (1997) and a non-technical account, so far as it relates to intelligence, in Bartholomew (2004, Chap. 12).

Let x denote the value of some quantity in whose heritability we are interested. This is presumed to be determined partly by the parents and partly by the environment. In other words there is both a genetic component and an environmental component. A very simple model expressing this fact might be written

$$x = g + e + (ge) \tag{11.6}$$

where g is the genetic contribution, e is the environmental component and (ge) is included to allow for the possibility that there might be an interaction, meaning that the environmental contribution depends on the genetic contribution and vice versa. The very simple points we wish to make do not depend on this term so we

shall treat it as zero. In practice, however, the existence and magnitude of any interaction effect is still controversial especially in the application to intelligence. The variance of x, ignoring the interaction effect, may be written,

$$\sigma_x^2 = \sigma_g^2 + \sigma_e^2 \tag{11.7}$$

The coefficient of heritability is then defined as

$$I = \sigma_g^2/\sigma_x^2 \tag{11.8}$$

The rationale behind this choice is that, if there is no environmental variation, the coefficient will be 1 and if there is no genetic variation it will be 0. The greater I, therefore, the greater the importance of the genetical component.

The first thing to be noticed is that this measure relates to a *population* and not to any *individual.* It is thus critically dependent on how the population is defined. The second thing is that the coefficient can only be calculated if the variances can be estimated. This is not a trivial matter but outside our present concern.

The main source of misunderstanding about measures like I, arises from the fact that they that they depend on things other than those they are supposed to be measuring. We have already seen that a wastage rate may depend on the length of service structure of the system as well as the propensity to leave which it was designed to measure. In the case of heritability the measure also depends on the characteristics of the population. In social applications it is particularly vulnerable to variations in the environment variance. If pressed to the limit, these variations can easily lead to paradoxical results. For example, intelligence as measured by IQ, often depends on educational opportunity and this may result in attempts to increase and equalise educational opportunity. But even if these attempts are only partially successful they will reduce environmental variation and so increase heritability as measured by I without there having been any actual change in the mechanism of inheritance itself. In general, I will depend on the degree of environmental variance and this may well vary from one population to another. It may also happen that some of the variation in heritability reported above may reflect that fact and have nothing to do with the inheritance of IQ itself.

All of the foregoing discussion relating to heritability was concerned with IQ. The intelligence quotient (IQ) is an index, calculated from test scores and not the underlying latent variable which it is intended to measure. The latter is often known as g and this is presumed to measure some characteristic of the brain. Ideally we want to know how the g-value of the offspring and parent is related and it is that relationship which we ought to be estimating. IQ is not, itself, inherited but is a property of the brain of the person to which it relates. One needs to set up a model in which heritability, appropriately defined, appears as an unobservable quantity measuring this relationship. This does not appear to have been attempted but the modelling approach highlights the problem and shows how it might be tackled.

References

Bartholomew, D. J. (1963). An approximate solution of the integral equation of renewal theory. *Journal of Royal Statistical Society B, 25*, 432–441.

Bartholomew, D. J. (1996). *The statistical approach to social measurement*. San Diego: Academic Press.

Bartholomew, D. J. (2004). *Measuring intelligence: Facts and fallacies*. Cambridge: Cambridge University Press.

Daniels, M., Bernie, D., & Kathryn, R. (1997). Of genes and IQ. In M. Daniels, S. E. Fienberg, D. P. Resnick, & R. Kathryn (Eds.), *Intelligence, genes and success* (pp. 41–70). New York: Springer.

Lane, K. F., & Andrew, J. F. (1955). A method of labour turnover analysis. *Journal of Royal Statistical Society A, 118*, 303–396.

Chapter 12
Bayesian and Computational Methods

Abstract In Bayesian inference the parameters are treated as random variables. Since they are necessarily unobserved, a Bayesian approach to inference appears to bring the whole of statistical inference within the purview of this Brief. In a formal sense, at least, this is the case. However, it is important to distinguish the different kinds of uncertainty with which we are dealing. First, there are the actual, observable, distributions of the manifest variables, secondly the hypothetical, and often arbitrary, distributions of unobservable variables, to these we must now be add the subjective distributions of unknown parameters. Ideally we would need to track and evaluate the part which each played in the final outcome. It is common, however, for our problems to involve many parameters and their actual distributions turn out not to be critical. For this reason, the pragmatic course is to use the Bayesian approach and obtain an empirical approximation to posterior distributions by Monte Carlo sampling. This may be effected by using the Hastings–Metropolis algorithm, or its simpler derivative, the Gibbs sampler. An outline discussion of these methods is given.

Keywords Bayes theorem · Frequentist inference · Gibbs sampling · Metropolis-Hastings algorithm · Inference · Markov chains · Markov chain Monte Carlo methods (MCMC) · Monte Carlo sampling

12.1 Approaches to Inference

In the Bayesian approach to inference, parameters are treated as random variables and this seems to put them on the same footing as other unobserved variables. At first sight this might make it possible to subsume all of the problems treated in this Brief under the Bayesian umbrella. Before rushing to this conclusion some cautionary remarks are in order.

The first is to take note of the computing revolution which has taken place in Statistics. Half a century ago the battle lines were drawn between Bayesians on the

D. J. Bartholomew, *Unobserved Variables*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-39912-1_12, © The Author(s) 2013

one hand and Frequentists on the other. Much of the debate turned on whether it was possible or desirable to introduce the subjective element into statistical inference which was required by the need to provide a prior distribution for the unknown parameter or parameters. Typically, there were only a small number of parameters—often only one or two. Simultaneously with the increase of computing power available there arose the need to handle problems with many more parameters and the need to solve the computational problems associated with them. This has changed the whole perspective within which inference is approached.

To begin with, it was practically impossible to elicit from individuals the multivariate prior distributions required. Secondly, the numerical evaluation of the multiple integrals, which the formal application of Bayes' theorem often required, was formidable. Even had it been possible to obtain the posterior distribution, the result would have been too complex to be absorbed by the user and radical summarisation would have been necessary. Such a summarisation would inevitably have concentrated on the region around the maximum of the posterior distribution in which the maximum of the likelihood and the shape of the likelihood around it would play a key role. Furthermore, if the prior distribution were fairly flat in the neighbourhood of its maximum, there would be little practical difference between the maximum of the posterior and the maximum of the likelihood. All of this blurs the distinction between Bayesian and likelihood approaches. For practical purposes, therefore, the difference between Bayesian inference and its competitors, which was never large, is now commonly disregarded, implicitly if not explicitly.

The relevance of all this to our present theme is that when unobserved, or latent, variables are added to the parameters of a model, the total number of unobserved variables moves the problem into the 'many parameter' class. But, before moving on, we make a number of basic points.

12.2 Preliminaries

The term 'Bayesian' can be somewhat ambiguous. If it refers to any approach which uses Bayes' theorem, then many of the methods used in latent variable modelling are certainly Bayesian because we typically determine the conditional distribution of the latent variables given the manifest variables, using Bayes' theorem. But the term is usually used more narrowly to refer only to the treatment of parameters as random variables.

As we noted in Chap. 1 we are dealing with three types of quantity: the manifest variables, the latent variables and the parameters. Even if these are treated individually as random variables, the variation which the probability distributions describe is of very different kinds and it is important to be clear at the outset what those differences are.

Manifest variables. These are observable and, in principle, their distributions are also observable. Their distributions may depend on parameters or latent variables or both. Hence there is no ambiguity, in principle, about what their probability distribution is.

Latent variables. Although these are assumed to vary, their distributions cannot be observed and may not, therefore, be estimable—even in principle. What we commonly do in these circumstances is to arbitrarily assume the form of their distribution so, in effect, we are determining the kind of scale on which the latent variable is measured. Thus, for example, if we use a normal distribution to describe the distribution of a human ability we are thereby saying that we have chosen a metric which renders the distribution normal.

Parameters. By definition these are fixed quantities but since we do not know their values our uncertainty about them may be expressed by probability distributions. We might try to do this in a formal way to represent total ignorance or we may summarise our state of knowledge which, unless it is agreed with other persons, is necessarily subjective.

(It is worth noting at this point that in our treatment so far, as in much practical work, we have actually proceeded as if the parameters are known. But the methods cannot be implemented unless numerical values are available for all unknown parameters. In practice unknown parameters are estimated either from data obtained beforehand or from the current data. The procedures which we have described will be imprecise but only to the extent that estimated values are being used instead of true values.)

In a formal sense there is no difficulty in expressing what we have said about the parametric situation in terms of probability distributions. All we have to do is to append the unknown parameters, $\boldsymbol{\theta}$, to the latent variables and find the expectation with respect to the θs. Thus

$$\mathrm{h}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \boldsymbol{h}(\boldsymbol{y}|\boldsymbol{\theta})\boldsymbol{h}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})/\boldsymbol{f}(\boldsymbol{x}|\boldsymbol{\theta}) \tag{12.1}$$

If we now wish to predict $\boldsymbol{y}$ in the absence of knowledge about $\boldsymbol{\theta}$, we shall have to average the distribution on the left hand side of (13.1) with respect to $\boldsymbol{\theta}$. Thus

$$h(\boldsymbol{y}|\boldsymbol{x}) = \int h(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{\theta})\boldsymbol{d}\boldsymbol{\theta} \tag{12.2}$$

On the other hand, if we wish to estimate the parameters, we shall need $f(\boldsymbol{x}|\boldsymbol{\theta})$.

One of the great breakthroughs of modern computational statistics has been the development of Monte Carlo algorithms for obtaining estimates of multi-parameter distributions. These methods are not usually automatic in the sense that there is an algorithm which leads inexorably to the correct solution. Instead they are user-guided iterative procedures which require the judgment of a human operator. It is interesting and, perhaps, significant that the methods were pioneered outside of Statistics, mainly by physicists and computer scientists. The methods treat latent variables and parameters alike as random variables, so they may be used to find the distributions of both simultaneously if necessary.

12.3 Markov Chain Monte Carlo Methods

The methods we have in mind are known as Markov Chain Monte Carlo methods (MCMC). They capitalise on the particular strength of modern computers to carry out a very large number of simple operations exceedingly quickly. The essential idea is to generate random samples from a probability distribution instead of determining the distribution itself. For sufficiently large samples, the characteristics of the sample distribution (e.g. mean, standard deviation etc.) will approximate those of the true distribution. Software packages such as R have built-in routines for generating samples from many common distributions and these have effectively replaced traditional statistical tables.

Markov chains are relevant because they may generate an equilibrium distribution which can be identified with a desired posterior distribution. To see this we need some basic properties of Markov chains the essentials of which are as follows.

For simplicity, we express the results required for discrete time-homogeneous Markov chains which can be thought of as approximating the continuous version in which our problem will often be expressed. At a given 'time' a Markov chain can be in one of a number of 'states', N, say. Between time t and time $t + 1$ the chain moves from state i to state j with probability p_{ij}, say,which does not depend on t This probability is known as a transition probability. For convenience the set of transition probabilities can be set out in a 'transition matrix' as follows,

$$\begin{array}{cccc} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{array}$$

We denote this matrix by **P**. Next we imagine an entity moving between states of the system in such a manner that it makes the transition from state i to state j with probability p_{ij}. Let $p_i(t)$ be the probability that the individual is in state i at time t and let $\boldsymbol{p}(t)$ be the row vector with these probabilities as elements. A direct probability argument then gives that

$$\boldsymbol{p}(\mathrm{t}+1) = \boldsymbol{p}(\mathrm{t})\boldsymbol{P} \tag{12.3}$$

Provided that every state can be reached from every other state,and that there is no cycling round a sub-set of states,the chain will eventually reach a limit with state probabilities, $\boldsymbol{p}$, satisfying

$$\boldsymbol{p} = \boldsymbol{p}\boldsymbol{P} \tag{12.4}$$

This limiting distribution is also the 'stationary' distribution because, once attained, the process remains there as Eqs. (12.3) and 12.4) show. In general, there will be many transition matrices which share the same stationary distribution. One of these is obtained by replacing $\boldsymbol{P}$ by its transpose. If Eq. (12.4) has the same

solution when $\boldsymbol{P} = \boldsymbol{P}'$ this means that the development of the chain is the same backwards as forwards in time and the chain is then said to be reversible.

The link with our problem of estimating a posterior distribution is made as follows. We think of the desired distribution as (finely) grouped into 'cells', each of which can be identified with a state of the chain. We aim to construct a process which takes the individual through a sequence of such cells in such a manner that the proportion of times it is ultimately found in any one state approaches the desired posterior probability. The means and other statistics of the posterior distribution can then be estimated from the proportion of times the individual eventually reaches any cell.

If we can find an appropriate transition matrix (or kernel) and if we can show that the process we have constructed is a Markov chain, then the general theory outlined above ensures that a stationary distribution will be reached and maintained indefinitely. Constructing a Markov chain with the desired properties might seem rather challenging but is relatively straightforward. The idea is to define a process by which changes of state might be supposed to have occurred and then show that it is indeed a Markov chain with the desired steady stationary distribution. The details are spelt out in some detail in Bartholomew et al. (2011, pp. 30–33 and Sect. 4.11, pp. 102–107). Here we merely give an indication of how the argument goes.

Let the unobserved variables be denoted by the vector $\boldsymbol{v}$, which here includes both parameters and any latent variables. The 'cells' referred to above will each be indexed by a different value of $\boldsymbol{v}$. The Metropolis- Hastings algorithm then proceeds as follows. The algorithm constructs a sequence of estimates, $\mathbf{v}_{(0)}, \mathbf{v}_{(1)}, \ldots \mathbf{v}_{(t)}$, say, which converges to the desired estimates. The algorithm begins by choosing a starting value $v_{(0)}$. This starting value is converted into an improved value, $v_{(1)}$by a realisation of a stochastic process constructed as follows. Imagine an entity which, initially, is assigned the value $\mathbf{v}_{(0)}(0)$. This designates its initial 'cell' The entity is now supposed to move to another cell (or state) which is selected by the 'jump' distribution (which is arbitrary). The choice of destination is chosen by reference to the ratio of the current value of its estimated probability to the corresponding value at the origin sate. (Since the choice depends only on the ratio, it is not necessary to know the value of the normalising constant because this cancels, being the same for both.) The choice is constrained by the need to ensure that the 'origin' state would be the one chosen if the destination state had been the starting point. This introduces a symmetry which ensures the process will be reversible. It is obvious that the process is Markovian because the transition probabilities depend only on the current state. Taking an overall view of the process we have described, we start with an approximation to the posterior distribution, obtained by inserting the initial parameter values and then modifying it as the estimates of the parameters are gradually improved.

When the process has been running sufficiently long to be close enough to its limiting value, the location of the entity may be regarded as a sample of size one from the posterior distribution. From that point onwards it will move among the

states according to a Markov chain, but these values do not constitute a random sample from that distribution because they are not independent. To obtain a genuinely random sample, the whole process must be repeated. However, it is sometimes suggested that we would obtain something which would serve as a random sample by taking every kth value where k is small. Another important practical question is to decide when we are sufficiently near to the limiting state. Such questions have received a good deal of attention in the literature but are not germane to our present concerns.

12.4 Gibbs Sampling

This is a variant of the procedure described above which is designed to greatly reduce the complexity of the calculations. Instead of having to compute the posterior distribution at each step, which may involve many variables, it works on one variable at a time. To this end we need the marginal distribution of each individual element of **v** rather than the whole vector and this may be much easier to obtain. The procedure thus takes each variable in turn and then iterates to a solution as if that were the only unknown parameter. In a sense, therefore, we are embedding one iteration within another. This is the routine used in the widely used program WinBUGS.

Reference

Bartholomew, D. J. (2011). *Martin Knott and Irini Moustaki, latent variable models and factor analysis* (3rd ed.). Chichester: Wiley.

Chapter 13
Unity and Diversity

Abstract This chapter summarises the content of the Brief, focussing on the key idea of the posterior distribution of the unobserved variables. It gives a chapter by chapter summary and provides a rationale for the order in which topics have been presented.

Keywords Analysis of variance · Bayesian paradigm · Categorical variables · Exponential family · Factor analysis · General linear model · Latent variable models · Mixtures · Maximum likelihood · Missing data · Posterior distribution · Rasch model · Social measurement · Structural equation models · Time series

Here we take a retrospective look at how the subject has been developed throughout the Brief by emphasising those elements which are common but without neglecting the points of difference.

The unifying theme of this Brief is that the information about unobserved variables in a statistical problem is properly conveyed by their posterior distribution. However, there is considerable diversity in the ways in which such variables may arise, and when they do, on whether they have intrinsic meaning or are merely intermediaries leading on to some more important aspect of the problem. Such observations may be real in the sense that, in principle at least, they can be observed. This is the case as with the unobserved variables in a time series where the variables of interest have not yet occurred. Where observations have simply been lost, it is not those observations themselves that we are interested in but the parameters to whose estimation they might otherwise have contributed. At the other extreme, unobserved variables may be hypothetical because they have been introduced to accommodate some simplifying feature which makes the model more intelligible. The latter alternative is more appropriate in many applications in sociology or psychology where the model is constructed to give expression to some hypothetical entity such as intelligence. These different types of unobserved variable are reflected in the subjects of successive chapters. To some extent the order of the chapters is dictated by the subject matter but there is a degree of arbitrariness which may have been noted and which needs explaining.

D. J. Bartholomew, *Unobserved Variables*, SpringerBriefs in Statistics,
DOI: 10.1007/978-3-642-39912-1_13, © The Author(s) 2013

The first two chapters lay the foundations by defining the notation and setting out conventions and basic results. Chapter 1 also serves the purpose of a Preface by commenting on the style and limitations of our treatment.

Chapter 3 performs an important double function by first introducing mixtures. By this means it treats a problem which is of practical interest in its own right, but, secondly it also provides an introduction to the key idea of a latent variable which occurs later. As we noted earlier, the term 'unobserved heterogeneity' is sometimes used to designate variation in the quantity being mixed and so this chapter provides, almost incidentally, a first example of a latent variable model.

Chapters 4–6 are closely linked in that they represent a progression from a very simple type of latent variable problem to the full generality of factor analysis and its ramifications.

The key elements which link Chaps. 4 and 5 and which prepare the ground for Chap. 6 are the data matrix, given first in Chap. 1, and the class of probability models which lie behind them first given in Eq. (2.4). The data matrix of observed variables may be set out as follows

$$\begin{array}{cccc} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{array}$$

which represents m independent random samples each of size n. As we progress through the chapters the xs are first of all treated as binary, then continuous and then categorical. Their distributions are supposed to all be members of the one-parameter exponential family with probability function

$$f(x_i|\mathbf{y}) = F(x_i)G(\alpha_i)\exp(\alpha_i x_i) \tag{13.1}$$

with

$$\alpha_i = \alpha_i(0) + \alpha_i(1)y_1 + \alpha_i(2)y_2 + \cdots + \alpha_i(m)y_m \tag{13.2}$$

By appropriate choice of parameters this distribution was made to represent the various models required. By this means it was possible to include a range of models, starting with the Rasch model and leading on to a full factor analysis model via the idea of random effects model as used in the analysis of the general linear model. Because of its greater familiarity to statisticians it is convenient to approach the general linear latent variable model, as here, by way of the analysis of variance. The essential unity of the methods thus developed has sometimes been concealed by the diversity of terminology which has been introduced, often reflecting the language of different disciplines. In particular, statisticians brought up on a diet of analysis of variance may be very familiar with the idea of 'random effects' but may not realise that a latent variable is basically the same as a random effect. This equivalence may often not have been noticed because the common use of random effects seldom gets beyond means and variances. It follows that the subtleties concerning the form of the latent distribution do not arise.

Chapters 7 and 8 deal with topics—identification and categorical variables—which pervade many branches of statistics but which are particularly relevant here. Indeed it is especially the notion of a statistical model that is particularly relevant because many misunderstandings have arisen because such models have been neglected or ignored. Indeed, the idea of a model is central to the whole Brief but the unifying thread would be much weaker without it. The fact that the exponential family includes both categorical and continuous distributions serves to focus attention on structure of the problems rather than on the form which those representations take for particular kinds of variable. Normality plays a central role in the theory of Statistics. It lies behind much of the analysis of variance although it is seldom remarked upon. It turns out, and it is well-known, that normality and linearity are closely bound up together. This linkage comes out very clearly when we consider linear structural equations models as in Chap. 7. Such models are usually formulated as a system of linear equations connecting manifest and latent variables. From these models one may deduce the covariances between the manifest variables and the models are then fitted by choosing as estimates of the parameter those values which bring the observed and theoretical covariances as close together as possible. This needs no normality assumption but if we also introduce the assumption (unverifiable empirically, of course) of normal residuals, the estimates turn out to be maximum likelihood estimates. Looked at the other way round; we may first specify the model with normal residuals, and then obtain the maximum likelihood estimators. The latter can then therefore be justified, in the absence of the normality assumption, as those values which bring the observed and expected covariance into closest agreement.

One might have expected the subjects of Chaps. 9 and 10 to occur much nearer the beginning because both deal with unobserved variables in their most rudimentary form. In Chap. 10, on Missing Data we suppose that the missing variables are real enough and the only reason we do not know them because, in the most obvious sense, they have not been lost or never observed. Likewise, Time Series, treated in Chap. 9 is a longstanding member of the family of statistical methods but it is not ordinarily thought of as having anything to do with unobserved variables. But variables which are unobserved because they lie in the future are just as 'unobserved' at the time of the analysis as those which lie in the past.

Chapter 11 on Social Measurement, might seem to deal with topics which do not take us far beyond 'common sense' but the reason for their inclusion is because it is precisely since 'common sense' can sometimes be a very poor substitute for a well-thought-out model! The reason for leaving these various topics to the latter part of the Brief is that their place and relevance can be more easily recognised once the general framework is clear and the reason for setting them in this context is more readily apparent.

The final chapter provides a fitting conclusion by subsuming, conceptually at least, all unobserved variable problems within the Bayesian paradigm. However, it is important not to lose sight of the diversity of practical contexts which give rise to unobserved variables and for this the Bayesian paradigm has no particular regard. In fact we have argued that the traditional divisions into families of

methods of inference are of little relevance or importance in many practical contexts, especially in the common situation where the number of 'parameters' is large and, under which condition, the differences between the methods become negligible.

Although unobserved variables have been with us since the dawn of statistics, they have often been dealt with in an *ad hoc* fashion which, if anything, has concealed the essential unity of the problems and has given rise to many misunderstandings. By following the model-based approach to statistical theory they can be seen to fit into a simple framework which means that the many of these misunderstandings can be eliminated or avoided altogether. The diversity of the problems is, in a sense, accidental; the unity is fundamental.