Tamás Rudas

Lectures on Categorical Data Analysis

Springer

Tamás Rudas
Center for Social Sciences
Hungarian Academy of Sciences
Budapest, Hungary

Eötvös Loránd University
Budapest, Hungary

ISSN 1431-875X ISSN 2197-4136 (electronic)
Springer Texts in Statistics
ISBN 978-1-4939-7691-1 ISBN 978-1-4939-7693-5 (eBook)
https://doi.org/10.1007/978-1-4939-7693-5

Library of Congress Control Number: 2018930750

Preface

This book offers a fairly self-contained account of the fundamental results in categorical data analysis. The somewhat old fashioned title (Lectures ...) refers to the fact that the selection of the material does have a subjective component, and presentation, although rigorous, aims at explaining concepts and proofs rather than presenting them in the most parsimonious way. Every attempt has been made to combine mathematical precision and intuition, and to link theory with the everyday practice of data collection and analysis. Even the notation was modified occasionally to better emphasize the relevant aspects.

The book assumes minimal background in calculus, linear algebra, probability theory, and statistics, much less than what is usually covered in the respective first courses. While the technical background required is minimal, there is, as often said, some maturity of thinking required. The text is fairly easy if read as mathematics, but occasionally quite involved, if read as statistics. The latter means understanding the motivation behind constructs and the relevance of the theorems for real inferential problems.

A great advantage of studying categorical data analysis is that many concepts in statistics are very transparent when discussed in a categorical data context, and at many places, the book takes this opportunity to comment on general principles and methods in statistics. In other words, the book does not only deal with "how?", but also with "why?".

Hopefully, the book will be used as a reference and for self-study, and it can be used as a textbook in an upper division undergraduate or perhaps a first year graduate class, too. To facilitate the latter use, the material is divided into 12 chapters, to suggest a straightforward pacing in a quarter-length course. The book also contains over 200 problems, which are mostly positioned a bit higher than simple exercises. Some of these problems could also be used as starting points for undergraduate research projects. In a less theory-oriented course, when providing the students with computational experience is a direct goal, and it uses some of the instructional time, the material in Chaps. 3 and 7 and perhaps in Chaps. 8 and 9 may be skipped.

The topics emphasized include the possibility of the existence of higher order interactions among categorical variables, as opposed to the assumption of multivariate normality, which implies that only pairwise associations exist; the use of the δ-method to correctly determine asymptotic standard errors for complex quantities reported in surveys; the fundamentals of the main theories of causal analysis based on observational data presented critically; a discussion of the Simpson paradox and a description of the usefulness of the odds ratio as a measure of association, and its limitations as a measure of effect; and a detailed discussion of log-linear models, including graphical models. Chapter 13 gives an informal overview of many current topics in categorical data analysis, including undirected and directed graphical models, path models, marginal models, relational models, Markov chain Monte Carlo, and the mixture index of fit. To include a detailed account of all these in the book would have doubled not only its length but, unfortunately, also the time needed to complete it.

The material in this book will be useful for students studying to achieve different goals. It can be seen as sufficient theoretical background in categorical data analysis for those who want to do applied statistical research—these students will need to familiarize themselves with some of the existing software implementations elsewhere. For those who want to go in the direction of machine learning and data science, the book describes, in addition to many of the fundamental principles of statistics, also a big part of the mathematical background of graphical modeling—obviously, these students will have to continue their studies in the direction of the various algorithmic approaches. Finally, for those who want to be engaged in research in the theory of categorical data analysis, the book offers a solid background to study the current research literature, some of which is mentioned in Chap. 13.

I would greatly appreciate if readers notified me of any typos or inconsistencies found in the book.

Budapest, Hungary
September 2017

Tamás Rudas

Contents

Chapter 1
The Role of Categorical Data Analysis

Abstract Any real data collection procedure may lead only to finitely many different observations (categories or measured values), not only in practice but also in theory. The various relationships that are possible between the observed categories are defined in the theory of levels of measurement. The assumption of continuous data, prevalent in several fields of applications of statistics, is an abstraction that may simplify the analysis but does not come without a price. The most common simplifying assumption is that the data have a (multivariate) normal distribution or their distribution belongs to some other parametric family. Another type of assumption, made in nonparametric statistics, is that of a continuous distribution function and essentially implies that all the observations are different. These assumptions have various motivations behind them, from substantive knowledge to mathematical convenience, but often also the lack of existence of appropriate methods to handle the data in categorical form or the lack of knowledge of these methods. In many scientific fields where data are being collected and analyzed, most notably in the social and behavioral sciences but often also in economics, medicine, biology, and quality control, the observations do not have the characteristics possessed by numbers, and assuming they come from a continuous distributions is entirely ungrounded. Further, several important questions in statistics, including joint effects of explanatory variables on a response variable, may be better studied when the variables involved are categorical, than when they are assumed to be continuous. For example, when three variables have a trivariate normal distribution, then the joint effect of two of them on the third one cannot be different from what could be predicted from a pairwise analysis. But in reality, if multivariate normality does not hold, the joint effect is a characteristic of the joint distribution of the three variables. In such a case, the assumption of normality makes it impossible to realize the true nature of the joint effect. For categorical data, structure and stochastics in statistical modeling are largely independent and are studied separately.

This introductory chapter deals with general measurement and inferential issues emphasizing the importance of categorical data analysis. The first section outlines the basic theory of levels of measurement.

1.1 Levels of Measurement

Statistical models try to capture essential characteristics of the population where the data came from. The information with respect to properties of the population is supplied by the data, and before the substantive part of this information could be investigated by fitting statistical models, certain formal aspects of the data need to be settled. Data are usually stored in computers as numbers, but not all observations possess all the characteristics of numbers. For example, a variable may contain the gender of a respondent of a survey and may be coded as 1 = "male" and 2 = "female". Apparently, 2 is twice as much as 1, but in no sense of the word would be a woman twice as much as a man. In this case, the observations are not numbers rather are denoted by numbers. Statistical models are often formulated as a functional relationship between variables, and the operations that are meaningful for numbers are not necessarily meaningful for the categories of gender or of other variables. The labeling of the categories with numbers is often arbitrary. For example, the respondents may be asked to what extent they agree with a certain statement, with categories "strongly disagree", "disagree", "neither agree nor disagree", "agree", and "strongly agree". These categories may be coded with the numbers 1, 2, 3, 4, 5, but they could be coded with the numbers 2, 4, 6, 8, 10, just as well. However, when the second coding is used, the standard deviation appears to be twice as big as the standard deviation with the first coding. Thus, it is not only complex statistical models, but also simple statistics may be sensitive to whether or not the numbers in the data are the actual values or just notations for the possible values or categories of a variable.

A formal treatment of the different characteristics the measurements may have is given in the theory of levels of measurement. A minimal definition of measurements is that it is a procedure, whereby about each pair of objects is decided whether they are identical or different. The theory of levels of measurement associates various characteristics with the differences that may exist between objects.

1.1.1 Categorical or Nominal Level of Measurement

In categorical, or nominal, measurement, one can decide whether two objects are identical, but if they are not, nothing can be said about the nature of their difference. For example, classifying respondents according to their gender, or classifying researchers according to the university where they obtained their PhD degree, are categorical measurement. Such data are sometimes referred to as count data. This kind of measurement is most common in the social and behavioral sciences but also occurs in many other fields where statistics is being used. In a trial to investigate the survival chances of patients who have undergone a certain medical treatment, the researchers classify the patients into categories "survived" or "deceased" after a certain number of years. Or quality control may classify the products of a factory as "perfect" or "defective". Two products, one in the "perfect" category and one in the "defective" category, are different, but this classification does not tell in what aspect is the second product defective. Of course, the definitions of these categories need to describe all the reasons for which a product may have to be classified into the latter category. In categorical measurement, two objects may be deemed not identical, but no information is given with respect to the nature of the difference.

In nominal measurement, the categories have to be disjoint and exhaustive, so that every observation falls into exactly one category. When data are collected from people using a survey, it is often not easy to set up an exhaustive system of categories in advance. Sometimes, in a so-called pilot study, the question is asked without response categories being offered (open-ended question), and the information collected is used to set up an appropriate system of categories. It is also customary to include an "other" option to make the system of categories exhaustive.

The categories used in nominal measurement are often denoted by numbers, but the choice of the numbers is entirely arbitrary. One has to pay attention that a category denoted by, say, 2 is not, in any sense of the word, twice as much as the category denoted by 1, although this would be true for the numbers themselves. Consequently, many of the common statistics used to describe the data have no real meaning and neither do many of the standard statistical models that posit a functional relationship among variables. If, for example, a variable denotes the region of the USA where a person was born, with categories East, South, Midwest, and West, then even if the categories are denoted by numbers, the average or the standard deviation of the variable obtained makes little sense. Using the numbers 1, 2, 3, 4 is not any better justified than using the numbers 7, −2, 11, 12, and the mean or the standard deviation will depend strongly on this choice. Similarly, in a regression analysis, one may assume that for variables X and Y,

$$E(Y|X=x) = ax + b,$$

for some constants a and b, but when X is region where the respondent was born and Y is the university where the PhD degree was obtained, assuming that

$$E([\textit{region where born}] \mid [\textit{university where PhD was obtained}]) =$$

$$a\,[\textit{university where PhD was obtained}] + b,$$

makes very little sense.

Therefore, while the research questions statisticians are faced with tend to be the same irrespective of the level of measurement of the variables available, the methods to answer these questions and the statistical models applied need to be different. This book presents statistical methods that are appropriate to analyze categorical variables.

1.1.2 Ordinal Level of Measurement

The ordinal level of measurement possesses the characteristics of categorical measurement, and, in addition, there is an ordering among the categories. That is, in addition to being different, one can also tell which one is larger, better, more expensive, or whatever is the aspect of ordering. For example, respondents in a survey are often asked about the highest degree they have obtained. The categories for the possible responses may be defined as elementary, high school, college, and graduate degree. It is clear that someone with a college degree has a higher level of education than someone who only finished elementary school, but the measurement itself does not tell anything about the nature of this difference. For example, no answer is given to the question how much higher. Or quality control may classify the products of a factory as "perfect" or "minor defect, may be sold at discount", or "defective". Two products, one in the "perfect" category and one in the "defective" category, are different; obviously a product in the perfect category is better than a product in the defective category, but ordinal measurement does not tell in what aspect is the second product defective or how much worse is the defective product than the perfect one.

It may look very attractive to assign numbers to the categories of the educational level variable, for example, 1, 2, 3, 4, as is often done in practice, and analyze the data based on that coding. But, again, similarly to the case of nominal measurement, the numbers 1, 2, 3, 4 have many characteristics not shared by the categories of the variable. For example, the numbers suggest that the difference between someone with a college degree and with an elementary degree is the same (namely, 2), as the difference between someone with a graduate degree and someone with a high school degree. Whether this is really the case may be considered as a serious research question but should not be assumed as an implication of the notation. Educational attainment may also be measured by the number of years of education completed. In this case, the categories are numbers and still do not necessarily possess all the characteristics of numbers. For most college graduates, it takes 16 years of schooling to obtain that degree. Can one say that the difference between individuals with 15 and 14 years of schooling is the same as the difference between individuals with 16 and 15 years of schooling? Well, on the one hand, one cannot deny that $15 - 14 = 16 - 15$. On the other hand, those with 14 or 15 years of schooling have not obtained a college degree, while those with 16 have, so the difference between 16 and 15 years of schooling seems more substantial than between 15 and 14 years. This is because the number of years of schooling in this example may not be relevant in itself rather as an indicator of educational attainment. The numbers of years of schooling completed behave as a number does, but educational attainment remains measured on an ordinal scale. This example also illustrates that the true level of measurement of a variable often cannot be decided just by the inspection of the measurement procedure that led to that variable, but the intended use of the variable in the analysis should also be taken into account.

1.1.3 Interval Scale

The next level of measurement is called interval scale. It is called a scale to refer to the fact that its properties are closer to those usually associated with standard measurement. If two objects are not identical from the perspective of the measurement, then there is an ordering between them, and, in addition, one can tell how much one of them is larger than the other one. In other words, differences between the categories are meaningful. The most common example of such a measurement is using centigrades to measure temperature. The difference between 2 and 1 °C is the same as between 7 and 6 °C, namely, 1 °C, but one cannot say that 2 °C is twice as warm as 1 °C, because the starting point of the scale was somewhat arbitrarily selected as the freezing temperature of water and cannot be identified with no heath, i.e., it does not constitute a real zero point. In general, measurements with an unspecified or arbitrarily selected zero point are on an interval scale. Examples include altitudes at which airplanes fly or longitudes on the Earth. Altitudes are usually given with reference to the mean sea level, and a plane flying at 8000 ft may not fly twice as much high measured from the ground than a plane at 4000 ft. A city at 60°E is not twice as much to the East as a city at 30°E, because counting starts at an arbitrarily selected position (Greenwich).

Readers of this book will often analyze data form surveys, where a popular method to measure attitudes or opinions is the application of a so-called Likert scale. This means asking respondents to express their levels of agreement with an item by choosing a number from 1 to 5 or from 1 to 7 or from 0 to 10. It is often assumed by survey scientists that this measurement constitutes an interval scale, that is, not only is, say, 8 a stronger agreement than 7, but also the difference between the levels of agreements expressed as 8 and 7 is the same as the difference between those expressed as 3 and 2. In most cases, this is only hoped to be the case, and to determine whether or not this assumption is true would require very involved methods.

Although it is quite common to denote the categories of an interval scale with numbers, just like in the case of all examples given so far, the categories do not possess all the characteristics of numbers, because, as a consequence of arbitrary zero points, ratios are not meaningful. The arbitrarily selected starting point means that if the values associated with the categories are $\{s_1,\ldots,s_k\}$, then for an arbitrary t, the values $\{t+s_1,\ldots,t+s_k\}$ could just as well be used. Therefore, any statistic that is invariant against this transformation, e.g., the standard deviation, is meaningful, but statistics that are sensitive to it, e.g., the mean, should be used with caution or not at all.

1.1.4 Ratio Scale

In the case of ratio scales, in addition to the characteristics of the previous levels of measurements, one can also say how many times a category is larger than another one. Categories of ratio scales possess all characteristics of (real) numbers, and the measurement results may be identified with numbers. In fact, the traditional concept of measurement involves a unit, and each object is compared to the unit. For example, the length of a desk may be compared to a yardstick or any unit of length and then told how many times (allowing fractions) the unit fits into the length of the desk. The foregoing examples were meant to illustrate that much of the data used in statistics are not results of such measurement procedures rather of more restricted procedures that have different characteristics, and these need to be taken into account when computing summary statistics or designing statistical models to be fitted to the data.

1.1.5 Changing the Level of Measurement

Analysts often wish to have data measured on a higher level of measurement than the actual data they have. Higher levels of measurement convey more information, than lower levels do, and one's favorite statistical method may be applicable to data at higher levels of measurement but not to data at lower levels of measurement. For example, it is customary (although not always with good reason) to assume that the categories of a Likert scale are equally spaced so that it is measurement on an interval scale, but it would be even better if the same measurement would be on a ratio scale. Then, one would also know that someone whose response to the question "to what extent do you agree with the following statement …?" is, say, 8, agrees twice as much as a person whose response is 4. Further, in this case, the agreement with the statement in the question could be used as a response variable in a linear regression, and perhaps good fit could be achieved, when approximated by some other variables.

Usually, there are good reasons why a variable is at the level of measurement where it is. Because the equator is (topologically) a circle, there is no natural starting point to measure how much to the East or West a location is; thus there seems to be no easy way to replace the system of longitudes (measured on an interval scale) by a "better" measurement on a ratio scale. Similarly, in a survey, the respondents may be asked to mark their levels of agreement with an item so that they reveal their feelings on a ratio scale, but, in fact, there is no guarantee at all that any set of instructions could lead to a measurement on the ratio level.

Certain types of data may only be accessible at lower levels of measurement. They cannot be transformed to a higher level of measurement by simply pretending – as one can so often see in applied data analysis – that the numbers that are used to denote the categories are actual numbers. This practice imputes the information not present in the measurement in an uncontrolled and ad hoc way. When information not present in the measurements is available from other sources, one may try to update the measurements with reference to this additional information. For example, one may try to find an embedding of the categories used in a Euclidean space of appropriate dimension subject to some externally defined property. Such methods are called scaling, see, e.g., [32], and will not be considered in this book.

Occasionally, data measured on a higher level of measurement are considered as if they were measured at a lower level. This results in loss of information and often involves combining the categories into fewer ones. The advantage of such procedures is that fewer parameters need to be estimated (see the discussion in the next section) and may be necessary to allow for similar analyses of a number of variables that have originally been measured at different levels.

1.2 Categorical or Continuous Data

Observations in statistical analysis are modeled by random variables. Usually, a random variable is defined as having values in some subset of the set of real numbers and being a measurable[1] function from a measure space into this subset. The behavior of a random variable X is described by its distribution function F_X, which, for every real a, gives the probability of the event $\{X \leq a\}$. When X can have any value on an interval (or half line or the entire real line) and all these values are taken up with probability zero, then F_X is a continuous function and X itself is also called continuous.

For a variable to be continuous, it has to have infinitely many values. More precisely, it has to have as many values as there are on an interval (called continuum cardinality). Obviously, there is no real data gathering procedure that could yield that many different results. Even if there was any physical procedure that could lead to infinitely many different values, one would not be able to record each and every one of them. Data collection can only lead to a finite number of different observations. This number may be very large and can be made larger by increasing the precision of the observations but will always be finite. Consequently, the idea of a continuous random variable, applied so often in statistical modeling[2], is an abstraction or approximation, the advantages and disadvantages of which need to be assessed, before the decision is made to approximate the data generating mechanism by a continuous random variable.

An apparent relaxation of the continuity assumption is to assume that the random variable may take on an infinite number of different values, not as many as the number of points on an interval but as many as the number of integers (called countable cardinality). Such a variable is usually called discrete, in the sense that its values may be ordered to have neighbors (like integers do). Such a discrete variable is a good model of many data collection procedures, in which there is a fixed unit of measurement (like millimeters to measure the length of a desk or dollars to measure someone's yearly income), but there is no theoretical maximum of the observable

[1] Understanding the concept of measurability, see, e.g., [16], is not needed to follow the presentation in the book, and this aspect will be suppressed later on.

[2] The use of continuous variables to model observations needs to be clearly distinguished from using continuous random variables to describe, e.g., the asymptotic behavior of test statistics. The latter is a mathematical construct that may well be continuous. In fact, asymptotic distributions that are continuous are very important in the analysis of categorical data.

values. Indeed, it would be very difficult to tell how many millimeters long is the longest desk and how many dollars is the largest yearly income. Yet, one may be quite certain that none of these figures can be higher than, say, the legendary[3] value of 2^{64} – leading, again, to finitely many different observations.

In every data collection procedure, the number of possible different observations is finite. In cases when this number is large, it is difficult, often impossible, to reliably estimate the probability of each outcome. That would require unrealistically large sample sizes. Very often, therefore, the assumption is made that the observations come from a continuous distribution that is characterized by a few parameters. When such an assumption is made, instead of the hundreds or thousands of individual probabilities, only the parameters of the distribution need to be estimated form the data. For example, if one observes 4 variables, each with 50 possible values, one has 6.25 million different outcomes, and some of these may be very unlikely, so a huge sample would be needed to estimate all these probabilities. If the assumption of, say, multivariate normality is made, one needs to estimate 4 expectations, 4 variances, and 6 covariances, 14 parameters in total, which may be done well, even if having a small or moderate sample size. In addition, in many fields of applications of statistics, experience or substantive arguments support the assumption of normality. The assumption of multivariate normality is, of course, a much stronger assumption than that of univariate normality of each of the variables involved, and tests of multivariate normality to have acceptable power may also need very large sample sizes.

The assumption of normality, although may greatly simplify (occasionally oversimplify) the analysis, is only meaningful if the variables concerned are measured on the ratio level. With lower levels of measurement, the categories do not possess the characteristics of numbers, and a real-valued random variable (e.g., one with a normal distribution) is not a good model.

A random variable is categorical, if it can take on values in a finite set. The set does not have to be a subset of the real numbers. Examples include variables taking values in the sets $\{male, female\}$ or $\{voted\ republican, voted\ democratic,$ *voted other*, *did not vote*$\}$. The usual measurability requirement is nil in this case, as probability may be associated with all subsets of a finite set. A categorical random variable is a good model of measurement at the nominal level. Categorical random variables may also be used when the measurement process occurred at a higher level of measurement, but the additional properties of the higher levels are disregarded, either to simplify the statistical analysis or because those additional features are not considered valid or reliable. This book concentrates on statistical methods for the multivariate analysis of categorical variables.

[3] The legend is about the inventor of the game of chess, who, as a present from the extremely pleased emperor, asked one grain of rice for having invented the first square of the chessboard, twice as many for the second one and always twice as many as for the previous one till the last one. This seemed like a modest compensation for the emperor but, in fact, amounted to $2^{64} - 1$ grains, by magnitudes more than the total number of atoms in the entire solar system.

When a variable is considered categorical, its distribution may be described by a finite number of parameters, and yet this approach implies no restriction on the distribution the variable may have. This is in contrast with the continuous case, where to be able to describe the distribution with a finite number of parameters, strong assumptions concerning a parametric family were needed to be made.

For the analysis of ordinal data, one may also consider nonparametric methods that do not rely on the actual values of the observations rather only on their ranks among all observations [52]. Unfortunately, many of these methods make use of the assumption that observing the same value twice has zero probability (no ties). This assumption is appropriate when ranks are derived for observations from a hypothetical continuous random variable (rather, a categorical variable with very many categories), but when there are only a small number of ordered categories possible, it does not seem to be an appropriate assumption. If, say, a Likert scale has 7 categories and the sample size is 1000, one cannot hope not to see ties.

Another interesting approach is to assume that a variable measured on the ordinal level is the manifestation of a continuous variable through certain cut-points. Every observable category is equivalent to the value of the unobservable (latent) variable being between two adjacent cut-points. For example, it may be assumed that job satisfaction is a continuous characteristic, and respondents are asked to report their positions on a Likert scale when prompted with the question “How happy are you with your current job?”. In such cases, some effort to recover certain properties of the underlying continuous variable may be made. Unfortunately, without making further assumptions about the latent variable, few of its characteristics can be deduced. For example, a continuous uniform latent variable, by appropriate choice of the cut-points, may be transformed into a unimodal or a bimodal ordinal variable, just like an underlying variable with a normal distribution may be cut into a bimodal or into a highly skewed distribution. There are situations, however, when knowledge available about the latent variable may be reliably incorporated into the analysis. For example, in the medical and psychological literature, it is often assumed that a latent trait is not only continuous but also normally distributed in the population but may only manifest itself if its value exceeds a threshold. This assumption is called the threshold model.

Sometimes, the expression of dichotomy of numerical versus categorical variables is used. The same concept is also referred to by the names of quantitative versus qualitative variables. In most cases, authors identify these concepts with ratio or interval scales and ordinal or categorical levels of measurement. The dichotomy is less precise than the categorization into four levels of measurement. The position taken in this book is that – as the precise level of measurement of a variable often depends on its intended role in the analysis – the statistician may make decisions as to what characteristics of the categories of a variable to rely on. A minimal assumption is that of a categorical level of measurement. The advantages and disadvantages of such an assumption need to be evaluated on a case-by-case basis.

1.3 Interaction in Statistical Analysis

The decision about categorical or continuous modeling of the variables is further motivated by the fact that the choice of continuous variables and the most often implied choice of multivariate normality, independently of its appropriateness from a substantive point of view or of the level of measurement defined by the data gathering procedure, also implies a simplification with respect to the statistical models that may be analyzed. To illustrate this point, consider a regression-type problem with response variable Y and explanatory variables X and Z. The research problem is called a regression-type problem, instead of a regression problem, to emphasize the fact that the variables concerned have predefined roles, but they are not necessarily continuous or multivariate normal, in which case a standard (linear) regression analysis might be appropriate. Rather, the possibility of treating them as categorical and the advantages and disadvantages of this choice are being investigated.

As in any regression-type problem, one wants to respond to three questions:

1. Which of the potential explanatory variables have an effect on the response variable?
2. Out of those explanatory variables which do have an effect on the response variable, which ones have strong and which ones have weak effects?
3. If there are several explanatory variables which have an effect, is their joint effect different from what one would expect based on their separate effects?

The methods that can be applied to answer these questions will be discussed in detail in Sect. 11.1, and we concentrate here on the third question which is, no doubt, the most intriguing out of the three. In fact, even the precise meaning of this question may require some clarification.

1.3.1 *Joint Effects in a Regression-Type Problem Under Joint Normality*

Assume first that $(X,Y,Z)' \sim N(\mu,\, \Sigma)$, with $\Sigma = (\sigma_{ij})$. Then, the conditional expectations of the response are as follows:

$$E(Y|X=x) = \mu_1 + \sigma_{12}\sigma_{22}^{-1}(x-\mu_2), \tag{1.1}$$

$$E(Y|Z=z) = \mu_1 + \sigma_{13}\sigma_{33}^{-1}(z-\mu_3), \tag{1.2}$$

and

$$E(Y|X=x, Z=z) = \mu_1 + \Sigma_{1(23)}\Sigma_{23}^{-1}((x,z)' - (\mu_2,\mu_3)'), \tag{1.3}$$

with

$$\Sigma_{1(23)} = (\sigma_{12}, \sigma_{13})$$

and

$$\Sigma_{23}^{-1} = \begin{pmatrix} \sigma_{22} & \sigma_{23} \\ \sigma_{32} & \sigma_{33} \end{pmatrix}^{-1} = C \begin{pmatrix} \sigma_{33} & -\sigma_{23} \\ -\sigma_{32} & \sigma_{22} \end{pmatrix},$$

where $C = 1/(\sigma_{22}\sigma_{33} - \sigma_{23}\sigma_{32})$.

As is well known, under multivariate normality, the conditional expectation of a variable given the values of all other variables is a linear function of the conditioning values. This is the basis of the linear model used almost exclusively to analyze variables with a joint normal distribution. The assumption of joint normality, however, has further implications on how the variables affect each other.

If the two explanatory variables X and Z are independent, that is, when $\sigma_{23} = \sigma_{32} = 0$, the last formula becomes

$$C \begin{pmatrix} \sigma_{33} & -\sigma_{23} \\ -\sigma_{32} & \sigma_{22} \end{pmatrix} = \frac{1}{\sigma_{22}\sigma_{33}} \begin{pmatrix} \sigma_{33} & 0 \\ 0 & \sigma_{22} \end{pmatrix},$$

so that

$$E(Y|X = x, Z = z) =$$

$$\mu_1 + (\sigma_{12}, \sigma_{13}) \frac{1}{\sigma_{22}\sigma_{33}} \begin{pmatrix} \sigma_{33} & 0 \\ 0 & \sigma_{22} \end{pmatrix} ((x,z)' - (\mu_2, \mu_3)') =$$

$$\mu_1 + \sigma_{12}\sigma_{22}^{-1}(x - \mu_2) + \sigma_{13}\sigma_{33}^{-1}(z - \mu_3) =$$

$$E(Y|X = x) + E(Y|Z = z) - E(Y),$$

using (1.1) and (1.2). So if the effect of a set of explanatory variables is identified with the deviation of the conditional expectation of response given the set of explanatory variables from the unconditional expectation, then the effect of X on Y is

$$E(Y|X = x) - E(Y) = \sigma_{12}\sigma_{22}^{-1}(x - \mu_2)$$

and the effect of Z on Y is

$$E(Y|Z = z) - E(Y) = \sigma_{13}\sigma_{33}^{-1}(z - \mu_3).$$

The joint effect of X and Z on Y is

$$E(Y|X = x, Z = z) - E(Y) = \sigma_{12}\sigma_{12}^{-1}(x - \mu_2) + \sigma_{13}\sigma_{33}^{-1}(z - \mu_3),$$

which is the sum of the individual effects of X and Z.

This leads to the following result:

Proposition 1.1. *In regression analysis under trivariate normality, if the two explanatory variables are independent, their joint effect on the response variable is equal to the sum of their individual effects.* □

This may seem quite obvious. If two variables are independent, how could they have a joint effect which is different from the sum of their individual effects? We will see in the next subsection that this may happen.

Further, if $\sigma_{12} = \sigma_{13} = 0$, then the second term in (1.3) is zero, implying the following result:

Proposition 1.2. *In regression analysis under trivariate normality, if both explanatory variables are independent from response, they do not have a joint effect on the response variable.* □

1.3.2 Joint Effects in a Regression-Type Problem with Categorical Variables

The situation described in the previous subsection is thought to be so obvious by many that the facts to be illustrated next may be considered as paradoxical.[4] Although regression analysis is usually interpreted as a method to analyze continuous, in particular normal, variables, researchers often face a similar problem when the variables are not continuous. The questions listed earlier in this section may be just as important, when the response and explanatory variables are categorical. A more detailed discussion of the regression problem for categorical variables is given in Chap. 11; here only some of the important properties are illustrated.

Let now the response variable be C and the explanatory variables A and B. The notation is different from the one used with continuous variables to emphasize the fact that these variables are categorical. The behavior of a binary variable is often described by the odds of one of its categories against the other category. The odds is the ratio of the two probabilities associated with the two categories of the variable. If the two categories are denoted as 1 and 2, the value of the odds for the response variable is $P(C=1)/P(C=2)$. The effect of an explanatory variable on the response can be best seen by comparing the odds of the response variable for those in different categories of the explanatory variable. If the two conditional odds are equal, the explanatory variable has no effect on response. A comparison of the two conditional odds is the odds ratio

$$\frac{P(C=1|A=1)/P(C=2|A=1)}{P(C=1|A=2)/P(C=2|A=2)} = \frac{P(C=1,A=1)P(C=2,A=2)}{P(C=1,A=2)P(C=2,A=1)},$$

shown here for the effect of A on C. The meaning, use, and properties of the odds ratio will be discussed in detail in Chap. 6; see also [72]. Conditional odds and odds ratios may be applied to handling the regression problem with categorical variables. As an example, a possible joint distribution of variables A, B, and C is shown in Table 1.1.

The marginal distribution of A and B derived from Table 1.1 is uniform, which implies that the variables A and B are independent. In spite of this, they do have a joint effect on C in the sense that the joint effect of A and B on C is not obtained from

[4] A paradox, of course, only means that the facts deviate from our expectations. Occasionally, our expectations prove to be ungrounded. For some of the paradoxes associated with probability, see [85].

Table 1.1 $2 \times 2 \times 2$ distribution with A and B independent and having a joint effect on C

	$C=1$		$C=2$	
	$B=1$	$B=2$	$B=1$	$B=2$
$A=1$	0.05	0.05	0.2	0.2
$A=2$	0.15	0.2	0.1	0.05

the individual effects by any simple rule. Indeed, from the $A \times C$ marginal table, one obtains that

$$P(C=1,A=1)/P(C=2,A=1) = 1/4$$

$$P(C=1,A=2)/P(C=2,A=2) = 7/3$$

and from the $B \times C$ marginal table, one obtains that

$$P(C=1,B=1)/P(C=2,B=1) = 2/3$$

$$P(C=1,B=2)/P(C=2,B=2) = 1.$$

The values of the conditional odds of C given both A and B are neither the products nor the sums nor any simple functions of the conditional odds above:

$$P(C=1,A=1,B=1)/P(C=2,A=1,B=1) = 1/4$$

$$P(C=1,A=1,B=2)/P(C=2,A=1,B=2) = 1/4$$

$$P(C=1,A=2,B=1)/P(C=2,A=2,B=1) = 3/2$$

$$P(C=1,A=2,B=2)/P(C=2,A=2,B=2) = 4$$

This shows that the joint effect of the two explanatory variables cannot be derived from their individual effects, rather they have a joint effect which needs to be determined separately, and Proposition 1.1 cannot be generalized to categorical variables.

Proposition 1.3. *When effect is measured by the conditional odds, categorical explanatory variables that are independent from each other may have a joint effect on a categorical response variable.* □

The data set in Table 1.2 is another example of the case described in Proposition 1.3. Further, it also shows a situation where neither A nor B has an effect on C. Indeed, both the $A \times C$ and the $B \times C$ marginals are uniform. In spite of this, A and B have a joint effect on C. Consequently, Proposition 1.2 cannot be extended to categorical variables.

Proposition 1.4. *When effect is measured by the conditional odds, categorical explanatory variables that individually have no effect on the response variable may have a joint effect on it.* □

Table 1.2 $2 \times 2 \times 2$ distribution with no A or B effect on C but joint A and B effect on C ($0 < a < \frac{1}{8}$)

	$C=1$		$C=2$	
	$B=1$	$B=2$	$B=1$	$B=2$
$A=1$	$\frac{1}{8}+a$	$\frac{1}{8}-a$	$\frac{1}{8}-a$	$\frac{1}{8}+a$
$A=2$	$\frac{1}{8}-a$	$\frac{1}{8}+a$	$\frac{1}{8}+a$	$\frac{1}{8}-a$

The structure of multivariate categorical data, consequently, may be more complex than that of multivariate normal data. More precisely, the assumption of multivariate normality imposes a strong simplification on the structure. When multivariate normality is assumed, Propositions 1.1 and 1.2 hold. When the data are categorical, the situations described in Propositions 1.3 and 1.4 may occur. The mechanism through which this simplification happens is that when multivariate normality is assumed, only means, variances, and covariances are estimated from the data and the analysis applies to a multivariate normal distribution with these parameters. These parameters may be estimated from pairwise bivariate distributions. Different data sets for which these parameters are the same will lead to exactly the same analyses, even if they differ in the empirical variant of, say, the following third-order mixed moment or interaction term:

$$\int\int\int XYZdxdydz.$$

Similarly, in the case of categorical data, the odds ratios between any two of the variables contain no information with respect to the following interaction among all three variables:

$$\frac{P(1,1,1)P(1,2,2)P(2,1,2)P(2,2,1)}{P(2,2,2)P(1,1,2)P(1,2,1)P(2,1,1)}, \tag{1.4}$$

which, for the data in Table 1.2, is

$$\frac{(1/8+a)^4}{(1/8-a)^4},$$

while the bivariate odds ratios all have the value of 1, irrespective of a. The quantity in (1.4) is the second-order odds ratio of the three variables and will be studied further in Chap. 6.

In the relatively rare cases when the data are *known* to have a multivariate normal distribution, the value of the above moment does not contain information in addition to the bivariate distributions: if the latter are known, the moment may be determined. It is an anomaly, if its value is different from what is implied by the bivariate distributions. But if normality is not known, by assuming it, the analysts disregard a potentially important feature of the data, one that is not recognized by looking at only bivariate distributions.

This implication of the assumption of multivariate normality seems more problematic than the inappropriateness of normality for data measured at lower levels of measurements.

1.4 Structure and Stochastics in Statistics

The separation of statistical activities into those dealing with structure and those dealing with stochastics is not exclusive for categorical data, but this separation is clearer and more useful in this case than under the assumption of multivariate normality. Statistical analyses aim to make inference with respect to a population underlying the observed data. In some applications of statistics, it is natural to assume that there is a population of individuals and the data were produced through observing a subset of this population, the sample. This is the case, for example, when a sample of the citizens of a country is interviewed in a survey concerning their intentions to vote at the upcoming elections. In other applications, the assumption of a preexisting population is less natural, and one rather assumes that the subsequent observations are produced by a data-generating mechanism. This is the natural view when, for example, a quality control engineer may inspect every n-th product that comes off an assembly line. These two situations are not fundamentally different, and in the sequel they both will be referred to by saying that the goal of the statistical analysis is to make inference with respect to a population.[5] The voting intentions of every citizen are supposed to exist before data are collected, and they constitute the population of interest. The future products of the assembly line do not exist when one of them is being investigated, and in such cases one usually considers all values that could be yielded by the data-generating mechanism, that is, all the products that could be manufactured on the assembly line as the population of interest.

Statistical inference concentrates on two questions. First, what statistical models are of interest and second, whether an interesting model fits. These are very different questions and should be treated as such. The assumption of multivariate normality largely determines both of these aspects. Multivariate normality specifies the parameters of interest and in most cases also the relevant structural models. For example, if a researcher has a regression-type problem, the assumption of multivariate normality implies that the regression is linear (because of the structure of the conditional means). When the population of interest is assumed to have a multivariate normal distribution, the data consist of observations from a multivariate normal distribution, so the sampling distribution is also multivariate normal. This implies optimality properties of certain estimation and testing procedures and also implies the distribution of certain test statistics. When working with categorical data, the two components of statistics are less strongly connected, and the structural and stochastic parts can often be considered separately from each other.

[5] See Chap. 2 for a more detailed discussion of the relationship between sampling procedures in these two situations.

In the categorical data framework, the assumption about the distribution of the population is not restrictive and is fully parametric. Models are specified by postulating certain characteristics of the population distribution. Most of the useful models assume some kind of simplicity. For example, when the population is characterized by two variables, their independence is a useful model. If independence holds, the structure of the population is simpler than in the case when the two variables are not independent. Simplicity may be interpreted as simplicity from a substantive point of view, taking into account the actual meanings of the variables, but also a more technical view may be taken that considers a model simpler than another one if the former one has fewer parameters. For more than two variables, the so-called log-linear models generalize independence in various ways, all assuming a certain simple structure among the variables. These models and some of their further generalizations (like graphical models, marginal models, relational models) are discussed in later chapters of this book. Statistical models for categorical data are often described with algebraic tools that are useful in understanding equivalent characterizations and implied properties of the models, which, in turn, are fundamental in choosing the appropriate model for a substantive problem and in giving its proper interpretation, which is a very important part of any statistical analysis.

The choice of the statistical model, which postulates desirable characteristics of the population, in the case of categorical data, has little implications on the stochastic part of the analysis. The sampling distribution is not implied by the choice of the model, rather it is determined by the sampling or observational procedure applied. Almost any population may be observed in ways that the sampling distribution will be multinomial or product multinomial or Poisson. The choice of the sampling procedure depends basically on the population to be studied and not on the model that is assumed to describe it. These sampling distributions will be discussed later in the book. Optimality of certain estimation procedures is a consequence of the sampling distribution and not of the model investigated, and so are the distributions of many test statistics.

In the case of categorical data, procedures for estimation and testing are also relevant without reference to a particular model, and the study of simple structures, that is, of statistical models, constitutes a separate and very important part of the existing knowledge. Accordingly, structure and stochastics are discussed in separate chapters or sections in this book. Chapters 2, 3, 4, 5, and 12 mostly deal with stochastics, and the remaining chapters mostly deal with structure.

1.5 Things to Do

1. What is the level of measurement of someone's income? What is the level of measurement of someone's income as an indicator of material well-being? What is the level of measurement if income is used as an indicator of the amount of disposable income?

2. Many universities use letter grades to assess student performance but produce a grade point average based on numbers associated with the letter grades. For example, an A- is associated with a value of 3.67 for the calculation of the mean. What are the respective levels of measurements? Under what assumptions is this procedure correct?
3. Suppose that a variable is measured on the ratio level but, of course, not without errors. It may be assumed that the error is a random variable with distribution $N(0,\sigma^2)$. Generate a rule to transform the variable into an ordinal variable with three categories so that the loss of information is minimal. How can the loss of information be measured here? Is something gained as a result of this transformation? Why is such a transformation determined by two cut-points?
4. The previous transformation in the previous item may result in very unequal numbers of observations in the three categories. An alternative method is to cut after the first third and before the last third of the observations. What are the advantages and disadvantages of this method?
5. Read about some of the properties and implications of discretizing continuous variables in [90].
6. Let X have the uniform distribution $U(a,b)$. Define cut-points $(c_1,\ldots,c_4)$ so that the resulting ordinal variable with five categories has the distribution $(p_1,\ldots,p_5)$, where the probabilities sum to 1. Repeat the exercise for $X \sim N(\mu, \sigma^2)$. Can one find out a and b or μ and σ^2 based on the observed frequencies $(f_1,\ldots f_5)$? Is maximum likelihood estimation of the parameters possible? Develop the likelihood function in both cases.
7. In spite of the negative results of the previous item, under strong assumptions for the latent (unobserved) variables, some of their characteristics may be estimated using the observed data. For example, if two latent variables with a joint bivariate normal distribution are cut into two binary variables during the observational process, their correlation coefficient may be estimated from the observed cross-classification of the binary variables. Read [89] about the tetrachoric correlation and related quantities.
8. Develop a derivation similar to that of leading to Proposition 1.1 in the case of three explanatory variables.
9. Show that the odds ratio is equal to 1 if and only if the conditional distribution of the response given a category c of the explanatory variable does not depend on c.
10. Categorical variables may be combined to give a new categorical variable with more categories. For example, A and B in Table 1.1 may be combined to yield a variable with four categories. Explain how the joint effect of A and B in Propositions 1.3 and 1.4 relates to the effect of the combined variable.
11. Suppose that to follow the course STAT520, the material covered in STAT450 and in STAT490 is helpful, but none of these classes are required to take STAT520. Assume further that the odds of getting an A in STAT520 (as opposed to getting another grade) is twice higher for those who have taken STAT450 than for those who have not taken this class. The same number for STAT 490 is 3. Based on this information, can one tell how many times is the odds of getting

an A (as opposed to any other grade) in STAT520 higher for those who have taken both STAT450 and STAT490, than for those who have not taken any of these classes?

12. When would you say in the situation described in the previous item that the effects of taking STAT450 or STAT490 are independent?
13. For two categorical variables, A and B, the model of independence is that $P(A=a, B=b) = P(A=a)P(B=b)$ for all of their categories a and b. When A is gender of the respondent and B is the respondent's reported voting behavior, explain why a population with independence of A and B is simpler than a population where independence does not hold.
14. In the above example, A has two categories, and B may have four categories. How many parameters are needed to describe the population with and without independence?
15. Assume that someone believes the population is characterized by a model that has ten parameters, while another researcher believes that another model, with 25 parameters, is the correct description of the population. If the researchers want to estimate the population parameters with the same precision, which one will require a larger sample size? What do you think, does sample size affect the cost of data collection?

Chapter 2
Sampling Distributions

Abstract The sampling distribution of categorical data is determined by the observational procedure applied and not by assumptions with regard to the statistical model that characterizes the population. The sampling distribution is said to be multinomial if the number of observations is fixed in advance (binomial if there are only two categories), product multinomial if certain subsamples have pre-specified sizes, and Poisson if not the sample size rather the time period or geographic extent of sampling is specified in advance. This chapter gives a precise definition of these sampling procedures and discusses their most important characteristics. Marginalization and conditioning are the most frequent transformations that are applied to categorical data when relationships among variables are studied, and their implications for the sampling distributions are also described. The relationship between the multinomial and the Poisson distributions is investigated in detail. Statistical modeling mostly concentrates on data obtained through one of the above sampling procedures, but most surveys of the human population apply sampling procedures with unequal selection probabilities of individuals. These procedures are reviewed briefly.

The chapter starts with a discussion of the binomial distribution, which is extended to the multinomial distribution but is also used to introduce the Poisson distribution.

2.1 The Binomial Distribution

When the population is divided into two groups, like economically active people may be employed or unemployed or eligible voters may or may not intend to cast their votes, the population is characterized by the respective probabilities of these groups, p_i, $i = 1,2$, so that $p_1 + p_2 = 1$ and $0 < p_i < 1$. In this case, $\{p_1, p_2\}$ is a probability distribution. The possibilities that the probabilities could be 0 or 1 are excluded, to make sure that there are two non-empty categories in the population.

The case when there is only one category in the population may be very interesting from a substantive point of view (e.g., no unemployment) but is irrelevant from the perspective of statistical analysis. In the case of a population with finite size, say N, the probabilities may be interpreted as the relative sizes of the two groups: $p_i = N_i/N$, where N_i is the number of those in the i-th group, $N_1 + N_2 = N$. In the case of an infinite population, it is assumed that observation of individuals from the population is possible and p_i is the probability of observing an individual belonging to group i.

The binomial distribution is associated with observing a sample from the population with a pre-specified size, say, n. The selection of the individuals into the sample is always done independently from each other. The exact formulation of the sampling procedure depends on whether or not the population is considered to be finite, and in either case sampling may occur with our without replacement of the already selected individuals. A more precise description of sampling procedures defines which subsets constitute a possible sample and what is the selection probability of any such subsets. Properties of the selection of individuals into the sample are derived from this.

2.1.1 Sampling Without Replacement

When the population is finite, the simplest sampling procedure is the so-called simple random sampling. Simple random sampling means that any subset of size n of the population is a possible sample and each such sample has the same probability of being selected. This view of the sampling procedure is slightly different from the approach when the procedure is defined through the selection of individual observations into the sample one after the other. In that case, in each step of the sampling procedure, each individual who is not yet selected has the same chance of being selected into the sample. The latter procedure is often referred to as sampling without replacement.

Under simple random sampling, the number of samples containing k individuals from group 1 and $n-k$ individuals from group 2 is

$$\binom{N_1}{k}\binom{N_2}{n-k}$$

and each of these samples has probability

$$\binom{N}{n}^{-1}$$

of being selected. So if a random variable X counts the number of individuals from group 1 in the sample, then the sample space of X is $\{0,1,\ldots,n\}$ and

$$P(X=k)=\frac{\binom{N_1}{k}\binom{N_2}{n-k}}{\binom{N}{n}}, \tag{2.1}$$

which can be written as

$$\frac{N_1!N_2!n!(N-n)!}{k!(N_1-k)!(n-k)!(N_2-n+k)!N!} \tag{2.2}$$

In the case of sampling without replacement, any sample of size n containing exactly k individuals from the first group may contain these individuals in

$$\binom{n}{k} \tag{2.3}$$

different positions, and each such sample has the same probability. For example, choosing first k individuals from the first group and then selecting $n-k$ individuals from the second group occur with probability

$$\frac{(N_1)\cdots(N_1-k+1)(N_2)\cdots(N_2-n+k+1)}{(N)\cdots(N-n+1)}. \tag{2.4}$$

If Y counts the number of observations from the first group in a sample of size n generated through sampling without replacement, then, for $k=0,\ldots,n$

$$P(Y=k)=\frac{(N_1)\cdots(N_1-k+1)(N_2)\cdots(N_2-n+k+1)}{(N)\cdots(N-n+1)}\frac{(n)\cdots(n-k+1)}{(k)\cdots(1)}, \tag{2.5}$$

which is obtained by multiplying (2.3) with (2.4).

The formulas (2.2) and (2.5) are obviously the same, leading to

Proposition 2.1. *Simple random sampling and sampling without replacement yield the sampling distribution given in (2.1).* □

The sampling distribution given in (2.1) is called the hypergeometric distribution.

The binomial distribution is obtained as the result of a different but closely related sampling procedure: sampling with replacement.

2.1.2 Sampling with Replacement

For a finite population, sampling with replacement means that an individual already selected for the sample is replaced into the population before the next individual is selected; thus samples containing the same individual several times are not impossible. Therefore, instead of (2.4), one has

$$\frac{(N_1)\cdots(N_1)(N_2)\cdots(N_2)}{(N)\cdots(N)}. \tag{2.6}$$

and

$$P(Y=k)=\frac{(N_1)\cdots(N_1)(N_2)\cdots(N_2)}{(N)\cdots(N-n+1)}\frac{(n)\cdots(n-k+1)}{(k)\cdots(1)} \tag{2.7}$$

or

$$P(Y=k)=\binom{n}{k}p_1^k p_2^{n-k}. \tag{2.8}$$

In the case of an infinite population, this sampling procedure is referred to by saying that the observations are independent and identically distributed (i.i.d.). The distribution given in (2.8) is called the binomial distribution.

In the practice of surveying the human population, sampling with replacement is not practical, as it would eventually involve interviewing the same person repeatedly. Instead, sampling without replacement is used. However, for large populations (like all inhabitants of a country), the binomial distribution is a good approximation of the sampling distribution in (2.1) in the following sense. Rewrite (2.5) as

$$P(Y=k)=\frac{\frac{N_1}{N}\cdots\frac{N_1-k+1}{N}\frac{N_2}{N}\cdots\frac{N_2-n+k+1}{N}}{\frac{N}{N}\cdots\frac{N-n+1}{N}}\frac{(n)\cdots(n-k+1)}{(k)\cdots(1)}. \tag{2.9}$$

For $N\to\infty$, so that p_i remains equal to N_i/N, (2.9) converges to (2.8), thus

Proposition 2.2. *The binomial probability in (2.8) is the limit of the sampling probability for simple random sampling (2.5), if the population size N converges to infinity, so that $p_i = N_i/N$ and $i = 1,2$ remain constant.* □

A heuristic interpretation of the proposition is that for large population sizes, sampling with or without replacement tend to behave similarly. Note that this result holds for fixed n and k.

The summary of the results of the last two subsections is that simple random sampling and sampling without replacement are identical. Sampling with replacement leads to i.i.d. samples. The sampling distribution in the latter case is binomial. Further, binomial is the limiting sampling distribution[1] for the former sampling procedures, if the population size converges to infinity.

2.1.3 *Properties of the Binomial Distribution*

For each observation $i=1,\ldots,n$, define the indicator variable Y_i, which is 1 if the observation is from group 1 and zero otherwise. Then, under sampling with replacement, the Y_i variables are i.i.d. with a so-called Bernoulli distribution and

[1] See Sect. 3.1 for a formal discussion of convergence in distribution. In the current setting, the sample space remains constant, and the probabilities associated with observing any sample converge.

$$E(Y_i) = 1p_1 + 0p_2 = p_1,$$

$$V(Y_i) = (1-p_1)^2 p_1 + (0-p_1)^2 p_2 = p_1 - p_1^2 = p_1(1-p_1),$$

using that $p_2 = 1 - p_1$. Independence of the observations means that Y_i and Y_j are independent, if $i \neq j$.

Obviously, if

$$Y = \sum_{i=1}^{n} Y_i,$$

then Y has a binomial distribution and

$$E(Y) = \sum_{i=1}^{n} E(Y_i) = \sum_{i=1}^{n} p_1 = np_1,$$

$$V(Y) = \sum_{i=1}^{n} V(Y_i) = \sum_{i=1}^{n} p_1(1-p_1) = np_1(1-p_1).$$

The usual notation for Y having the binomial distribution is $Y \sim \mathscr{B}(n, p_1)$. The sample space of Y is $\{0, \ldots, n\}$, and the probability of $Y = k$, for $k = 0, \ldots, n$, was given in (2.8).

The asymptotic approximation of a binomial distribution with a normal distribution will be discussed in the next chapter.

2.2 The Multinomial Distribution

The binomial distribution is related to the observation of a categorical variable with two categories: someone belongs to one group or to the other. Such a variable is called binary or dichotomous. A categorical variable with several categories is called a polytomous variable. The multinomial (also called polynomial) distribution is related to observing such a variable. Suppose the population of interest is divided into c mutually exclusive and exhaustive groups. The division of the population is characterized by a vector[2] $\mathbf{p}$, with $p_i > 0$ and $\sum_i p_i = 1$. In the case of a finite population, the value of p_i is the fraction of the population that is in group i, and when sampling is made with equal selection probabilities, it is the probability of selecting an individual from group i. For an infinite population, p_i is the probability that an individual randomly selected from the population belongs to group i. For the polynomial distribution, only sampling with replacement, i.e., the i.i.d. case, is considered in the text; for sampling without replacement, see the second task in the Things to Do section.

When the sample size is fixed at n observations and a vector-valued variable $\mathbf{Y}$ counts the number of observations in each category $i = 1, \ldots, c$, then the sample space of $\mathbf{Y}$ consists of vectors $\mathbf{y}$, such that

[2] Vectors in this book are column vectors.

$$\sum_{i=1}^{c} y_i = n,\; y_i \geq 0,\; i = 1, \ldots c.$$

Let the collection of such vectors $\mathbf{y}$ be denoted by $S_{c,n}$.

Theorem 2.1. *For all $\mathbf{y} \in S_{c,n}$,*

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \frac{n!}{\prod_{i=1}^{c} y_i!} \prod_{i=1}^{c} p_i^{y_i}. \tag{2.10}$$

Further,

$$E(\boldsymbol{Y}) = n\boldsymbol{p}. \tag{2.11}$$

$$\mathrm{Cov}(\boldsymbol{Y}) = \mathrm{n}\boldsymbol{D_p} - \mathrm{n}\boldsymbol{pp'}, \tag{2.12}$$

where $\mathrm{Cov}(\boldsymbol{Y})$ *is the covariance matrix of* $\boldsymbol{Y}$ *and* $\mathbf{D_p}$ *is a diagonal matrix with the probabilities of* $\mathbf{p}$ *in the diagonal positions.*

Proof. Although a proof similar to that of given for (2.8) is possible for (2.10), the following induction argument on c gives more insight.

For $c = 2$, (2.8) and (2.10) coincide; thus the latter formula is true. Assume now that for some $c-1$, (2.10) is proved, and it is shown now that (2.10) also holds for c. Let the parameters n and a c-dimensional $\mathbf{p}$ be given. Let $\mathbf{q}$ be a $c-1$ dimensional vector, so that $q_i = p_i$, for $i \leq c-2$ and $q_{c-1} = p_{c-1} + p_c$. For any $\mathbf{y} \in S_{c,n}$, let $\mathbf{z}$ be a $c-1$-dimensional vector, so that $z_i = y_i$, for $i \leq c-2$ and $z_{c-1} = y_{c-1} + y_c$. Denote this vector as $\mathbf{z}(\mathbf{y})$, and define a random variable $\mathbf{Z}$, with possible values $\mathbf{z}$, such that

$$P(\mathbf{Z} = \mathbf{z}) = \sum_{\mathbf{y}:\mathbf{z}(\mathbf{y})=\mathbf{z}} P(\mathbf{Y} = \mathbf{y}).$$

Then, using the induction assumption,

$$P(\mathbf{Y} = \mathbf{y}) = P(\mathbf{Z} = \mathbf{z})P(Y_{c-1} = y_{c-1}, Y_c = y_c | \mathbf{Z} = \mathbf{z}) =$$

$$P(\mathbf{Z} = \mathbf{z})P(Y_{c-1} = y_{c-1}, Y_c = y_c | y_{c-1} + y_c = z_{c-1}) =$$

$$\left(\frac{n!}{\prod_{i=1}^{c-1} z_i!} \prod_{i=1}^{c-1} q_i^{z_i} \right) \frac{z_{c-1}!}{y_{c-1}! y_c!} \left(\frac{p_{c-1}}{p_{c-1} + p_c} \right)^{y_{c-1}} \left(\frac{p_c}{p_{c-1} + p_c} \right)^{y_c} =$$

$$\left(\frac{n!}{\prod_{i=1}^{c-1} z_i!} \prod_{i=1}^{c-1} q_i^{z_i} \right) \frac{z_{c-1}!}{y_{c-1}! y_c!} \frac{p_{c-1}^{y_{c-1}} p_c^{y_c}}{q_{c-1}^{z_{c-1}}} =$$

$$\frac{n!}{\prod_{i=1}^{c} y_i!} \prod_{i=1}^{c} p_i^{y_i}.$$

To see (2.11), note that for $i = 1, \ldots, c$, $Y_i \sim \mathscr{B}(n, p_i)$, so $E(Y_i) = np_i$. This also implies that $V(Y_i) = np_i(1-p_i)$. To prove (2.12), one also needs to see that $Cov(Y_i, Y_j) = -np_i p_j$, for $i \neq j$.

For $k = 1, \ldots, n$, define the indicator variables Z_{kl} as $Z_{kl} = 1$ if the k-th observation is in category l and zero otherwise. Then

$$Y_l = \sum_{k=1}^{n} Z_{kl}$$

and

$$Cov(Y_i, Y_j) = E((Y_i - np_i)(Y_j - np_j)) = E\left(\left(\sum_{k=1}^{n} Z_{ki} - np_i\right)\left(\sum_{k=1}^{n} Z_{kj} - np_j\right)\right) =$$

$$E\left(\left(\sum_{k=1}^{n} Z_{ki}\right)\left(\sum_{k=1}^{n} Z_{kj}\right)\right) - n^2 p_i p_j. \tag{2.13}$$

To evaluate the first term in (2.13), note that for any k, $Z_{ki}Z_{kj} = 0$, because at most, one of them may be nonzero, and so is its expectation. For different k values, say k_1 and k_2, $Z_{k_1 i}$ and $Z_{k_2 j}$ are independent, so the expected value of their product is the product of their expected values. The product of the sums in the last expression of (2.13) has n^2 terms, n of them are zero, and the remaining $n(n-1)$ are all equal to $p_i p_j$. Thus

$$Cov(Y_i, Y_j) = n(n-1)p_i p_j - n^2 p_i p_j = -np_i p_j.$$

□

The variable $\mathbf{Y}$ is said to have a multinomial distribution with parameters n and $\mathbf{p}$ if its distribution is as given in (2.10), and this is denoted as

$$\mathbf{Y} \sim \mathscr{M}(n, \mathbf{p}).$$

The fact that the covariance between two components of a multinomial distribution is negative may be given the following intuitive explanation. When the value of one component is higher than expected, fewer observations, than expected, remain for the other components, so they will tend to be less than their respective expected values. This effect is stronger if the variables have large probabilities. For example, with $n = 100$ and $\mathbf{p}' = (0.5, 0.4, 0.1)$, if $Y_1 = 65$, then Y_3 may be equal to or more than its expectation, but Y_2 will be certainly less than its expectation. This observation will be made more precise in Sect. 2.4, where conditional distributions are discussed.

Facts mentioned in the proof of Theorem (2.1) also imply that

$$Cor(Y_i, Y_j) = \frac{-np_i p_j}{\sqrt{np_i(1-p_i)np_j(1-p_j)}} = -\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}}. \tag{2.14}$$

2.2.1 Stratified Sampling: The Product Multinomial Distribution

Although simple random samples based on moderate sample sizes usually provide fairly accurate estimates for the fraction p of those with a given characteristic in the population (see Chap. 4), samples with controlled composition are used frequently. To illustrate the need for such samples, suppose that one wishes to estimate from the sample the unemployment rate in the population. It may very well be the case that the unemployment rates among men and among women are different. In such a case, one expects a better estimate of the unemployment rate from samples with a gender distribution identical to that of the population, which may be known quite precisely from official records. It will be shown in Sect. 4.2.2 that this is really the case. In fact, many social surveys use samples that precisely represent the composition of the population according to gender, age, educational level, and some other aspects. Another relevant situation is when the goal of the study is to compare a minority group to the rest of the population. In such cases, a design when, say, half of the sample is selected from members of the minority of interest, and the other half from the rest of the population, leads to better estimates of the characteristics of the minority and a better comparison, than a simple random sample that is likely to contain only very few observations from this minority.

Random samples with controlled composition are called stratified samples. The population is divided into groups, called strata in this context, according to one or several characteristics. It is determined, before the sample is selected, how many observations are to be taken from each stratum. These numbers are determined either to precisely represent the population distribution in these categories or to have some other pre-specified composition. Then a simple random sample is selected from each stratum, and these are combined to form a sample from the entire population. Not only the selection of the individuals from within the same stratum that is done independently of each other but also the selection from the different strata. So if there are s strata, and the determined sample size in stratum i is $n(i)$, further the observations are grouped into c groups, and the distribution of these c groups in stratum i is described by the probabilities $\mathbf{p}(i)$, and then the distribution of the observations $\mathbf{Y}(i)$ in stratum i is $\mathbf{Y}(i) \sim \mathscr{M}(n(i), \mathbf{p}(i))$. When the variables $\mathbf{Y}(i)$ are considered together for all $i = 1, \ldots, s$, then the sample space is

$$S = \times_{i=1}^{s} S_{n(i),c}$$

and

Proposition 2.3.

$$P((\boldsymbol{Y}(1), \ldots, \boldsymbol{Y}(s)) = (\boldsymbol{y}(1), \ldots, \boldsymbol{y}(s))) =$$

$$\prod_{i=1}^{s} \left(\frac{n(i)!}{\prod_{j=1}^{c} y(i)_j!} \prod_{j=1}^{c} p(i)_j^{y(i)_j} \right), \; (\boldsymbol{y}(1), \ldots, \boldsymbol{y}(s)) \in S,$$

further

$$E((\boldsymbol{Y}(1),\ldots,\boldsymbol{Y}(s))) = (n_1\boldsymbol{p}(1),\ldots n_s\boldsymbol{p}(s)),$$

and the covariance matrix of the random vector $(\boldsymbol{Y}(1)',\ldots,\boldsymbol{Y}(s)')'$ *is block diagonal with the diagonal blocks being the covariance matrices of* $\boldsymbol{Y}(i)$ *given in (2.12).* □

The distribution described in Proposition 2.3 is called product multinomial distribution.

The expectation vector of a product multinomial distribution consists of the expectation vectors of the strata. Its covariance matrix, because of the independence of the observations in different strata, is block diagonal, with the blocks being the covariance matrices of the multinomial distributions in the strata.

2.3 Contingency Tables

Sampling distributions cannot be properly studied without studying the sample space, of which the observations are elements. So far, it has been assumed that the population is divided into c categories and each observation belongs to one and only one of these. This is an appropriate model of the observation of a categorical variable. But in most real problems, the sample space is more structured than this. In fact, usually several categorical variables are observed and analyzed at the same time. For example, in a study concerning the health status of the population, in addition to a classification whether or not an individual is healthy, also classifications regarding past diseases, whether medication was taken in the past, and socioeconomic status and other characteristics may be relevant. Each of these characteristics defines a categorical variable, and the joint analysis of these variables is called multivariate analysis. Each individual in the population belongs to one and only one category of each variable. An individual may be healthy now but may have had a certain disease in the past, but may have taken no medication for it, and may be characterized by being an upper middle class professional. Classification according to all these aspects is called cross-classification.

More formally, let the categorical variables $X_1,\ldots,X_m$ be of interest, with X_i having categories $1,\ldots,c_i$. The categories may be denoted by natural numbers without restriction of generality, but one should keep in mind that the categories do not have the properties which numbers do. When the individuals in the population are cross-classified according to the m variables, the possible number of different observations is $c = \prod_{i=1}^{m} c_i$. The sample space of observing variable X_i is $\{1,\ldots,c_i\}$, and as in theory every category of any variable may occur together with any category of any other variable, the joint sample space of the variables is the Cartesian product

$$\Omega = \times_{i=1}^{m}\{1,\ldots,c_i\}.$$

There are situations, when not all combinations of the categories of the individual variables may occur. Such situations will be briefly discussed in Chap. 13.

The joint sample space Ω is called a contingency table. The contingency table has dimension m, and it is also called an m-way table. Each combination $(j_1,\ldots,j_m)$, with $j_i \in \{1,\ldots,c_i\}$, is called a cell of the contingency table. Note that when the categories of a variable are given, the order of these categories is arbitrary, and also when a joint category is considered, the order of the variables is arbitrary.

For example, in the case of $m = 2$ and $c_1 = 2$, $c_2 = 4$, Ω has the structure shown in Table 2.1. The cells of the table may contain, in the case of a finite population, the number of individuals in each combination (frequency distribution) or the fraction of the population in that specific category combination. In the latter case, the entries in the table may be interpreted as probabilities and constitute a probability distribution. Frequency or probability distributions may be derived from a sample and may also represent estimates based on a model and data.

Table 2.1 The structure of a 2 × 4 contingency table

(1,1)	(1,2)	(1,3)	(1,4)
(2,1)	(2,2)	(2,3)	(2,4)

A two-way table is said to contain rows (those cells with the same index of the first variable) and columns (those cells with the same index of the second variable). When the table contains frequencies or probabilities, the distribution of the first variable is obtained by summing over the second index and may be conveniently written into an additional column on the marginal. Similarly, the distribution of the second variable is obtained by summing over the first index and may be written into a marginal row, as illustrated in Table 2.2 for frequencies. A $+$ sign instead of an index means summation for all values of that index. The frequency $f(+,+)$ is obtained by summing all entries in the table and is the total population or sample size. Table 2.2 also shows the categories of the two variables, say A and B,

Table 2.2 A 2 × 4 frequency table with marginal frequencies added

	B = 1	B = 2	B = 3	B = 4	
A = 1	f(1,1)	f(1,2)	f(1,3)	f(1,4)	f(1,+)
A = 2	f(2,1)	f(2,2)	f(2,3)	f(2,4)	f(2,+)
	f(+,1)	f(+,2)	f(+,3)	f(+,4)	f(+,+)

When $m = 3$, the contingency table is a three-dimensional object with rows, columns, and layers. If the table contains the joint distribution of variables A, B, and C, it is best represented as cross-classifications according to variables A and B of those who are in a particular category of C. Such cross-classifications are conditional distributions, which will be discussed in the next section. For example, when $c_1 = 2$, $c_2 = 4$, $c_3 = 3$, the structure of the frequency distribution is illustrated in Table 2.3.

Note that when a contingency table like the one in Table 2.3 contains probabilities, the values $\{p(i,j,k), i = 1,\ldots,c_1, j = 1,\ldots c_2\}$ are not the probabilities of the

Table 2.3 The structure of a $2 \times 4 \times 3$ frequency distribution

	C = 1			
	B = 1	B = 2	B = 3	B = 4
A = 1	f(1,1,1)	f(1,2,1)	f(1,3,1)	f(1,4,1)
A = 2	f(2,1,1)	f(2,2,1)	f(2,3,1)	f(2,4,1)

	C = 2			
	B = 1	B = 2	B = 3	B = 4
A = 1	f(1,1,2)	f(1,2,2)	f(1,3,2)	f(1,4,2)
A = 2	f(2,1,2)	f(2,2,2)	f(2,3,2)	f(2,4,2)

	C = 3			
	B = 1	B = 2	B = 3	B = 4
A = 1	f(1,1,3)	f(1,2,3)	f(1,3,3)	f(1,4,3)
A = 2	f(2,1,3)	f(2,2,3)	f(2,3,3)	f(2,4,3)

conditional distribution of A and B given $C = k$. The latter conditional probabilities are of the form $p(i, j, k)/p(+,+,k)$.

Four-dimensional tables may be defined as consisting of three-dimensional conditional tables, and higher-dimensional tables may be defined through a similar recursion. On the other hand, when a higher-dimensional table is given, lower-dimensional tables may be generated by conditioning and/or marginalization.

Conditioning on specific categories of certain variables involves the selection of those in the given categories and cross-classifying them according to the remaining variables. Marginalization over some of the variables means disregarding these variables and cross-classifying only according to the remaining variables. For example, with the structure shown in Table 2.3, the conditional distribution of A and C, given $B = 3$, is a 2×3 table, and the $A \times C$ marginal table also has this structure. In the case of frequencies, the former table contains only those in $B = 3$, while the latter does contain all observations. In the case of probabilities, the conditional distribution is obtained if the probabilities are divided by $p(+,3,+)$, while in the latter case, the marginal probabilities sum to 1 and constitute a probability distribution.

2.4 Conditional and Marginal Distributions

When the population is structured as a contingency table Ω and the observations are independent and identically distributed, the frequencies resulting from the observation of n individuals may be considered as having a multinomial distribution with c categories. In this section, the marginal and conditional distributions derived from this multinomial distribution are studied. For this, slightly more general concepts of marginalization and conditioning will be used, than those discussed in the previous section.

When the table has c categories, and $\mathbf{Y}$ is a c-dimensional vector in lexicographic order of the cells, let $\mathbf{A}$ be a $k \times c$ matrix of 0's and 1's, such that each column contains exactly one 1. $\mathbf{A}$ may be selected in such a way that $\mathbf{AY}$ gives a marginal distribution of $\mathbf{Y}$. For example, when $\boldsymbol{\Omega}$ is 2×4, as in Table 2.2, $\mathbf{Y}$ contains the frequencies in the order

$$(1,1),\,(1,2),\,(1,3),\,(1,4),\,(2,1),\,(2,2),\,(2,3),\,(2,4)$$

and the following matrix generates the row marginals:

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

while the column marginals are generated by

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Obviously, such matrices define a partition of the cells of the table by their rows. In certain applications, more general partitions are needed. For example, a social mobility table cross-classifies men according to their own and their fathers' social status. Those who have a higher status than their fathers are upward mobile, those in a lower status than their fathers are downward mobile, and those in the same position as their fathers are immobile. A 3×3 social mobility table has the structure shown in Table 2.4. The upward mobile sons are in the lower triangular positions, the immobile sons are in the main diagonal positions, and the downward mobile sons are in the upper triangular positions. To get the number of the upward mobile, immobile, and downward mobile sons, respectively, the following matrix may be used:

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

Table 2.4 A 3×3 social mobility table

		Son		
		High	Medium	Low
Father	High	f(1,1)	f(1,2)	f(1,3)
	Medium	f(2,1)	f(2,3)	f(2,4)
	Low	f(3,1)	f(3,2)	f(3,3)

Theorem 2.2. *If* $Y \sim \mathscr{M}(n,\boldsymbol{p})$ *and* $\boldsymbol{A}$ *is a 0-1 matrix with exactly one 1 in every column of it, then*

$$\boldsymbol{A}Y \sim \mathscr{M}(n,\boldsymbol{A}\boldsymbol{p}).$$

Proof. If $\mathbf{A}$ is $k \times c$, then both $\mathbf{AY}$ and $\mathbf{Ap}$ are k-dimensional and the sample space of $\mathbf{X}$ is $S_{n,k}$. Let the categories of $\mathbf{Y}$ be $\mathbf{c} = (c_1,\ldots,c_c)'$ and the categories of $\mathbf{X}$ $\mathbf{d} = (d_1,\ldots,d_k)'$. The transformation matrix $\mathbf{A}$ defines a surjective mapping between $\mathbf{c}$ and $\mathbf{d}$. Now select a sample of size n from the population, and consider two observational procedures. One classifies the observations into the categories of $\mathbf{c}$, to obtain $\mathbf{y}$, and then considers $\mathbf{Ay}$. The other one uses the same observations and classifies them into $\mathbf{d}$ and it obtains, say, $\mathbf{x}$. For every sample, $\mathbf{Ay} = \mathbf{x}$; therefore $\mathbf{AY} = \mathbf{X}$ holds for all samples, so $\mathbf{AY}$ and $\mathbf{X}$ have the same distribution. □

Although various conditioning operations are possible, for the applications in this book, it is sufficient to consider the case, when $\mathbf{Y}$ is conditioned upon $\mathbf{AY} = \mathbf{Ay}$. This includes conditional distributions conditioned on fixed lower-order marginal distributions but also on more general partition sums, including certain cell frequencies being fixed. Before giving the general result, an interesting special case will be considered.

For $c = 4$, let $\mathbf{Y} \sim \mathscr{M}(n,\mathbf{p})$. Choose $\mathbf{A}$ so that $\mathbf{AY}$ is the sum of the first two and of the last two components. Then, the conditioning event is $\mathbf{AY} = \mathbf{Az}$ for a fixed $\mathbf{z} \in S_{n,c}$.[3] The sample space of the conditional distribution of $\mathbf{Y}|(\mathbf{AY} = \mathbf{Az})$ is the subset, say, $S_{n,c}(\mathbf{Az})$, such that for all $\mathbf{y} \in S_{n,c}(\mathbf{Az})$, $\mathbf{Ay} = \mathbf{Az}$. Then, for $\mathbf{y} \in S_{n,c}(\mathbf{Az})$,

$$P(\mathbf{Y} = \mathbf{y}|\mathbf{AY} = \mathbf{Az}) = \frac{P(\mathbf{Y} = \mathbf{y},\ \mathbf{AY} = \mathbf{Az})}{P(\mathbf{AY} = \mathbf{Az})} =$$

$$\frac{P(\mathbf{Y} = \mathbf{y})}{P(\mathbf{AY} = \mathbf{Az})} =$$

$$\frac{\frac{n!}{y_1!(z_1+z_2-y_1)!y_3!(z_3+z_4-y_3)!} p_1^{y_1} p_2^{z_1+z_2-y_1} p_3^{y_3} p_4^{z_3+z_4-y_3}}{\frac{n!}{(z_1+z_2)!(z_3+z_4)!}(p_1+p_2)^{z_1+z_2}(p_3+p_4)^{z_3+z_4}},$$

using Theorem 2.2, and then by reordering the terms, one obtains that this is equal to

$$\frac{(z_1+z_2)!}{y_1!(z_1+z_2-y_1)!}\left(\frac{p_1}{p_1+p_2}\right)^{y_1}\left(\frac{p_2}{p_1+p_2}\right)^{z_1+z_2-y_1} \times$$

$$\frac{(z_3+z_4)!}{y_3!(z_3+z_4-y_3)!}\left(\frac{p_3}{p_3+p_4}\right)^{y_3}\left(\frac{p_4}{p_3+p_4}\right)^{z_3+z_4-y_3} =$$

$$P(\mathbf{Y}_1 = \mathbf{y}_1, \mathbf{Y}_2 = \mathbf{y}_2|\mathbf{y}_1 + \mathbf{y}_2 = \mathbf{z}_1 + \mathbf{z}_2)P(\mathbf{Y}_3 = \mathbf{y}_3, \mathbf{Y}_4 = \mathbf{y}_4|\mathbf{y}_3 + \mathbf{y}_4 = \mathbf{z}_3 + \mathbf{z}_4),$$

that is, the subsets, the sums of which are fixed in the condition, become independent of each other, and one obtains a product multinomial joint distribution. The

[3] It would be correct to write the conditioning event as $\mathbf{AY} = \mathbf{Ay}$, but the argument is easier to read if $\mathbf{z}$ is used to denote a specific value of $\mathbf{y}$.

general result is essentially the same, except that the subsets with fixed sums may be of different sizes. To illustrate the usefulness of this generalization, suppose that the social mobility table considered is not 3×3, as it was in the example prior to Theorem 2.2, rather, say, 7 categories are used to describe social status. The immobile group contains 7, and both the upward mobile and the downward mobile groups contain 21 cells. The next result, when applied to this example, provides the distribution of the cell frequencies, given the total numbers of those immobile and upward or downward mobile.

Theorem 2.3. *If $\boldsymbol{Y} \sim \mathscr{M}(n,\boldsymbol{p})$ is c-dimensional and $\mathbf{A}$ is a $k \times c$ matrix of* 0*'s and* 1*'s, with exactly one 1 in every column, then for every $\boldsymbol{z} \in S_{n,c}$ and $\boldsymbol{y} \in S_{n,c}(\boldsymbol{Az})$,*

$$P(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{AY}=\boldsymbol{Az}) = \prod_{j=1}^{k}\left(\frac{(\boldsymbol{Az})_i!}{\prod_{i:\boldsymbol{A}_{ji}=1} y_i!}\prod_{i:\boldsymbol{A}_{ji}=1}\left(\frac{p_i}{(\boldsymbol{Ap})_j}\right)^{\boldsymbol{y}_i}\right). \tag{2.15}$$

Proof. The proof goes exactly like the argument in the introductory example. For $\mathbf{y} \in S_{n,c}(\mathbf{Az})$,

$$P(\mathbf{Y}=\mathbf{y}|\mathbf{AY}=\mathbf{Az}) = \frac{P(\mathbf{Y}=\mathbf{y},\mathbf{AY}=\mathbf{Az})}{P(\mathbf{AY}=\mathbf{Az})} =$$

$$\frac{P(\mathbf{Y}=\mathbf{y})}{P(\mathbf{AY}=\mathbf{Az})} = \frac{\frac{n!}{\prod_{i=1}^{c} y_i!}\prod_{i=1}^{c} p_i^{y_i}}{\frac{n!}{\prod_{j=1}^{k}(\mathbf{Az})_j!}\prod_{j=1}^{k}(\mathbf{Ap})_j^{(\mathbf{Az})_j}} =$$

$$\frac{\frac{n!}{\prod_{j=1}^{k}\prod_{i:\mathbf{A}_{ji}=1} y_i!}\prod_{j=1}^{k}\prod_{i:\mathbf{A}_{ji}=1} p_i^{y_i}}{\frac{n!}{\prod_{j=1}^{k}(\mathbf{Az})_j!}\prod_{j=1}^{k}(\mathbf{Ap})_j^{(\mathbf{Az})_j}} = \prod_{j=1}^{k}\left(\frac{(\mathbf{Az})_j!}{\prod_{i:\mathbf{A}_{ji}=1} y_i!}\prod_{i:\mathbf{A}_{ji}=1}\left(\frac{p_i}{(\mathbf{Ap})_j}\right)^{y_i}\right).$$

□

The distribution described in Theorem 2.3 is a slightly more general product multinomial distribution than the one described in Proposition 2.3. Its expectation vector may be assembled from the expectations of the multinomial distributions on the right-hand side of (2.15), and its covariance matrix is block diagonal because of the independence in (2.15), and the blocks are the covariance matrices of the multinomial distributions on the right-hand side.

2.5 The Poisson Distribution

The sampling procedures discussed so far assumed that a sample of a fixed size could be selected from the population. When the population is not of a finite size, this meant taking observations, as long as the desired sample size was obtained. In many applications, the temporal nature of the observations is relevant: the sampling units to be observed may occur one after the other or the observation or measurement procedure may take time and may be done in a certain order of the units. For example, patients arrive in a hospital in a given order, or even if some happen to arrive at the same time, their admission will happen in a certain order. Or in an exit poll after election, interviewing of the voters as they leave the polling station happens in a given order. Of course, there may be several nurses working on hospital admission or pollsters collecting data in parallel, but even in those cases, order may be determined. Often in such data collection procedures, the total sample size is not determined in advance, rather observations are taken for a given period of time. For example, patients seeking hospital admission with a given medical condition are observed for, say, a week.

In such cases, the sampling procedure is best characterized by the intensity with which observations of interest occur. The intensity is a nonnegative number and refers to a certain time period. It is the expected number of observations (of interest) during that period. If the intensity during period T is, say, λ, and the random variable X counts the occurrences for that period, then $E(X) = \lambda$. One assumption regarding the distribution of X is that if T is divided into very small disjoint time intervals $T_1, \ldots, T_n$ of equal lengths, then the probability of an observation occurring during the period T_i is p_i. Under a constant intensity assumption, p_i does not depend on i. If for fixed T n is large enough so that $p_i = p$ is so small that the probability of having two or more observations during T_i is very small, then the number of observations during T_i is 0 or 1, and the expected number of observation during T_i is approximately p. Thus, the expected number of total observations during T is about np, so $p = \lambda/n$. In this case, the total number of observations, X, has, approximately, a binomial distribution $\mathscr{B}(n, \lambda/n)$, so

$$P(X = k) \approx \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(\frac{n-\lambda}{n}\right)^{n-k}.$$

The above formula is only an approximation, when the possibility of having more than 1 observation in each period T_i is neglected. The shorter the periods T_i are, that is, the larger is n, the more realistic is this assumption. This suggests considering the following limit

$$\lim_{n\to\infty} \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(\frac{n-\lambda}{n}\right)^{n-k} =$$

$$\lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n^k}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k} =$$

$$\lim_{n\to\infty} \left(1(1-\frac{1}{n})\cdots(1-\frac{k-1}{n})\right)\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k}.$$

In the above expression, the first and the fourth terms converge to 1, and the third term converges to $\exp(-\lambda)$, so one obtains that

$$\lim_{n\to\infty} \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(\frac{n-\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!}e^{-\lambda}.$$

Based on the above heuristic argument, the Poisson distribution is defined as

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}, \tag{2.16}$$

for every nonnegative integer k. The notation for the random variable X having a Poisson distribution is $X \sim \mathscr{P}(\lambda)$.

The argument presented above that leads to the definition of the Poisson distribution is often referred to by saying that the Poisson distribution is the limit of the binomial distribution, when n goes to infinity so that np remains constant. This should not be interpreted as if a given experiment or sampling procedure would be repeated with increasing sample size, because p does not remain constant as n increases. Rather, the correct interpretation is to think of a series of experiments, each characterized by its own n_i and p_i parameters, such that $n_i \to \infty$ and $n_i p_i$ remains constant.

The convergence argument is also useful in proving properties of the Poisson distribution, but it needs to be made more precise. Let $X \sim \mathscr{P}(\lambda)$. A binomially distributed random variable $Y_n \sim \mathscr{B}(n, \lambda/n)$ has $\{0,1,\ldots,n\}$ as its sample space. Define a variable $\tilde{Y}_n$ with a sample space of all nonnegative integers and

$$P(\tilde{Y}_n=k)=P(Y_n=k),\ k=0,1,\ldots,n,$$

$$P(\tilde{Y}_n=k)=0,\ k=n+1,n+2,\ldots.$$

Then the argument above implies that

Proposition 2.4.

$$\lim_{n\to\infty} P(\tilde{Y}_n=k)=P(X=k)$$

for all nonnegative integers k. □

Theorem 2.4. *The Poisson distribution is a probability distribution, i.e.,*

$$\sum_{k=1}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda}=1,$$

for all $\lambda > 0$. *If* $X \sim \mathscr{P}(\lambda)$, *then* $E(X)=\lambda$ *and* $Var(X)=\lambda$.

Proof. Proposition 2.4 implies that

$$\sum_{k=1}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda}=\sum_{k=1}^{\infty}\lim_{n\to\infty}P(\tilde{Y}_n=k)=$$

$$\lim_{n\to\infty}\sum_{k=1}^{\infty}P(\tilde{Y}_n=k)=\lim_{n\to\infty}\sum_{k=1}^{\infty}P(Y_n=k)=$$

$$\lim_{n\to\infty}1=1$$

where the infinite sum and the limit can be interchanged because of the positivity of all terms and the existence of the limit.

Further, $E(Y_n) = E(\tilde{Y}_n)$ and $Var(Y_n) = Var(\tilde{Y}_n)$. The expectation and the variance are continuous functions of the probabilities associated with the values of a random variable, so Proposition 2.4 implies that

$$\lim_{n\to\infty} E(\tilde{Y}_n) = \lim_{n\to\infty} \lambda = \lambda = E(X)$$

and

$$\lim_{n\to\infty} Var(\tilde{Y}_n) = \lim_{n\to\infty} n\frac{\lambda}{n}\left(\frac{n-\lambda}{n}\right) = \lim_{n\to\infty} \lambda\left(\frac{1-\lambda/n}{1}\right) = \lambda = Var(X).$$

□

Multivariate Poisson distributions may be easily defined by assuming that the components are independent from each other. The counts may represent observations falling into different categories or different cells of a contingency table. So if there are c categories, and each is characterized by an intensity λ_i, and X_i counts the number of observations in each category, then

$$P(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{c} P(X_i = x_i) = \prod_{i=1}^{c} \frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i},$$

for all nonnegative integer vectors $\mathbf{x}$. Then X has multivariate Poisson distribution with parameter $\boldsymbol{\lambda}$, denoted as $\mathbf{X} \sim \mathscr{P}(\boldsymbol{\lambda})$.

Marginalization of a multivariate Poisson distribution leads to a Poisson distribution:

Theorem 2.5. *Suppose that $\mathbf{X} \sim \mathscr{P}(\boldsymbol{\lambda})$ and $\mathbf{A}$ is a $k \times c$ $0-1$ matrix with exactly one 1 in every column. Then $\mathbf{AX} \sim \mathscr{P}(\mathbf{A}\boldsymbol{\lambda})$.*

Proof. As the independence part of the claim is straightforward, it is sufficient to show that if $X_i \sim \mathscr{P}(\lambda_i)$ and X_1 and X_2 are two independent one-dimensional Poisson variables, then $X_1 + X_2 \sim \mathscr{P}(\lambda_1 + \lambda_2)$. Indeed,

$$P(X_1 + X_2 = k) = \sum_{l=0}^{k} P(X_1 = l, X_2 = k-l) = \sum_{l=0}^{k} P(X_1 = l)P(X_2 = k-l) =$$

$$\sum_{l=0}^{k} \frac{\lambda_1^l}{l!} e^{-\lambda_1} \frac{\lambda_2^{k-l}}{(k-l)!} e^{-\lambda_2} = e^{-(\lambda_1+\lambda_2)} \frac{1}{k!} \sum_{l=0}^{k} \frac{k!}{l!(k-l)!} \lambda_1^l \lambda_2^{k-l} =$$

$$\frac{(\lambda_1 + \lambda_2)^k \cdot}{k!} e^{-(\lambda_1+\lambda_2)}.$$

□

The following proposition is a simple consequence of Theorem 2.5. As its proof shows, a simple induction proof can be given.

Proposition 2.5. *Let $X_i \sim \mathscr{P}(\lambda_i)$ be pairwise independent for $i = 1, \ldots, k$. Then*

$$\sum_{i=1}^{k} X_i \sim \mathscr{P}(\sum_{i=1}^{k} \lambda_i).$$

□

More surprising is the behavior of the multivariate Poisson distribution upon conditioning. When one conditions on the total, the joint distribution becomes multinomial.

Theorem 2.6. *If $\boldsymbol{X} \sim \mathscr{P}(\boldsymbol{\lambda})$, then*

$$\boldsymbol{X} | \left(\sum_{i=1}^{c} X_i = k \right) \sim \mathscr{M}(k, \frac{1}{\sum_{i=1}^{c} \lambda_i} \boldsymbol{\lambda}).$$

Proof. For $\mathbf{x} \in S_{c,n}$,

$$P\left(\mathbf{X} = \mathbf{x} | \left(\sum_{i=1}^{c} X_i = k\right)\right) = \frac{P(\mathbf{X} = \mathbf{x}, (\sum_{i=1}^{c} X_i = k))}{P((\sum_{i=1}^{c} X_i = k)))} =$$

$$\frac{P(\mathbf{X} = \mathbf{x})}{P((\sum_{i=1}^{c} X_i = k)))} = \frac{\prod_{i=1}^{c} \frac{\lambda_i^{x_i}}{x_i!} e^{-\lambda_i}}{\frac{(\sum_{i=1}^{c} \lambda_i)^k}{k!} e^{-\sum_{i=1}^{c} \lambda_i}} = \frac{k!}{\prod_{i=1}^{c} x_i!} \prod_{i=1}^{c} \left(\frac{\lambda_i}{\sum_{j=1}^{c} \lambda_j} \right)^{x_i}.$$

□

An interesting implication of Theorem 2.6 is that conditioning on the total introduces negative correlations among the coordinates which without conditioning are independent. It is going to be a very important topic in this book that joint distributions may change their properties fundamentally upon conditioning.

The next result is a converse of Theorem 2.6. If the total has a Poisson distribution and the joint conditional distribution is multinomial, then the joint distribution is multivariate Poisson.

Theorem 2.7. *Suppose that $\boldsymbol{X}$ is a c-dimensional vector-valued variable on the nonnegative integers, such that*

$$\sum_{i=1}^{c} X_i \sim \mathscr{P}(\lambda)$$

and

$$\boldsymbol{X} | \left(\sum_{i=1}^{c} X_i = k \right) \sim \mathscr{M}(k, \boldsymbol{p}).$$

Then

$$\boldsymbol{X} \sim \mathscr{P}(\boldsymbol{\lambda}), \text{ with } \lambda_i = \lambda p_i.$$

Proof. For any c-dimensional nonnegative integer vector $\mathbf{x}$,

$$P(\mathbf{X}=\mathbf{x}) = P(\mathbf{X}=\mathbf{x}|\sum_{i=1}^{c} X_i = \sum_{i=1}^{c} x_i)\; P(\sum_{i=1}^{c} X_i = \sum_{i=1}^{c} x_i) =$$

$$\frac{(\sum_{i=1}^{c} x_i)!}{\prod_{i=1}^{c} x_i!} \prod_{i=1}^{c} p_i^{x_i} \; \frac{\lambda^{\sum_{i=1}^{c} x_i}}{(\sum_{i=1}^{c} x_i)!} e^{-\lambda} = \prod_{i=1}^{c} \frac{(\lambda p_i)^{x_i}}{x_i!} e^{-\lambda p_i}.$$

□

The last two theorems indicate that multinomial and Poisson samples are related. In fact, both kinds of samples may be seen as results of the following sampling procedure. Let the population of interest be divided into c categories, with respective probabilities in the components of the vector $\mathbf{p}$.

If the population is finite, $\mathbf{p}$ may be considered to consist of the relative sizes of the categories, and sampling with replacement (which, for large populations, is a good approximation of sampling without replacement) may be modeled by assuming that a procedure selects individuals one after the other from the population, so that each individual has the same probability of being selected, then determines which category the individual belongs to, then replaces this individual, and the procedure repeats itself, independently of the previous outcomes. For populations which are not finite, one assumes that the sampling procedure produces individuals, one after the other, independently from each other, with probabilities equal to the components of $\mathbf{p}$.

If the researcher takes, as sample, the first n individuals yielded by this procedure, the distribution of the variable $\mathbf{X}$ counting the observed frequencies of the categories is $\mathscr{M}(n,\mathbf{p})$. Indeed, whether the n individuals are selected at the same time or one after the other makes no difference, if no distinction is made between sampling with and without replacement. Another procedure is to keep observing individuals for a predetermined time period, yielding a vector of frequencies $\mathbf{Y}$ with a Poisson distribution. In this model of the sampling procedure, the difference between multinomial and Poisson samples is that in the first case, sampling is stopped after a certain number of observations were made, and in the second case, it is stopped after a certain period of time passed.

The sampling procedure works independently of the intention of the researcher to stop it after a certain number of observations were taken or after a certain amount of time passed. Therefore, if it was stopped after a certain time had passed, then given that the total that far was n, the distribution should be the same, as if the procedure was stopped because n observations had been collected.

This argument and Theorem 2.7 imply the following:

Proposition 2.6. *Assume the sampling procedure from a population with fixed sample size n yields the frequencies*

$$\boldsymbol{X} \sim \mathscr{M}(n,\boldsymbol{p}),$$

and the same sampling procedure with fixed sampling period yields the frequencies $\boldsymbol{Y}$. *Then*

$$\boldsymbol{Y}|(\sum_{i=1}^{c} y_i = n) \sim \mathscr{M}(n,\boldsymbol{p})$$

and

$$E(\boldsymbol{Y}|(\sum_{i=1}^{c} y_i = n)) = n\boldsymbol{p}.$$

Further, if also $\sum_{i=1}^{c} Y_i \sim \mathscr{P}(\lambda)$*, then*

$$\boldsymbol{Y} \sim \mathscr{P}(\boldsymbol{\lambda}), \text{ with } \lambda_i = \lambda p_i.$$

□

Based on this result, the sampling procedure described in the paragraphs before Proposition 2.6 is called (multivariate) Poisson sampling.

2.6 Sampling with Unequal Selection Probabilities

Most of the results in this book apply to data collected through multinomial or Poisson sampling, which imply equal selection probabilities of all individuals in the population. Unfortunately, most of the surveys of the human population are based on sampling procedures with unequal selection probabilities. Every real large-scale sampling procedure is a combination of three methods: stratification (discussed in Sect. 2.2.1), multistage selection, and cluster sampling. They are combined to optimize the sample with respect to reduction of data collection costs and of the standard error of the estimates. Often, knowledge about the population is used in this optimization procedure.

These sampling techniques are not discussed in this book. An excellent classical text describing these methods and the principles behind them is [38].

In general, for a finite population, a sampling procedure is specified by listing which subsets of the population are samples under the procedure and giving the probability with which any of these subsets is selected. For example, simple random sampling without replacement, with sample size n, is specified by allowing all subsets of size n of the population as samples and selecting each of these samples with the same probability. The selection probability of an individual is the sum of the selection probabilities of the samples which contain this individual. Obviously, in simple random sampling, each individual has the same selection probability.

Another sampling procedure which is often used in practice is a two-stage procedure, where first households are selected, with each household having the same selection probability, and then from each selected household, one individual is selected, again, with equal selection probabilities. This procedure, although it is the combination of two steps, both with equal selection probabilities, does not assign equal selection probabilities to the individuals in the population. Someone living in a household of size 4 has only half of the selection probability of an individual living in a household of size 2.

Unequal selection probabilities of individuals need to be taken into account in the statistical analysis of the data, although this is often a difficult task. For a fundamental result in estimation, see Sect. 4.1.4.

2.7 Things to Do

1. Determine the expected value and the variance of a variable with hypergeometric distribution.
2. Study the approximation given in Proposition 2.2. Determine the probabilities in (2.2) and (2.8) for $p_1 = 0.3$, $n = 30$, $k = 11$, for $N = 100$ and $N = 1{,}000{,}000$.
3. Define sampling without replacement in the multinomial case. Formulate and prove results similar to Propositions 2.1 and 2.2.
4. Generate plots of the correlation between two components of a multinomial distribution, when one has probability 0.1, 0.2, 0.5, 0.8, 0.9, for all possible probabilities of the other.
5. Is the absolute value of the correlation in (2.14) less than or equal to 1, as it should be? Why?
6. Assume that the population of interest is distributed according to gender and educational level as given in Table 2.5. You want to select a stratified sample of size 2000 so that the gender by education distribution of the population is preserved. What are the strata? Determine the respective sample sizes in each stratum.

Table 2.5 Population distribution according to gender and educational level (hypothetical)

	Elementary school	High school	College	Graduate degree
Men	3,225,113	18,341,666	17,455,671	5,221,456
Women	3,978,110	20,011,457	15,473,882	4,992,155

7. Suppose that in a state there are 8,310,441 automobiles registered; out of these, 7517 are vintage cars. A sample of 500 is taken to compare owners of vintage cars to ordinary car owners. It is decided that 250 observations will be selected from each group, with simple random sampling. What is the selection probability of a vintage car owner and of an ordinary car owner?
8. In a stratified sampling procedure, using 3 strata and selecting 200, 400, 200 observations, respectively, the observations are classified into 2 groups, with the following probabilities: $\mathbf{p}^1 = (0.3, 0.7)'$, $\mathbf{p}^2 = (0.5, 0.5)'$, $\mathbf{p}^3 = (0.8, 0.2)'$. What is the distribution of the resulting frequencies? Determine their covariance matrix.
9. Give examples of categorical variables such that their joint sample space is not the Cartesian product of their individual sample spaces.

10. For the three-dimensional probability distribution in Table 2.6, determine the $A \times C$ and the $B \times C$ marginal distributions. For the same table, determine the conditional distribution specified by $A = 2$, the one specified by $C = 2$, and the one specified by $A = 2$ and $C = 2$.

Table 2.6 A 2 × 4 × 3 probability distribution

	C = 1				C = 2				C = 3			
	B = 1	B = 2	B = 3	B = 4	B = 1	B = 2	B = 3	B = 4	B = 1	B = 2	B = 3	B = 4
A = 1	0.01	0.03	0.01	0.01	0.05	0.05	0.1	0.01	0.01	0.04	0.01	0.02
A = 2	0.1	0.02	0.2	0.02	0.05	0.01	0.01	0.15	0.02	0.01	0.03	0.03

11. In an $A \times B$ two-way contingency table, the conditional distributions $P(A|B = b)$ do not depend on the conditioning category b of B. Prove that in this case, the marginal distribution of A is the same as any of the conditional distributions $P(A|B = b)$. Prove that in this case the conditional distribution $P(B|A = a)$ does not depend on a.
12. Formulate the condition which implies that the distribution described in Theorem 2.3 becomes identical to the one described in Proposition 2.3.
13. Check the precision of the approximation leading to the definition of the Poisson distribution by calculating the binomial and Poisson probabilities for $\lambda = 5$, $k = 8$, $n = 10, 20, 50, 100, 500$.
14. Give an induction proof of Proposition 2.5.
15. There are famous examples of real data sets where the Poisson distribution seems to provide a good approximation of the observed data. One such data set is about the number of death by horse kicks in the Prussian army. Find this data set.
16. Perform simulations to check Proposition 2.6. Choose c, $\mathbf{p}$, and n. Simulate the behavior of $\mathbf{Y}$, without assuming that the total has a Poisson distribution. Generate 10,000 samples of size n. Compare the distribution of $\mathbf{Y}$ to $\mathscr{M}(n, \mathbf{p})$.
17. Find information about large international surveys of the human population. Study their sampling procedures. What is the distribution of the categorical variables observed?

Chapter 3
Normal Approximations

Abstract For the case of sufficiently large sample sizes, many properties of the binomial and multinomial distributions may be well approximated by normal distributions. The theoretical framework for such approximations is convergence in distribution. It is implied by the central limit theorem that when the sample size goes to infinity, appropriately normed binomial and multinomial distributions converge to the normal distribution. If, however, in the case of binomial distributions, also p converges to zero, so that np remains constant, the binomial converges to a Poisson distribution. The most important use of normal approximations is a very useful method to obtain asymptotic variance and covariance formulas for functions of binomial or multinomial variables. This δ-method is widely used in estimation and testing. Asymptotic normality applies to sampling distributions and refers to the deviation of the observed probabilities from their respective expectations. It does not affect the fundamental difference between categorical and normal assumptions with respect to the population distribution.

This chapter starts with the introduction of the concept of convergence in distribution, which is the basis of asymptotic considerations given in the chapter and later in the book.

3.1 Convergence in Distribution

Sequences of random variables may converge to a random variable in many senses. Out of these, convergence in distribution is the weakest concept. It does not guarantee that the values of the random variables in the sequence converge or become similar to the value of the limit random variable, only that their typical behavior, as summarized by their distributions, become similar. These concepts will be first illustrated with a simple example.

Let, for every n, X_n be a random variable taking on the values $0, \frac{1}{n}, \frac{2}{n}, \dots \frac{n-1}{n}$, each with probability $\frac{1}{n}$. Further, let X be a random variable that can take on any value on the interval $[0,1]$, with density function $f(x) = 1$. All these variables have a certain variant of the uniform distribution.

Intuitively, one would think that the variables X_n become more and more similar to X, as $n \to \infty$. In the current setup, one has to be careful about the precise meaning which may be associated with this claim. The joint distribution of $(X, X_1, \dots, X_n, \dots)$ is not defined, only their individual (also called marginal) distributions. Therefore, in general, no claim may be made about the joint behavior of these variables. There is, however, a certain characteristic of X_n, which converges to that of X, when $n \to \infty$. It is the distribution of X_n which converges to that of X, when $n \to \infty$. The values of random variables are stochastic quantities and no statement involving the values of many of them can be made without knowing their joint distribution. The distributions, as characterized by the respective distribution functions, depend on the behavior of the individual variables, and statements about them may be made without knowing the joint distribution. See also the comments after the proof of Theorem 3.6.

The distribution function of X_n, F_n, is[1]

$$F_n(x) = P(X_n \leq x) = \frac{k}{n}, \text{ if } 0 \leq x \leq \frac{n-1}{n} \text{ and } k-1 \leq nx < k,$$

with $F_n(x) = 0$ if $x < 0$ and $F_n(x) = 1$, if $\frac{n-1}{n} < x$. The distribution function of X is

$$F(x) = P(X \leq x) = x, \text{ if } 0 \leq x \leq 1,$$

with $F(x) = 0$ if $x < 0$ and $F_n(x) = 1$, if $1 < x$. Then, it is easily seen that

$$F_n(x) \to F(x), \text{ if } n \to \infty, \text{ for all } x.$$

Indeed, when $x < 0$ or $1 < x$, the convergence is trivial. For $0 \leq x \leq 1$, $F(x) - F_n(x) = x - k/n$, and $0 \leq |x - k/n| < 1/n$, thus $F(x) - F_n(x)$ converges to zero as $n \to \infty$.

This result means that the distribution of X_n becomes similar to the distribution of X, as $n \to \infty$. The definition of convergence in distribution generalizes the convergence established above. A sequence of random variables (X_n) is said to converge to a random variable X, if for any x, where the distribution function of X, F_X is continuous, the sequence $F_n(x)$ converges to $F(x)$. Convergence in distribution is denoted as

$$X_n \xrightarrow{d} X,$$

where d refers to the distribution of a random variable. In the above example, F was continuous everywhere, and convergence was established for all x.

[1] Sometimes the distribution function is defined as $F_n(x) = P(X_n < x)$ but the two definitions lead to the same results in this case.

One of the most fundamental results of probability theory, the central limit theorem, is a result about convergence in distribution. For a proof, see, e.g., [86].

Theorem 3.1. *Let X_n, $n = 1,2,3,\ldots$ be independent and identically distributed random variables with expectation $E(X)$ and (finite) variance $V(X)$. Then, with $S_n = X_1 + \cdots + X_n$,*

$$\frac{S_n - E(S_n)}{\sqrt{V(S_n)}} = \frac{S_n - nE(X)}{\sqrt{nV(X)}} \overset{d}{\to} Y,$$

if $n \to \infty$, where Y has a standard normal distribution. □

The central limit theorem implies that under a relatively mild assumption (finite variance), averages, when appropriately standardized, tend to be normally distributed. The speed of convergence depends on a number of factors.

The main use of convergence in distribution is to make approximate statements about probabilities of events associated with the random variables X_n. If such probabilities are hard to determine, a probability calculation associated with X_n may be replaced by the same calculation for their limiting in distribution. However, it may also occur that the distribution of X_n depends on an unknown parameter, but the limiting distribution does not. In such a case, the approximation based on convergence in distribution may be calculated, while the true value cannot. The most important such example is the asymptotic distribution of the so-called chi-squared statistics, like the Pearson statistic (see Sect. 5.2), which are often used to test model fit for categorical data. Here, the sample size plays the role of n. For any (finite) sample size, the distribution of the statistic depends on the population distribution. But the statistics converge in distribution to a random variable with a distribution that does not depend on the population distribution, only on whether or not it is in the model which is being tested.

The following simple result allows one to infer convergence in distribution for transformed variables.

Proposition 3.1. *Suppose that a sequence of random variables X_n converges in distribution to a random variable X and that f is a continuous real function. Then*

$$f(X_n) \overset{d}{\to} f(X).$$

□

A general consequence of converging in distribution to a variable with finite variance is given in the next result.

Proposition 3.2. *Suppose that $X_n \overset{d}{\to} X$, $E(X) = 0$, $V(X)$ is finite and X has a continuous distribution function. Then, for every $\varepsilon > 0$, there exist a $K(\varepsilon)$ and an $n(\varepsilon)$, such that*

$$P(|X_n| \leq K(\varepsilon)) \geq 1 - \varepsilon, \text{ if } n > n(\varepsilon). \tag{3.1}$$

Proof. Suppose that, to the contrary, there exists an $\varepsilon > 0$, for which no such bound $K(\varepsilon)$ exists. Let K_i, $1,2,\ldots$ be a sequence that converges monotone increasingly to ∞, and, for every i, (3.1) with $K(\varepsilon) = K_i$ does not hold. Then, for every i, there is a subsequence of $1,2,\ldots$, say (n_i), such that

$$a_{n_i} = P(|X_{n_i}| > K_i(\varepsilon)) \geq 1 - \varepsilon.$$

Because the values of the distribution functions converge, every subsequence (a_{n_i}) converges to the same value, implying that

$$P(|X| > K_i(\varepsilon)) \geq 1 - \varepsilon$$

for all i, while K_i converges to infinity. Then $V(X)$ cannot be finite. □

Throughout the chapter, further properties related to convergence in distribution will be discussed.

3.2 Normal Approximation to the Binomial

The normal approximation to the binomial distribution is a result about the convergence in distribution of a sequence of binomial variables to a normal variable. It is a straightforward consequence of the central limit theorem.

Theorem 3.2. *Let*

$$X_n \sim \mathscr{B}(n,p), \text{ and } X \sim N(0,1).$$

Then

$$\frac{X_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} X, \tag{3.2}$$

if $n \to \infty$.

Proof. Let Z_i be the indicator variable for the i-th observation. Then $E(Z_i) = p$, $Var(Z_i) = p(1-p)$, and $X_n = \sum_{i=1}^n Z_i$. Theorem 3.1 applied to the indicator variables yields that

$$\frac{\sum_{i=1}^n Z_i - nE(Z_i)}{\sqrt{nVar(Z_i)}} \xrightarrow{d} X,$$

where X is a standard normal random variable. But

$$\frac{\sum_{i=1}^k Z_i - nE(Z_i)}{\sqrt{nVar(Z_i)}} = \frac{X_n - np}{\sqrt{np(1-p)}}$$

□

A variant of (3.2) is that

$$\frac{X_n - np}{\sqrt{n}}$$

converges in distribution to a normal variable with expectation zero and variance $p(1-p)$. This version generalizes conveniently to the normal approximation of the multinomial distribution, see Sect. 3.3.

Formula (3.2) implies that

$$P\left(\frac{X_n - np}{\sqrt{np(1-p)}} \leq a\right) \to P(X \leq a)$$

and thus that

$$P\left(a < \frac{X_n - np}{\sqrt{np(1-p)}} \le b\right) \to P(a < X \le b). \tag{3.3}$$

Often, the result in (3.3) is used to approximate $P(a < X_n \le b)$ by

$$\boldsymbol{\Phi}\left(\frac{b - np}{\sqrt{np(1-p)}}\right) - \boldsymbol{\Phi}\left(\frac{a - np}{\sqrt{np(1-p)}}\right), \tag{3.4}$$

where $\boldsymbol{\Phi}$ is the standard normal distribution function. A limitation of the approximation in (3.4) is that while for $0 \le \varepsilon,\ \delta < 1$,

$$P(a < X_n \le b) = P(a - \varepsilon < X_n \le b + \delta), \tag{3.5}$$

the approximations of these quantities given by (3.4) are different. Therefore, it is usual to apply to both sides of (3.5), instead of (3.4), the approximation

$$\boldsymbol{\Phi}\left(\frac{b + 0.5 - np}{\sqrt{np(1-p)}}\right) - \boldsymbol{\Phi}\left(\frac{a - 0.5 - np}{\sqrt{np(1-p)}}\right), \tag{3.6}$$

which is called the continuity-corrected approximation. The formula in (3.6) also has the advantage, that the approximation of $P(X_n = b)$ as the limit of $P(a < X_n \le b)$, when $a \to b$ (for fixed n) is not zero, as it would be when using (3.4). Of course, as $n \to \infty$, the importance of the continuity correction diminishes.

The normal approximation to the binomial given in Theorem 3.2 refers to the case when n goes to infinity with a fixed p. This means that the same observational or sampling procedure is repeated for larger and larger sample sizes. Recall that it was shown in Sect. 2.5 that when n goes to infinity but at the same time p goes to zero so that $np = \lambda$ remains constant, the probability $P(X = k)$ converges to the probability that a Poisson variable with parameter λ is equal to k. Obviously, the same is true if instead of the probability of being equal to k, the probability of being less than or equal to k is being considered, implying

Proposition 3.3. *For a fixed positive λ, Let $X_n \sim \mathscr{B}(n, \lambda/n)$ and $X \sim \mathscr{P}(\lambda)$. Then*

$$X_n \overset{d}{\to} X.$$

□

Proposition 3.3 is about the behavior of a series of binomial variables describing different sampling or observational procedures, in contrast with Theorem 3.2. These two results may be summarized as saying that if $n \to \infty$, a series of binomial variables converges in distribution, to a standard normal, if standardized and p remains constant, and to a Poisson, if p goes to zero proportionally to n.

3.3 Normal Approximation to the Multinomial

A normal approximation to the multinomial distribution, similar to the one for the binomial, may be obtained. The main use of this approximation is not the easy calculation of probabilities associated with the multinomial distribution, rather the derivation of asymptotic distributions for various functions of a variable with multinomial distribution, as will be illustrated in Sect. 3.4.

Let $\mathbf{X}_n \sim \mathscr{M}(n, \mathbf{p})$ have a c dimensional multinomial distribution. Then $\mathbf{X}_1$ may be considered an indicator variable, and $\mathbf{X}_n$ is the sum of n independent copies of such a variable. Then (2.12) implies that the covariance matrix of $\mathbf{X}_1$ is

$$\boldsymbol{\Sigma} = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 & \dots & -p_1p_c \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 & \dots & -p_2p_c \\ -p_3p_1 & -p_3p_2 & p_3(1-p_3) & \dots & -p_3p_c \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ -p_cp_1 & -p_2p_2 & -p_cp_3 & \dots & p_c(1-p_c) \end{pmatrix} \tag{3.7}$$

The following result is a generalization of Theorem 3.2 for multinomial distributions. It uses the concept of convergence in distribution for vector-valued random variables. As convergence in distribution is defined by distribution functions, the concept extends word by word to the multivariate case.

Theorem 3.3. *For $\boldsymbol{X}_n \sim \mathscr{M}(n, \boldsymbol{p})$,*

$$\frac{\boldsymbol{X}_n - n\boldsymbol{p}}{\sqrt{n}} \xrightarrow{d} \boldsymbol{X},$$

where $\boldsymbol{X} \sim N(0, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ given in (3.7).

Proof. The result is a straightforward application of the multivariate central limit theorem (see, e.g., [86]) to the indicator variables $\mathbf{X}_1$. □

The formulation of Theorem 3.3 exposes the role of appropriate normalization in obtaining a limiting distribution. A similar reformulation of Theorem 3.2 was also given in Sect. 3.2. Concentrating on the univariate case, if $X_n \sim \mathscr{B}(n, p)$, $X_n - np$ has expectation zero, and $(X_n - np)/\sqrt{n}$ has a limiting normal distribution. What happens to $X_n - np$, if it is normalized by dividing it by a quantity less or more than $\sqrt{n}$?

If normalization is too weak, that is, instead of $\sqrt{n}$, normalization is by $\sqrt{n}^{1-\delta}$, for some positive δ, then the variance of $(X_n - np)/\sqrt{n}^{1-\delta} = (n^{\delta/2})(X_n - np)/\sqrt{n}$, is $p(1-p)n^{\delta}$, and this converges to infinity, so $(X_n - np)/\sqrt{n}^{1-\delta}$ does not have a limiting distribution.

If normalization is too strong, that is, $X_n - np$ is divided, instead of $\sqrt{n}$, by $\sqrt{n}^{1+\delta}$ for some positive δ, then $(X_n - np)/\sqrt{n}^{1+\delta}$ is not going to have a limiting distribution, rather it will converge to zero in the following sense. Let ε be an arbitrary positive number. Then

$$\left|\frac{X_n - np}{\sqrt{n}^{1+\delta}}\right| > \varepsilon \tag{3.8}$$

if and only if

$$\left|\frac{X_n - np}{\sqrt{n}}\right| > n^{\delta/2}\varepsilon. \tag{3.9}$$

The left hand side of (3.9) is the absolute value of a variable which converges in distribution to a normal with expectation zero and variance $p(1-p)$, while the right-hand side converges to infinity. Therefore, the probability of the event (3.9) and thus of (3.8) converges to zero, as $n \to \infty$, for any positive ε. This property is called convergence in probability to zero. Convergence of a sequence of random variables Z_n to zero is defined in general as

$$P(|Z_n| > \varepsilon) \to 0, \text{ if } n \to \infty \text{ for every positive } \varepsilon$$

and is denoted as

$$Z_n \xrightarrow{p} 0.$$

The previous argument is not restricted to the binomial distribution and it implies, with $\delta = 1$, the next result.

Proposition 3.4. *If $\sqrt{n}Z_n$ converges in distribution to a normal variable with zero expectation, then $Z_n \xrightarrow{p} 0$.* □

Applying Proposition 3.4 for $Z_n = X_n/n - p$ with $X_n \sim \mathscr{B}(n,p)$ yields the following result

$$\frac{X_n}{n} - p \xrightarrow{p} 0,$$

or, as often expressed,

$$\frac{X_n}{n} \xrightarrow{p} p.$$

This fact is the (weak) law of large numbers for binomial variables. It means that the relative frequency of an event, after n repetitions of the experiment, converges to its probability p. This fact is of central importance in statistics but also in probability theory by linking an observable quantity (the relative frequency) to an unobservable quantity (the probability) which plays a fundamental role in our theory of occurrences of the event.[2]

An interesting property of convergence in probability to zero is that if such a sequence of random variables is added to another sequence that converges in distribution, then the sum will converge in distribution to the same limit.

Theorem 3.4. *Let X_n, $n = 1, 2, \ldots$ be such that $X_n \xrightarrow{d} X$ and Y_n, $n = 1, 2, \ldots$ be such that $Y_n \xrightarrow{p} 0$. Then $X_n + Y_n \xrightarrow{d} X$.*

[2] This is the so-called frequentist view of probability. For chapter-length introductions to probability theory, see, e.g., [30] or [86], and for alternative approaches to probability, see [46].

Proof. For any random variables T and U, and $\varepsilon > 0$,

$$P(T \le a) \le P(U \le a+\varepsilon) + P(|T-U| > \varepsilon). \tag{3.10}$$

This is seen from the following series of inequalities:

$$\begin{aligned} P(T \le a) = P(T \le a, U \le a+\varepsilon) + P(T \le a, U > a+\varepsilon) \\ \le P(U \le a+\varepsilon) + P(T-U \le a-U, a-U < -\varepsilon) \\ \le P(U \le a+\varepsilon) + P(T-U < -\varepsilon) \\ \le P(U \le a+\varepsilon) + P(T-U < -\varepsilon) + P(T-U > \varepsilon) \\ = P(U \le a+\varepsilon) + P(|T-U| > \varepsilon) \end{aligned}$$

Let x be any value at which the distribution function of X, F, is continuous. Then applying (3.10) with $T = X_n + Y_n$ and $U = X_n$ yields

$$P(X_n + Y_n \le a) \le P(X_n \le a+\varepsilon) + P(|Y_n| > \varepsilon)$$

and with $T = X_n$ and $U = X_n + Y_n$ for $a - \varepsilon$ yields

$$P(X_n \le a-\varepsilon) \le P(X_n + Y_n \le a) + P(|Y_n| > \varepsilon).$$

The last two inequalities imply that

$$P(X_n \le a-\varepsilon) - P(|Y_n| > \varepsilon) \le P(X_n + Y_n \le a) \le P(X_n \le a+\varepsilon) + P(|Y_n| > \varepsilon). \tag{3.11}$$

If now we let n converge to infinity, the second terms in the first and third expressions in (3.11) converge to zero for every positive ε because Y_n converges to zero in probability, yielding that

$$\lim_{n\to\infty} P(X_n \le a-\varepsilon) \le \lim_{n\to\infty} P(X_n + Y_n \le a) \le \lim_{n\to\infty} P(X_n \le a+\varepsilon),$$

for every $\varepsilon > 0$. Because a is a point of continuity of F, this is only possible if

$$P(X \le a) = \lim_{n\to\infty} P(X_n \le a) = \lim_{n\to\infty} P(X_n + Y_n \le a).$$

□

The following result is a straightforward consequence of Theorem 3.4 and of the definition of convergence in distribution.

Proposition 3.5. *Let X_n, $n = 1,2,\ldots$ be such that $X_n \xrightarrow{d} X$ and Y_n, $n = 1,2,\ldots$ be such that $Y_n - b \xrightarrow{p} 0$, for some constant b. Then, for any constant a,*

$$aX_n + Y_n \xrightarrow{d} aX + b.$$

□

3.3.1 Normal Approximation to the Product Multinomial

The product multinomial distribution (see Sect. 2.2.1) is often used in practice, and the normal approximation to the multinomial distribution given above will now be extended to that case. As is clear from Proposition 2.3, the probabilities associated with the possible values of a variable with product multinomial joint distribution are products of multinomial probabilities. Therefore, if $\mathbf{X}_{n(i)}$ has the multinomial distribution $\mathbf{X}_{n(i)} \sim \mathscr{M}(n(i), \mathbf{p}(i))$ in the i-th stratum, then, by Theorem 3.3,

$$P\left(\frac{\mathbf{X}_{n(i)} - n(i)\mathbf{p}(i)}{\sqrt{n(i)}} \leq \mathbf{a}(i)\right) \to P(\mathbf{Y}_i \leq \mathbf{a}),\ \text{if } n(i) \to \infty,$$

where

$$\mathbf{Y}_i \sim N(n(i)\mathbf{p}(i), \mathbf{D}_{\mathbf{p}(i)} - \mathbf{p}(i)\mathbf{p}(i)').$$

Then the following result is immediate.

Theorem 3.5. *Let $(\boldsymbol{X}_{1,n(1)}, \boldsymbol{X}_{2,n(2)}, \dots, \boldsymbol{X}_{k,n(k)})$ have a product multinomial distribution with parameters $(n(1), n(2), \dots, n(k))$ and $(\boldsymbol{p}(1), \boldsymbol{p}(2), \dots, \boldsymbol{p}(k))$. Then*

$$\left(\frac{\boldsymbol{X}_{1,n(1)} - n(1)\boldsymbol{p}(1)}{\sqrt{n(1)}}, \frac{\boldsymbol{X}_{2,n(2)} - n(2)\boldsymbol{p}(2)}{\sqrt{n(2)}}, \dots, \frac{\boldsymbol{X}_{k,n(k)} - n(k)\boldsymbol{p}(k)}{\sqrt{n(k)}}\right) \xrightarrow{d} \boldsymbol{Y},$$

if $n(i) \to \infty$, $i = 1, 2, \dots, k$, where

$$\boldsymbol{Y} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

with[3]

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{D}_{\boldsymbol{p}(1)} - \boldsymbol{p}(1)\boldsymbol{p}(1)' & \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_{\boldsymbol{p}(2)} - \boldsymbol{p}(2)\boldsymbol{p}(2)' & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D}_{\boldsymbol{p}(3)} - \boldsymbol{p}(3)\boldsymbol{p}(3)' & \dots & \boldsymbol{0} \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{D}_{\boldsymbol{p}(k)} - \boldsymbol{p}(k)\boldsymbol{p}(k)' \end{pmatrix}$$

Proof. The components of $\mathbf{Y}$ are independent, just like the components of $(\mathbf{X}_{1,n(1)}, \mathbf{X}_{2,n(2)}, \dots, \mathbf{X}_{k,n(k)})$, and each component of $\mathbf{Y}$ has the asymptotic distribution of one of the components of $(\mathbf{X}_{1,n(1)}, \mathbf{X}_{2,n(2)}, \dots, \mathbf{X}_{k,n(k)})$. The joint distribution of independent normally distributed variables is normal. □

[3] Note that the $\mathbf{0}$ submatrices in $\boldsymbol{\Sigma}$ are of different sizes.

3.3.2 Interaction Does Not Disappear Asymptotically

In Sect. 1.3 of this book, a fundamental difference between categorical and continuous, normally distributed data was illustrated in terms of the possible interaction structure. Categorical variables may have a more complex interaction structure than variables with a normal distribution. What is the implication of the results of this section for the difference in the interaction structure? Does asymptotic normality of the multinomial distribution mean that the higher-order interactions that may be present in a multinomial distribution, but not in a multivariate normal distribution, disappear asymptotically? No, this is not the case.

In categorical data analysis, one has to distinguish clearly between the population distribution and the sampling distribution. There is no restriction on the structure within the population, and the population distribution is entirely characterized by the vector of probabilities $\mathbf{p}$. This approach is fully parametric, because the entire population is described by a finite number of parameters. However, the parametric approach in this case is not restricted, as it is in many other setups, by the often restrictive choice of a family of distribution, within which the parameters apply. Any of the sampling distributions discussed in Chap. 2 may be combined with any probability vector $\mathbf{p}$. The former only depends on the observational procedure but not on the structure of the population. The population distribution means structure; the sampling distribution is the stochastic component (see Sect. 1.4).

When (multivariate) normality is assumed to hold, it is often considered to apply both to the population and to the sampling distributions. In these cases, the assumption is that sampling means the observation of independent and identically distributed realizations (see Sect. 2.1.2). More precisely, in these cases, the population is characterized by its ability to produce i.i.d. observations of a given distribution. In this sense, the population and sampling distributions are not distinguished.

The asymptotic normality discussed in this chapter applies to the stochastic component, that is, the sampling variation of the (relative) frequencies around their expected values (which belong to the structural part). How the observations deviate from their respective expectations is what can be characterized by asymptotic normality. This does not affect the structure the vector $\mathbf{p}$ may have, including the lack or presence of higher-order interactions.

3.4 The δ-Method

The method to be discussed in this section is the most important procedure to derive asymptotic variance formulas for variables obtained from transformations of a multinomial variable. The same formulas are often used as approximate variance for finite but sufficiently large sample sizes. In these formulas, partial derivatives, usually denoted by ∂, play a central role, and thus it would be more precise to call the method the ∂-method.

To illustrate the kind of questions that may be answered by the application of the δ-method, consider a survey that is carried out to predict the results of the presidential elections. Assuming simple random sampling[4] and that everybody who was selected into the sample responded,[5] the number X of those who say they would vote for a given candidate has a binomial distribution with parameters n (the sample size) and p (the population fraction of those who would vote for the candidate). Therefore, the relative frequency of the yes responses is a reasonable estimator[6] of p. The properties of this estimator are largely influenced by the variance of the relative frequency, which is $p(1-p)/n$. However, the elections are not won by the candidate who has the larger fraction of votes from among all those eligible to vote but by the candidate who has the majority of the votes that are actually cast. These may be different, because many who are eligible to vote do not actually cast a vote. Thus, the fraction of votes cast for a given candidate is better estimated by Y/Z, where Z is the observed number of those who will vote and Y is the observed number of those, from among these, who will vote for the candidate of interest. To implement this procedure, the respondent would have to be classified into one of the following categories: will not vote (X_1), will vote for this candidate (X_2), and will vote for another candidate (X_3). Then $(X_1,X_2,X_3)'$ has a multinomial distribution and $Y = X_2$ and $Z = X_2 + X_3$. Thus $Y/Z = X_2/(X_2 + X_3)$ is a transformation of the multinomial. The δ-method may be used to determine the asymptotic (and approximate) variance of this quantity.

A simple case of transformation is when a one-dimensional random variable X with a (finite) variance $V(X)$ is subjected to a linear transformation. Then the variance of $aX + b$ is $a^2V(X)$. The δ-method is based on a related idea. When the function f applied to X may be differentiated, then a good approximation of it at x is $f'(x)x + c$, where x is a possible value of X and c is a constant. Then, if X is close to x, then the variance of $f(X)$ may be approximated by $[f'(x)]^2V(X)$. The following theorem (the δ-method) says that under certain circumstances this idea does work.

Theorem 3.6. *Let X_n, $n = 1,2,\ldots$ be a sequence of random variables such that for some parameter e*

$$\sqrt{n}(X_n - e) \xrightarrow{d} Y,\ Y \sim N(0,V(e)), \tag{3.12}$$

and let the real function f be differentiable at e. Then

[4] Because in such cases the population size ranges from a couple of millions to hundreds of millions of people and the sample size is a few thousands, the issue of replacement is irrelevant for practical purposes; see Sect. 2.1.

[5] This is a very unrealistic assumption. In real surveys, usually 40–70% of the selected sample persons respond. Some are not found, some reject to participate in the survey, and some choose not to respond to the relevant question. Therefore, considerations that disregard nonresponse, like the ones that follow, are important but do not reflect upon reality sufficiently precisely. Their role in comparing different surveys is similar to that of official fuel consumption figures for cars: one cannot expect to make as many miles per gallon in real life as indicated, but a car with a higher official consumption figure is likely to have a higher real consumption than another one with a lower figure.

[6] This will be justified in the next chapter of the book.

$$\sqrt{n}(f(X_n)-f(e)) \overset{d}{\to} Z,\ Z \sim N(0,[f'(e)]^2V(e)). \tag{3.13}$$

Proof. By Proposition 3.2, condition (3.12) implies that for every $\alpha > 0$, there exist a $K(\alpha)$ and an $n(\alpha)$, such that

$$P(\sqrt{n}|X_n - e| \leq K(\alpha)) \geq 1-\alpha, \text{ if } n > n(\alpha). \tag{3.14}$$

For any $\beta > 0$, let $n(\alpha,\beta)$ be the smallest integer so that $n(\alpha,\beta) \geq n(\alpha)$ and $K(\alpha)/\sqrt{n(\alpha,\beta)} < \beta$. Thus

$$P(|X_n - e| < \beta) \geq 1-\alpha, \text{ if } n > n(\alpha,\beta). \tag{3.15}$$

Denote

$$U(x_n) = (f(x_n)-f(e)) - f'(e)(x_n - e), \tag{3.16}$$

where x_n is a possible value of X_n. Then the assumption that f is differentiable at e implies that

$$\frac{U(x_n)}{|x_n - e|} \to 0, \text{ if } |x_n - e| \to 0. \tag{3.17}$$

Of course, the meaning of the condition that $|x_n - e| \to 0$ is that all values of $|x_n - e|$ are arbitrarily small for sufficiently large values of n. This is exactly what (3.15) shows to be true with probability exceeding $1-\alpha$ Therefore, for arbitrary $\gamma > 0$, there exists an $n(\alpha,\beta,\gamma) \geq n(\alpha,\beta)$, such that

$$P\left(\frac{U(x_n)}{|X_n - e|} < \gamma\right) \geq 1-\alpha \text{ if } n > n(\alpha,\beta,\gamma). \tag{3.18}$$

Because for $n > n(\alpha,\beta,\gamma)$, (3.14) implies that

$$P\left(\frac{\sqrt{n}|U(X_n|}{K(\alpha)} \leq \frac{|U(X_n)|}{|X_n - e|}\right) > 1-\alpha, \text{ if } n > n(\alpha,\beta,\gamma),$$

it follows from (3.18) that

$$P\left(\frac{\sqrt{n}|U(X_n|}{K(\alpha)} < \gamma\right) > 1-2\alpha, \text{ if } n > n(\alpha,\beta,\gamma).$$

Finally, for arbitrary $\varepsilon > 0$, choose γ in such a way that $K(\alpha)\gamma \leq \varepsilon$. Then, the last formula gives that

$$P(\sqrt{n}|U(X_n| \geq \varepsilon) \leq 2\alpha, \text{ if } n > n(\alpha,\beta,\gamma).$$

This means that

$$\sqrt{n}U(X_n) \overset{p}{\to} 0. \tag{3.19}$$

But then, by rearranging (3.16), one obtains that

$$\sqrt{n}(f(X_n)-f(e)) = \sqrt{n}f'(e)(x_n - e) + \sqrt{n}U(X_n).$$

so (3.19) implies that Theorem 3.4 may be applied and, together with Proposition 3.5 and (3.12), implies (3.13). □

A proper understanding of Theorem 3.6 and its proof requires to realize that the probability statements do not give the probability that certain sequences will be of given characteristics. Rather, for any property, they give the probability that the variable for a given n will be of the characteristic required by that property. For example, formula (3.15) is not the probability that a sequence of random variables will be of the property that, for large enough n, the members of the sequence will not deviate from e by more than β. Rather, it gives the probability that the random variable X_n does not deviate more from e than β. These statements apply to each X_n, not affected by the other members of the sequence. But if all members of the sequence, at least, for large enough n, do have a property, it becomes an asymptotic property of the entire sequence.

To illustrate the application of the δ-method, consider a variable $X_n \sim \mathscr{B}(n,p)$. If one is interested in $\ln X_n$, then the machinery of the δ-method is set up as follows. Because

$$\frac{X_n - np}{\sqrt{n}} \overset{d}{\to} Y, \text{ with } Y \sim N(0, p(1-p)),$$

the form one has to use to obtain (3.12) is

$$\sqrt{n}\left(\frac{X_n}{n} - p\right) \overset{d}{\to} Y,$$

which is about the deviation of the relative frequency from the probability (multiplied by $\sqrt{n}$). The derivative of the logarithm function at p is $1/p$, so (3.13) takes the form of

$$\sqrt{n}(\ln(X_n/n) - \ln(p)) \overset{d}{\to} Z,\ Z \sim N(0, (1-p)/p).$$

More interesting and useful results may be obtained using a multivariate version of the δ-method. Just like the univariate version given above, it also relies on connecting differentiability of the transforming function with the stochastic behavior of a sequence of random variables. In the multidimensional case, differentiability also means that the function can be well approximated locally by a linear function, but the linear approximation is multidimensional, and it is essentially multiplication by the matrix of partial derivatives.

To illustrate the idea, consider the random variables X and Y, and assume they have a joint distribution. Then it is well known that

$$Var(X+Y) = E(X+Y-E(X+Y))^2 =$$

$$E(X-E(X))^2 + E(Y-E(Y))^2 + 2E((X-E(X))(Y-E(Y))) =$$

$$Var(X) + Var(Y) + 2Cov(X,Y)$$

and

$$Var(X-Y) = E(X-Y-E(X-Y))^2 =$$

$$E(X-E(X))^2+E(Y-E(Y))^2-2E((X-E(X))(Y-E(Y)))=$$
$$Var(X)+Var(Y)-2Cov(X,Y).$$

The partial derivative vector of $X+Y$ is $(1,1)'$, and the partial derivative vector of $X-Y$ is $(1,-1)'$. Further,

$$Var(X+Y)=(1,1)\begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and

$$Var(X-Y)=(1,-1)\begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

In general, let $\mathbf{f}:\mathbb{R}^k\rightarrow\mathbb{R}^l$ be differentiable. Such a function has l coordinates f_i, and each of them is a function of k variables. Let $\nabla(\mathbf{f}))$ denote the partial derivative of $\mathbf{f}$. $\nabla(\mathbf{f}))$ is an $l\times k$ matrix, which has one row for every coordinate of $\mathbf{f}$ and one column for every variable of $\mathbf{f}$. The derivative function contains the partial derivatives of $\mathbf{f}$. More precisely,

$$\nabla(\mathbf{f}))_{(i,j)}=\frac{\partial f_i}{\partial x_j}.$$

Theorem 3.7. *Let $\boldsymbol{X}_n$, $n=1,2,\ldots$ be a sequence of k-dimensional vector-valued random variables such that for some parameter $\boldsymbol{e}$*

$$\sqrt{n}(\boldsymbol{X}_n-\boldsymbol{e})\xrightarrow{d}\boldsymbol{Y},\ \boldsymbol{Y}\sim N(\boldsymbol{0},\boldsymbol{Cov}(\boldsymbol{e})),\tag{3.20}$$

and let the real function $\boldsymbol{f}:\mathbb{R}^k\rightarrow\mathbb{R}^l$ be differentiable at $\boldsymbol{e}$. Then,

$$\sqrt{n}(\boldsymbol{f}(\boldsymbol{X}_n)-\boldsymbol{f}(\boldsymbol{e}))\xrightarrow{d}\boldsymbol{Z},\ \boldsymbol{Z}\sim N(\boldsymbol{0},\nabla(\boldsymbol{f}))(\boldsymbol{e})\boldsymbol{Cov}(\boldsymbol{e})\nabla(\boldsymbol{f}))'(\boldsymbol{e})).\tag{3.21}$$

Proof. The proof is essentially the same as that of Theorem 3.6, just all statements need to be made in a multivariate version. To facilitate this, Proposition 3.2 and Theorem 3.4 have to be reformulated for multidimensional random variables. This requires replacing absolute values by vector norms, and the inequalities between vectors need to be interpreted component-wise, just like in the definition of multidimensional distribution functions. Further, the multivariate version of (3.17) has to be used. □

To illustrate the application of the multivariate δ-method, we return to the problem discussed at the beginning of the section. To forecast the fraction of votes that will be cast for a given candidate during presidential elections, the respondents in a survey with a sample size of n are classified into the categories: will not vote ($T_{n,1}$), will vote for another candidate ($T_{n,2}$), and will vote for the given candidate ($T_{n,3}$). Then, assuming simple random sampling, $\mathbf{T}_n=(T_{n,1},T_{n,2},T_{n,3})'$ has a multinomial distribution with parameters n and $\mathbf{p}$, and the quantity of interest is $T_{n,3}/(T_{n,2}+T_{n,3})$, which is a transformation of the multinomial variable. The δ-method may be used to determine the asymptotic (and approximate) variance of this quantity. The central limit theorem implies that $(\mathbf{T}_n - n\mathbf{p})/\sqrt{n}$ converges in distribution to a normal. The δ-method applies to a random variable that, when multiplied by $\sqrt{n}$, converges in distribution to a normal. Therefore, it is $\sqrt{n}(\mathbf{T}_n/n - \mathbf{p})$ that appears in (3.20). Set

$$\mathbf{X}_n = \frac{\mathbf{T}_n}{n}.$$

The parameter $\mathbf{e}$ is $\mathbf{p}$, and the covariance matrix of $\mathbf{T}_n$ is $n\mathbf{D}_\mathbf{p} - n\mathbf{p}\mathbf{p}'$, so the covariance matrix of $\sqrt{n}(\mathbf{X}_n/n - \mathbf{p})$ is $\mathbf{D}_\mathbf{p} - \mathbf{p}\mathbf{p}' = \mathbf{Cov}(\mathbf{p})$.

The function $\mathbf{f}$ takes the form

$$\mathbf{f}(T_{n,1}, T_{n,2}, T_{n,3}) = \mathbf{f}(X_{n,1}, X_{n,2}, X_{n,3}) = \frac{X_{n,3}}{X_{n,2} + X_{n,3}},$$

so $k = 3$, $l = 1$. The partial derivative matrix of $\mathbf{f}$ is

$$\nabla(\mathbf{f})(\mathbf{X}_n) = \begin{pmatrix} 0 & \frac{-X_{n,3}}{(X_{n,2}+X_{n,3})^2} & \frac{X_{n,2}}{(X_{n,2}+X_{n,3})^2} \end{pmatrix}$$

This partial derivative matrix has to be evaluated at $E(\mathbf{X}_n) = \mathbf{p}$, yielding

$$\nabla(\mathbf{f})(\mathbf{p}) = \begin{pmatrix} 0 & \frac{-p_3}{(p_2+p_3)^2} & \frac{p_2}{(p_2+p_3)^2} \end{pmatrix}$$

Then (3.21) says that the asymptotic covariance matrix of

$$\sqrt{n}\left(\frac{X_{n,3}}{X_{n,2} + X_{n,3}} - \frac{p_3}{p_2 + p_3}\right)$$

is

$$\begin{pmatrix} 0 & \frac{-p_3}{(p_2+p_3)^2} & \frac{p_2}{(p_2+p_3)^2} \end{pmatrix} \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 \\ -p_3p_1 & -p_3p_2 & p_3(1-p_3) \end{pmatrix} \begin{pmatrix} 0 \\ \frac{-p_3}{(p_2+p_3)^2} \\ \frac{p_2}{(p_2+p_3)^2} \end{pmatrix}$$

$$= \frac{p_2p_3}{(p_2+p_3)^3}.$$

Results of similar calculations will be used in the next chapter to obtain margins of errors for various quantities routinely reported from surveys.

3.5 Things to Do

1. Use the convergence of the discrete uniform random variables to the continuous uniform random variable on $[0,1]$, to approximate $P(0.53 < X_n < 0.86)$. Why is the approximation easier to determine than the exact value of probability for a given n?

2. Determine the precision of the approximation in the previous item for $n = 10, 50, 100$.
3. Let X_k be the discrete uniform variable on $0, \frac{1}{k}, \frac{2}{k}, \dots \frac{k-1}{k}$, and define the standardized S_n as in Theorem 3.1. Determine the maximum difference between its distribution function and the standard normal distribution function for $k = 10$ and $n = 5$, 10, 50.
4. How does the difference in the previous item depend on k?
5. Give a rigorous proof of Proposition 3.1. How is the continuity of f used in the proof?
6. Illustrate with an example that Proposition 3.1 is not true if f is not continuous.
7. For $X_n \sim B(n, p)$, with $n = 5$, 10, 50 and $p = 0.1, 0.5,,$ determine $P(X_n \leq 3)$ and the approximation suggested by Theorem 3.2. How do n and p determine the accuracy of the approximation?
8. Illustrate the importance of the continuity correction by plotting the difference between the approximations in (3.4) and (3.6) for different choices of p, a, and b, as a function of n.
9. Suppose that in a survey, 2000 respondents are interviewed about their employment status. It is decided in advance that 1000 men and 1000 women will be interviewed and independent simple random samples will be selected for both men and women. In the population, 8% of women and 10% of men are unemployed. Use a product multinomial distribution to model the responses and determine the covariance matrix.
10. In the example in the previous item, 8% and 10% are not the unemployment rates for women and men, respectively. This is because the unemployment rate is not determined as a fraction of the entire population, rather as a fraction of the economically active part of the population. There are people who neither work nor look for a job, for example, those who are still students or who have already retired. Assume that a simple random sample of size 2000 is used to ask people about their employment status. The responses are classified as inactive, employed, not employed but looking for a job. The unemployment rate is the ratio of the last category to the sum of the last two. Apply the δ-method to obtain the asymptotic variance formula of the estimated unemployment rate.
11. Assume now that a stratified sample of 1000 men and 1000 women is used in the survey in the item above. Determine the asymptotic variance formula for the estimated unemployment rate.
12. Formulate and prove the multivariate version of Proposition 3.2.
13. Formulate and prove the multivariate version of Theorem 3.4.
14. Prove Proposition 3.5.
15. Prove Theorem 3.7.

Chapter 4
Simple Estimation for Categorical Data

Abstract This chapter summarizes several simple procedures often used in the analysis of categorical data. These include maximum likelihood estimation of parameters of binomial, multinomial, and Poisson distributions and also unbiased estimation with unequal selection probabilities. The Lagrange multiplier method is introduced, and maximum likelihood estimation in general parametric models is considered. In addition to the usual formula for the standard error of an estimated probability, the δ-method is used to derive asymptotic standard errors for estimates of more complex quantities, which are routinely reported in surveys. Standard errors of estimates of fractions based on stratified samples are compared to standard errors obtained from simple random samples.

This chapter starts with an introduction to maximum likelihood estimation, which is then applied to various sampling procedures and statistical models.

4.1 Introduction to Maximum Likelihood Estimation

Estimation is a procedure to make inference for parameters of distributions using the observed data. There are several procedures and principles that are used in statistics to obtain estimates. Estimation procedures are characterized by various properties. This section gives an overview of estimation in the context of categorical data, with an emphasis on maximum likelihood estimation.

A straightforward estimation procedure is the method of moments. If X is a random variable, $M_k = E(X^k)$ is called its k-th moment. Several parameters of interest may be expressed as functions of moments. For example, $E(X)$ is the first moment, and $V(X) = E(X^2) - (E(X))^2 = M_2 - (M_1)^2$. The moments of a random variable may be estimated by its empirical moments. The k-th empirical moment, based on a sample of $x_1, \ldots, x_n$, is

$$m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

The method of moments may be applied to parameters p of (the distribution of) X that may be written in the form of $p = f(M_1, \dots, M_k)$ for some function f. Then the method of moments estimate of p is

$$\hat{p} = f(m_1, \dots, m_k).$$

For example, if $X \sim \mathscr{B}(1, p)$ is an indicator variable, then $E(X) = p$, so $p = M_1$. If one has n independent observations of X, say $x_1, \dots, x_n$, then m_1, their average, is an estimate of M_1, so $\hat{p} = m_1$. Of course, the sum of the observations $x_1, \dots, x_n$ may be considered as a single observation of a $\mathscr{B}(n, p)$ variable, and the relative frequency m_1 is, also in this approach, the method of moments estimator of p based on a single observation.

If Y has a (discrete) uniform distribution on the integers in the interval $[0, u]$, where the integer u is a parameter, then its expectation is $u/2$, so the method of moments estimator of u is twice the average of the observations. If one observes $1, 1, 3, 5$, then u is estimated as 5. If the observed values are $1, 2, 3, 5$, then twice the average is 5.5. Or, if the observations are $4, 4, 4, 4$, then twice the average is 8. In the last two cases, one may debate whether the estimate should be 5 or 6, and having observed 4 all the time, whether or not 8 is reasonable as an estimate of u. These examples illustrate that principles need to be applied carefully and may not apply to all possible cases equally well.

The principle of maximum likelihood is used very widely in statistics, often leading to estimators with desirable properties but also occasionally leading to situations that are being debated, and not without reason. The principle of maximum likelihood suggests to choose the parameter that maximizes the likelihood of the observed sample compared to all other parameters. This principle applies in a setting where the distribution of the variable observed and thus the likelihood[1] of the sample depend on a finite number of parameters. To formalize the setup, assume that the probability[2] that a random variable X takes on one of its values, say, x, depends, in addition to x, on a parameter p: $P(X = x) = l(x, p)$. If n independent replications of this variable $x_1, \dots, x_n$ are observed, then the probability, or likelihood, of the sample is

$$l_n(x_1, \dots, x_n, p) = \prod_{i=1}^{n} l(x_i, p). \tag{4.1}$$

The form of the likelihood function given in (4.1) is meaningful, when the observations have already been made. Sometimes, a view of the likelihood function before any observations were taken, that is, as a random variable, is useful. In that case,

$$P(l_n(X_1, \dots, X_n, p) = l_n(x_1, \dots, x_n, p)) = P(X_1 = x_1, \dots, X_n = x_n).$$

[1] In a categorical setup, the likelihood of the sample is the same as its probability. The word likelihood is used to also refer to situations where all observations have zero probability (as with continuous random variables) but different likelihoods (as, e.g., with a normal distribution).

[2] In a more general setting, the density.

The principle of maximum likelihood suggests to choose the value of p, usually denoted by $\hat{p}$ as the estimate, for which

$$l_n(X_1,\ldots,X_n,\hat{p}) > l_n(X_1,\ldots,X_n,p),$$

for all possible values of the parameter $p \neq \hat{p}$.

In the case of $X \sim \mathscr{B}(n,p)$, $0 < p < 1$ is assumed; otherwise, only one of the categories would exist in the population. Then the likelihood of $X = x$ is

$$\binom{n}{x} p^x (1-p)^{(n-x)},$$

and the part that depends on p is

$$p^x (1-p)^{(n-x)}, \tag{4.2}$$

called the kernel of the likelihood. In fact, (4.2) is also $l_n(X_1,\ldots,X_n,p)$ for the n indicator variables $X_i \sim \mathscr{B}(1,p)$. This shows that the likelihood depends on the data only through the number of "yes" responses, so it is a sufficient statistic. To maximize the likelihood, often its logarithm is taken[3] that usually leads to a simpler problem. The kernel of the log-likelihood is

$$L(x,p) = x\ln p + (n-x)\ln(1-p),$$

with

$$x = \sum_{i=1}^{n} x_i.$$

The kernel of the log-likelihood $L(x,p)$ is considered for fixed data x, as a function of p. Its derivative in p is

$$\frac{x}{p} - \frac{n-x}{1-p},$$

and by setting it to zero, one obtains that

$$p = \frac{x}{n}.$$

The second derivative of the kernel of the log-likelihood is

$$-\frac{x}{p^2} - \frac{n-x}{(1-p)^2} = \frac{-x(1-p)^2-(n-x)p^2}{p^2(1-p)^2} = \frac{2px-x-np^2}{p^2(1-p)^2},$$

the numerator of which at $p = x/n$ is $2x^2/n - x - x^2/n = x^2/n - x = x(x/n-1) < 0$. Thus if $0 < x/n < 1$, that is, if $p = x/n$ is in the interior of the interval $(0,1)$, then $p = x/n$ uniquely maximizes $L(x,p)$, leading to the following result

[3] For the time being, it is assumed that this is possible, that is, $p > 0$. Section 4.1.1 gives a more detailed discussion that includes the case of $p = 0$ and that also applies here.

Proposition 4.1. *For $X \sim \mathscr{B}(n, p)$, if $0 < x < n$, the maximum likelihood estimate of p is*

$$\hat{p} = \frac{x}{n}.$$

□

Thus, the relative frequency is not only the method of moments but also the maximum likelihood estimate (MLE) of the probability. On the other hand, if $x = 0$ or $x = n$, there exists no MLE for $0 < p < 1$. If, for example, $x = 1$, the larger is p, the larger is the likelihood. Indeed, the likelihood

$$p^n + (1-p)^0 = p^n + 1$$

is a monotone increasing function of p, so it has no maximum on the open interval $(0, 1)$ containing p.

The above argument gave the MLE of p, and $1 - p$ is usually not considered as a parameter of the binomial, because it is a simple function of p, and it seems reasonable to think that

$$1 \hat{-} p = 1 - \hat{p}.$$

Indeed, this may be formalized using the following simple result.

Proposition 4.2. *Let $\hat{\mathbf{p}}$ the maximum likelihood estimate of the parameter $\mathbf{p}$, and let $\boldsymbol{f}$ be a one-to-one function, the domain of which contains the entire range of the parameter. Then $\boldsymbol{f}\widehat{(\mathbf{p}}) = \boldsymbol{f}(\hat{\mathbf{p}})$.*

□

4.1.1 Maximum Likelihood Estimation for the Multinomial Distribution

To extend the foregoing results to the multinomial distribution, define the maximum likelihood estimate of a vector-valued parameter, just like that of a scalar-valued parameter: its value is the one that associates a higher likelihood with the data than any other parameter value. In a k-dimensional multinomial setup, it is more common to consider all components of $\mathbf{p}$ as a parameter, as opposed to the binomial case, when out of p and $1 - p$, usually only the first one is considered to be a parameter. If $\mathbf{X} \sim \mathscr{M}(n, \mathbf{p})$ is k-dimensional, then it is usually assumed that $p_i > 0$ for all $i = 1, \ldots, k$; otherwise, $\mathbf{X}$ would be of lower dimension. The kernel of the log-likelihood is

$$L(\mathbf{x}, \mathbf{p}) = \sum_{i=1}^{k} x_i \ln p_i = \mathbf{x}' \ln \mathbf{p},$$

where the logarithm of a vector is interpreted component-wise. $L(\mathbf{x}, \mathbf{p})$ has to be maximized in $\mathbf{p}$. Note that this function does not have a maximum on all vectors $\mathbf{p}$, because $L(\mathbf{x}, t\mathbf{p}) = L(\mathbf{x}, \mathbf{p}) + n \ln t$. This underlines the fact that the domain of the function $L(\mathbf{x}, \mathbf{p})$ is the Cartesian product of the sample space of the multinomial (for $\mathbf{x}$) and of the set of k-dimensional positive probability distributions (for $\mathbf{p}$). The latter requires that $\sum_{i=1}^{k} p_i = 1$ and $p_i > 0$.

The first requirement on $\mathbf{p}$ is imposed[4] on the maximization problem. The positivity requirement is used when logarithms are taken. Finding the MLE requires solving the following constrained maximization problem:

$$\text{maximize } \mathbf{x}' \ln \mathbf{p} \text{ in } \mathbf{p}, \text{ subject to } \sum_{i=1}^{k} p_i = 1. \tag{4.3}$$

Such constrained maximization problems may be solved by the so-called Lagrange multiplier method (see, e.g., [11]), which is summarized in the following proposition.

Proposition 4.3. *If $\boldsymbol{p}$ maximizes $g(\boldsymbol{p})$ subject to $\boldsymbol{h}(\boldsymbol{p}) = \boldsymbol{0}$, then $\boldsymbol{p}$ is a stationary point of $g(\boldsymbol{p}) + \boldsymbol{\lambda}'\boldsymbol{h}(\boldsymbol{p})$ for some $\boldsymbol{\lambda}$.* □

The meaning of stationary point is a point where all partial derivatives are zero. The components of $\boldsymbol{\lambda}$ are the Lagrange multipliers, and $\mathbf{h} = \mathbf{0}$ contains the restrictions.

The Lagrange multiplier method establishes a necessary condition for a point (vector) to maximize the function g subject to constraints. This means that the vector which maximizes g subject to the given constraints is among the vectors that are stationary points of $g(\mathbf{p}) + \boldsymbol{\lambda}'\mathbf{h}(\mathbf{p})$. Whether or not a stationary point does maximize g needs to be checked, but the Lagrange multiplier method gives candidates.

In the case of the multinomial, the Lagrange multiplier method suggests solving the following unconstrained maximization problem instead of (4.3)

$$\text{maximize } \sum_{i=1}^{k} x_i \ln p_i + \lambda \left(\sum_{i=1}^{k} p_i - 1 \right).$$

The function to be maximized is called the (kernel of the) augmented log-likelihood function. Its partial derivative according to p_j for a fixed j is

$$\frac{x_j}{p_j} + \lambda, \tag{4.4}$$

and its partial derivative according to λ is

$$\sum_{i=1}^{k} p_i - 1. \tag{4.5}$$

Setting (4.4), for all j, and also (4.5) equal to zero is a system of equations. It is immediate from (4.5) that the solution fulfills the condition in (4.3). Consider the equations derived from (4.4). Multiplying both sides by p_j yields the equations

[4] Later on, maximum likelihood estimates under statistical models defined in terms of further restrictions on $\mathbf{p}$ will be determined. In those cases, the additional restrictions implied by the model need to be imposed, too.

$$x_j + \lambda p_j = 0. \tag{4.6}$$

Summing these for all values of $j = 1, \ldots, k$ yields that

$$\sum_{i=1}^{k} x_i + \lambda \sum_{i=1}^{k} p_i,$$

or $n + \lambda = 0$, implying that $\lambda = -n$. Plugging this into (4.6) yields that $p_j = x_j/n$. Is x_j/n the MLE of p_j?

Before answering this question, a treatment without Lagrange multipliers is given. This relies on a slightly different view of the likelihood function and of the constraints under which one wants to maximize it. So far, the kernel of the log-likelihood was considered without its domain given explicitly, implying that it was a function defined everywhere where it made sense, that is, for all vectors **p** that were positive (and it was an additional restriction that the components of **p** summed to 1). Now, the domain of the kernel of the log-likelihood will be only those vectors **p** that are positive and sum to 1. Then, the constrained maximization problem becomes an unconstrained maximization problem on the domain of the kernel. To see whether or not x_j/n is the MLE of p_j, one has to see if x_j/n is a(n inner) point of the domain, if the partial derivatives of the kernel of the log-likelihood are zero at x_j/n and if its second partial derivative matrix is negative definite at x_j/n.

The boundary of the domain are the vectors with some of their components equal to zero. The value of $\mathbf{p} = \mathbf{x}/n$ is an inner point if and only if $\mathbf{x} > \mathbf{0}$.

The inner points of the domain of the kernel are the vectors with all components being positive, that is, the domain is an open set in the $k-1$ dimensional space. This can be used to reparameterize the maximization problem, so that the kernel is written as a function of $p_1, \ldots, p_{k-1}$, and the domain of the k parameters is the $k-1$-dimensional simplex, that is,

$$0 < p_i,\ i = 1, \ldots, k-1,\ \sum_{i=1}^{k-1} p_i < 1,$$

and the kernel of the log-likelihood function may now be written as

$$\sum_{i=i}^{k} x_i \ln p_i = \sum_{i=i}^{k-1} x_i \ln p_i + x_k \ln(1 - \sum_{i=1}^{k-1} p_i).$$

The partial derivative according to p_j, $j = 1, \ldots, k-1$ is

$$\frac{x_j}{p_j} + \frac{x_n}{\sum_{i=1}^{k-1} p_i - 1}. \tag{4.7}$$

When $p_j = x_j/n$, the above partial derivative is

$$n + \frac{x_n}{\frac{n-x_n}{n} - 1} = n + \frac{x_n}{\frac{-x_n}{n}} = 0.$$

Thus, when the observed frequencies are positive, all the partial derivatives of the kernel of the log-likelihood on the simplex are zero at the points x_j/n, for $j = 1, \ldots, k-1$. The converse of this argument proves Proposition 4.3 in the present case.

The second derivative is a matrix consisting of partial derivatives of the first derivatives. The derivative of (4.4) according to p_j is

$$-\frac{x_j}{p_j^2}$$

and is zero according to any of the other components of $\mathbf{p}$. Thus, the matrix of second partial derivatives is

$$\boldsymbol{\Sigma} = \begin{pmatrix} -\frac{x_1}{p_1^2} & 0 & 0 \ldots & 0 \\ 0 & -\frac{x_2}{p_2^2} & 0 \ldots & 0 \\ 0 & 0 & -\frac{x_3}{p_3^2} \ldots & 0 \\ . & . & .... & . \\ . & . & .... & . \\ . & . & .... & . \\ 0 & 0 & 0 \ldots & -\frac{x_{k-1}}{p_{k-1}^2} \end{pmatrix} \tag{4.8}$$

and this matrix is negative definite.

Therefore, by a standard result in calculus, if $\mathbf{x}/n$ is a(n inner) point of the domain of the likelihood function (or of the kernel of the log-likelihood), then $\mathbf{p} = \mathbf{x}/n$ maximizes its value. This yields the following result.

Theorem 4.1. *Let $\boldsymbol{X}$ be $\mathscr{M}(\boldsymbol{p}, n)$ with $\boldsymbol{p} > \boldsymbol{0}$. If $\boldsymbol{x} > \boldsymbol{0}$, then*

$$\hat{\boldsymbol{p}} = \frac{\boldsymbol{x}}{n}$$

is the MLE of $\boldsymbol{p}$.

□

If some of the components of the observed vector $\mathbf{x}$ are zero, $\mathbf{x}/n$ is not the MLE of $\mathbf{p}$, because it is outside of the domain.

In fact, the method of the proof presented for Theorem 4.1 applies more generally. In the analysis of categorical data, one very often wishes to test statistical hypotheses. These are restrictions on $\mathbf{p}$, in addition to the restrictions considered so far. An important step in testing the hypothesis that such a model holds in the population of interest is to find maximum likelihood estimates under the model. Although such models are going to be considered only later on in the book, the following important result can be proved with the methods already developed.

Theorem 4.2. *Let $H = \{\boldsymbol{p} : \boldsymbol{h}(\boldsymbol{p}) = \boldsymbol{0}\}$ be a statistical model, such that $\boldsymbol{h}$ is a $k-l$-dimensional function and that there exists a partitioning of $\boldsymbol{p}' = (p_1, \ldots, p_l, p_{l+1}, \ldots, p_k) = (\boldsymbol{p}'_1, \boldsymbol{p}'_2)'$ and a function $\boldsymbol{f} : \boldsymbol{R}^l \to \boldsymbol{R}^{k-l}$, such that*

$$\boldsymbol{h}(\boldsymbol{p}_1,\boldsymbol{p}_2)=\boldsymbol{0} \text{ if and only if } \boldsymbol{p}_1 \in S \text{ and } \boldsymbol{p}_2=\boldsymbol{f}(\boldsymbol{p}_1),$$

for some open set S in the l-dimensional Euclidean space. Assume, further, that the partial derivative matrix $\boldsymbol{h}$ *is of full rank.*

Consider the kernel of the augmented log-likelihood function

$$\boldsymbol{x}'\ln\boldsymbol{p}+\boldsymbol{\lambda}'\boldsymbol{h}(\boldsymbol{p}) \tag{4.9}$$

and the kernel of the reduced log-likelihood function

$$\boldsymbol{x}_1'\ln\boldsymbol{p}_1+\boldsymbol{x}_2'\ln\boldsymbol{f}(\boldsymbol{p}_1). \tag{4.10}$$

Then,

(i) if $(\boldsymbol{q},\boldsymbol{\lambda})$ is a stationary point of (4.9), then $\boldsymbol{q}_1$ is a stationary point of (4.10)

(ii) if $\boldsymbol{q}_1 \in S$ is a stationary point of (4.10), then there exists a $\boldsymbol{\lambda}$, such that $((\boldsymbol{q}_1',\boldsymbol{f}(\boldsymbol{q}_1)')',\boldsymbol{\lambda})$ is a stationary point of (4.9).

Proof. The partial derivatives of (4.9) are, for all λ_j, $j=1,\dots,k-l$,

$$h_j(\mathbf{p}) \tag{4.11}$$

and, for any p_i, $i=1,\dots,k$,

$$\frac{x_i}{p_i}+\boldsymbol{\lambda}'\frac{\partial\mathbf{h}}{\partial p_i}. \tag{4.12}$$

The partial derivatives of (4.10) for any p_i, $i=1,\dots,l$, are

$$\frac{x_i}{p_i}+\sum_{j=1}^{k-l}\frac{x_{l+j}}{f_j(\mathbf{p}_1)}\frac{\partial f_j(\mathbf{p}_1)}{\partial p_i}. \tag{4.13}$$

To see (i), note that by assumption, $\mathbf{h}(\mathbf{q})=\mathbf{0}$ and, thus $\mathbf{q}_1\in S$, and $\mathbf{q}_2=\mathbf{f}(\mathbf{q}_1)$. Given that (4.12) is zero at $(\mathbf{q},\boldsymbol{\lambda})$, one obtains that for $i=1,\dots,k$,

$$\left(\frac{x_i}{p_i}\right)(\mathbf{q})=\left(-\boldsymbol{\lambda}'\frac{\partial\mathbf{h}}{\partial p_i}\right)(\mathbf{q}),$$

where the notation emphasizes that both sides need to be evaluated at $\mathbf{q}$. Plugging this into (4.13), one obtains that for $i=1,\dots,l$,

$$\left(-\boldsymbol{\lambda}'\frac{\partial\mathbf{h}}{\partial p_i}-\sum_{j=1}^{k-l}p_{l+j}\boldsymbol{\lambda}'\frac{\partial\mathbf{h}}{\partial p_{l+j}}\frac{1}{f_j(\mathbf{p}_1)}\frac{\partial f_j(\mathbf{p}_1)}{\partial p_i}\right)(\mathbf{q}).$$

Using that when evaluated at $\mathbf{p}=\mathbf{q}$,

$$f_j(\mathbf{p}_1)=p_{l+j},$$

for $j=1,\dots,k-l$, (4.13) evaluated at $\mathbf{q}$ is

$$-\boldsymbol{\lambda}'\left(\frac{\partial \mathbf{h}}{\partial p_i}+\sum_{j=1}^{k-l}\frac{\partial \mathbf{h}}{\partial p_{l+l}}\frac{\partial f_j(\mathbf{p}_1)}{\partial p_i}\right)(\mathbf{q}), \tag{4.14}$$

for $i=1,\ldots,l$. Define now on the open set S

$$\mathbf{h}^*(\mathbf{p}_1)=\mathbf{h}(\mathbf{p}_1,\mathbf{f}(\mathbf{p}_1)).$$

By assumption, $\mathbf{h}^*=\mathbf{0}$ on S, and, thus, its partial derivatives according to p_i, for $i=1,\ldots,l$ are all zero on S, including $\mathbf{p}_1=\mathbf{q}_1$. This means that for all $i=1,\ldots,l$,

$$\frac{\partial \mathbf{h}^*}{\partial p_i}(\mathbf{q}_1)=\left(\frac{\partial \mathbf{h}}{\partial p_i}+\sum_{j=1}^{k-l}\frac{\partial \mathbf{h}}{\partial p_{j+l}}\frac{\partial f_j(\mathbf{p}_1)}{\partial p_i}\right)(\mathbf{q_1},\mathbf{f}(\mathbf{q_1}))=\mathbf{0}. \tag{4.15}$$

Finally, by recalling that $\mathbf{f}(\mathbf{q}_1)=\mathbf{q}_2$, by assumption, one obtains that (4.14) is zero.

To see (ii), note first that $\mathbf{q}_1\in S$; thus $\mathbf{h}(\mathbf{q}_1,\mathbf{f}(\mathbf{q}_1))=\mathbf{0}$ and (4.11), is zero for $\mathbf{p}=(\mathbf{q}_1',\mathbf{f}(\mathbf{q}_1)')'$.

That (4.12) is also zero under the conditions will be proved first for $i=l+1,\ldots,k$, then for $i=1,\ldots,l$. When considered for $i=l+1,\ldots,k$, (4.12) at $\mathbf{f}(\mathbf{q}_1)$ is a system of linear equations in the components of $\boldsymbol{\lambda}$, with a full rank matrix of coefficients; thus, there exists a choice of $\boldsymbol{\lambda}$, such that (4.12) is zero for all $i=l+1,\ldots,k$, when evaluated at $\mathbf{f}(\mathbf{q}_1)$.

In the case of $i=1,\ldots,l$, (4.13) is zero when evaluated at $\mathbf{q}_1$ thus

$$\frac{x_i}{p_i}(\mathbf{q}_1)=-\sum_{j=1}^{k-l}\frac{x_{l+j}}{f_j(\mathbf{p}_1)}\frac{\partial f_j(\mathbf{p}_1)}{\partial p_i}(\mathbf{q}_1),$$

and using this, (4.12) at $(\mathbf{q}_1)$ may be written as

$$-\sum_{j=1}^{k-l}\frac{x_{l+j}}{f_j(\mathbf{p}_1)}\frac{\partial f_j(\mathbf{p}_1)}{\partial p_i}+\boldsymbol{\lambda}'\frac{\partial \mathbf{h}}{\partial p_i}(\mathbf{q}_1).$$

It is implied by (4.15) that for $i=1,\ldots,l$,

$$\frac{\partial \mathbf{h}}{\partial p_i}(\mathbf{q}_1)=-\sum_{j=1}^{k-l}\frac{\partial \mathbf{h}}{\partial p_{l+j}}\frac{\partial f_j(\mathbf{p}_1)}{\partial p_i}(\mathbf{q}_1),$$

and plugging this into the previous formula, one obtains for (4.12) that

$$\left(-\sum_{j=1}^{k-l}\frac{x_{l+j}}{f_j(\mathbf{p}_1)}\frac{\partial f_j(\mathbf{p}_1)}{\partial p_i}-\sum_{j=1}^{k-l}\boldsymbol{\lambda}'\frac{\partial \mathbf{h}}{\partial p_{l+j}}\frac{\partial f_j(\mathbf{p}_1)}{\partial p_i}\right)(\mathbf{q}_1)=$$

$$-\left(\frac{\partial f_j(\mathbf{p}_1)}{\partial p_i}\sum_{j=1}^{k-l}\left(\frac{x_{l+j}}{f_j(\mathbf{p}_1)}+\boldsymbol{\lambda}'\frac{\partial \mathbf{h}}{\partial p_{l+j}}\right)\right)(\mathbf{q}_1)=$$

$$-\left(\frac{\partial f_j(\mathbf{p}_1)}{\partial p_i}\sum_{j=1}^{k-l}\left(\frac{x_{l+j}}{p_{l+j}}+\boldsymbol{\lambda}'\frac{\partial \mathbf{h}}{\partial p_{l+j}}\right)\right)(\mathbf{q}_1)=\mathbf{0},$$

because, the values of the terms between the inner parentheses are zero, as implied by (4.12) being zero for $j = l+1,\dots,k$. □

The rank condition of the partial derivative matrix of the model function $\mathbf{h}$ was only used to imply that the system of linear equations in (4.12) for $j = l+1,\dots,k$ has a solution in $\boldsymbol{\lambda}$. If this is assured in another way, the theorem still holds. The usefulness of the Lagrange multiplier method and of Theorem 4.2 lies in the fact that the statistical models of interest are often specified implicitly, as restrictions on the cell probabilities, and the function $\mathbf{f}$ is not always easy to find. In such cases, the augmented likelihood function yields the stationary points without a need to determine $\mathbf{f}$. The value of l is the dimension of the model. On the other hand, the quantity $k-l$ is called the (residual) degree of freedom of the model and plays an important role in testing the fit of models. Note that the proof of Theorem 4.2 did not rely on the proof of the Lagrange multiplier method (Proposition 4.3).

For example, if $k = 3$, let the model assume that $\mathbf{p} > \mathbf{0}$ and $p_2 = p_3$. Then,

$$\begin{aligned} h_1(\mathbf{p}) &= p_1 + p_2 + p_3 - 1 \\ h_2(\mathbf{p}) &= p_2 - p_3 \end{aligned} \tag{4.16}$$

and for $p_1 \in S = (0,1)$,

$$\mathbf{f}(p_1) = \left(\frac{1-p_1}{2}, \frac{1-p_1}{2}\right)',$$

so that $\mathbf{p}_1 = (p_1)$.

The partial derivatives of the kernel of the augmented log-likelihood function are

$$\frac{x_1}{p_1} + \lambda_1$$

$$\frac{x_2}{p_2} + \lambda_1 + \lambda_2 \tag{4.17}$$

$$\frac{x_3}{p_3} + \lambda_1 - \lambda_2 \tag{4.18}$$

$$p_1 + p_2 + p_3 - 1 \tag{4.19}$$

$$p_2 - p_3 \tag{4.20}$$

according to p_1, p_2, p_3, λ_1, λ_2, respectively. All these expressions are set equal to zero, and the unique solution of the system of equations obtained is

$$q_1 = \frac{x_1}{n}, q_2 = \frac{1-p_1}{2}, q_3 = \frac{1-p_1}{2}. \tag{4.21}$$

The partial derivative of the kernel of the reduced log-likelihood function is

$$\frac{x_1}{p_1} + \frac{2x_2}{1-p_1}\frac{-1}{2} + \frac{2x_3}{1-p_1}\frac{-1}{2},$$

and this function at $p_1 = q_1$ is

$$\frac{nx_1}{x_1} + \frac{2nx_2}{n-x_1}\frac{-1}{2} + \frac{2nx_3}{n-x_1}\frac{-1}{2}$$

$$= n - \frac{n(x_2+x_3)}{n-x_1} = n - \frac{n(x_2+x_3)}{x_2+x_3} = 0.$$

That is, the stationary point of the augmented log-likelihood function is also a stationary point of the reduced log-likelihood function. Conversely, when the partial derivative of the kernel of the reduced log-likelihood function is set to zero, for $p_1 \in (0,1)$, one obtains that

$$x_1(1-p_1) - x_2p_1 - x_3p_1 = 0$$

or

$$x_1 - np_1 = 0;$$

thus, the only solution is q_1.

For $0 < x_1 < n$, the second derivative at q_1 is negative; thus, the reduced form of the likelihood function is maximized at the stationary point. However, if $x_1 = 0$, the likelihood of the model cannot be maximized: $p_1 = 0$ is not in the model, but the closer is q_1 to zero, the larger is the maximized likelihood associated with $(q_1, (1-q_1)/2), (1-q_1)/2)'$.

We return now to maximum likelihood estimation without any model imposed and reconsider the assumption of $\mathbf{p} > \mathbf{0}$. The positivity assumption has the consequence that the log-likelihood function can be maximized instead of the likelihood function. One may, however, ask whether the likelihood function could be maximized directly, and then the assumption would not be needed.

In the case of domains that are not open, the machinery presented above does not work. If the likelihood function has its maximum on the boundary, the derivative will not be zero at the maximum, because it only exists for inner points. In such cases, other methods need to be applied to find the MLE. To illustrate this situation, let $k = 3$ and $\mathbf{p} \geq \mathbf{0}$, so the domain is not open. If the observations are $(0, 10, 20)'$, the likelihood function (disregarding a constant term) is

$$(1-p_2-p_3)^0 p_2^{10} p_3^{20} \tag{4.22}$$

and this function has a unique maximum at $(0, 1/3, 2/3)'$, but its derivative is never zero (the derivative exists for positive $\mathbf{p}$ values only). The maximum is obtained because $p_1^0 = 1$, and the other two terms are maximized when $p_2 + p_3 = 1$. On the other hand, if $\mathbf{p} > \mathbf{0}$ is assumed, no MLE exists.

A more formal argument for finding the MLE in the case of $\mathbf{p} \geq \mathbf{0}$ yields the following generalization of Theorem 4.1.

Theorem 4.3. *Let $\boldsymbol{X}$ be $\mathscr{M}(\boldsymbol{p},n)$. Then*

$$\hat{\boldsymbol{p}} = \frac{\boldsymbol{x}}{n}$$

is the MLE of $\mathbf{p}$.

Proof. Because $\sum_{i=1}^{k} x_i = n$, there are positive observed frequencies. Without restriction of generality, it may be assumed that the observations with indexes $1,\dots,j$ are positive, and the remaining observations are zero, with j being at least 1 and at most k. Then, the kernel of the likelihood is

$$\left(\prod_{i=1}^{j} p_i^{x_i}\right)\left(\prod_{i=j+1}^{k} p_i^{0}\right), \tag{4.23}$$

with the understanding that in the case of $j = k$, the second term does not exist, or, equivalently, its value may be taken as 1. But even if $j < k$, that is, if zero observations exist, the value of the second term of (4.23) is 1. Therefore, no $\mathbf{p}$ can maximize (4.23) unless $p_i = 0$ for $i \geq j+1$, because reallocating probability to the first term from the second increases value of the former and does not change the value of the latter . Thus, the value of (4.23) for any choice of $\mathbf{p}$ that may maximize it is the value of the first term. If the first term is maximized by an appropriate choice of the first j components of $\mathbf{p}$, then the same values and zeros for the remaining components will maximize the likelihood, because the second term is always 1.

To maximize the first term is the same as maximizing a j dimensional multinomial with positive observations. If any of the first j components of $\mathbf{p}$ is zero, the likelihood is zero, because having nonzero observations from a cell with zero probability has zero probability. Therefore, if the likelihood can be maximized, it will be maximized for positive values for the first j components of $\mathbf{p}$. By Theorem 4.1, the relative frequencies maximize the first component.

Finally observe that the zeros in the case of probabilities in the second component are the relative frequencies, too. □

The MLE $\hat{\mathbf{p}}$ is, of course, a random variable. This fact is emphasized by writing

$$\hat{\mathbf{p}} = \frac{\mathbf{X}}{n}.$$

Then, the result after Proposition 3.4 implies that

$$\hat{\mathbf{p}} \xrightarrow{p} \mathbf{p}$$

and Theorem 3.3 implies the asymptotic normality of the MLE.

Theorem 4.4. *Let $\boldsymbol{X}$ be $\mathscr{M}(\boldsymbol{p},n)$. Then*

$$\sqrt{n}(\hat{\boldsymbol{p}} - \boldsymbol{p}) \xrightarrow{d} \boldsymbol{Y},$$

where $\boldsymbol{Y} \sim N(\boldsymbol{0},\boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ given in (3.7). □

Often, in particular for tests of fit of statistical models, instead of maximum likelihood estimates of cell probabilities, maximum likelihood estimates of the cell frequencies under the sampling distribution are needed. More precisely, if $\mathbf{X}$ is a random variable, its expectation $E(\mathbf{X})$ is a parameter of its distribution and may be estimated by maximum likelihood.

Proposition 4.4. *Let $\boldsymbol{X}$ be $\mathscr{M}(\boldsymbol{p},n)$. Then, the MLE of $E(\boldsymbol{X})$ is*

$$E(\hat{\boldsymbol{X}}) = n\hat{\boldsymbol{p}}.$$

Proof. $E(\mathbf{X}) = n\mathbf{p}$, and Proposition 4.2 implies the result. □

4.1.2 Maximum Likelihood Estimation for the Poisson Distribution

If $X \sim \mathscr{P}(\lambda)$, then for any positive integer k, $P(X = k) = \lambda^k e^{-\lambda}/k!$. If $\lambda > 0$ is assumed (otherwise, only $X = 0$ would have positive probability), the kernel of the log-likelihood is

$$k\ln(\lambda) - \lambda$$

and its derivative in λ is

$$\frac{k}{\lambda} - 1.$$

If this is set to zero, one obtains the solution

$$\lambda = k. \tag{4.24}$$

As the second derivative of the kernel of the log-likelihood is $-k/\lambda^2$, (4.24) gives the maximum. This implies the following result:

Theorem 4.5. *If $X \sim \mathscr{P}(\lambda)$, then*

$$\hat{\lambda} = x$$

is the MLE of λ. □

In the multivariate case, the components of the Poisson distribution were defined to be independent. The joint probability is the product of the probabilities of the individual components, and, consequently, the joint likelihood of a sample is the product of the individual likelihoods. To see this in more detail, let a c-dimensional variable $\mathbf{X}$ have the distribution $\mathscr{P}(\boldsymbol{\lambda})$. Every component of $\boldsymbol{\lambda}$ is nonnegative, so $\boldsymbol{\lambda} \in S$, where

$$S = \times_{i=1}^{c}[0,\infty).$$

This means that the joint domain of the individual parameters $\lambda_1,\ldots,\lambda_c$ is the Cartesian product of their individual domains. This property is referred to as the parameters being variation independent, meaning that they can vary independently of each other. This is a very desirable property that will be further discussed in Sect. 6.1.2. Variation independence does not hold for every set of parameters. For example, the parameters $\mathbf{p} = (p_1, \dots p_k)'$ of a multinomial distribution are not variation independent because they have to sum to 1. So while $p_1 = 0.4$ is a possible choice of that parameter, just like $p_2 = 0.7$ is a possible choice, these two choices are not possible at the same time.

The likelihood function of the multivariate Poisson distribution is

$$P(\mathbf{X} = \mathbf{k}) = \prod_{i=1}^{c} P(X_i = k_i) = \prod_{i=1}^{c} \frac{\lambda_i^{k_i}}{k_i!} e^{-\lambda_i}.$$

In this case, the likelihood function is the product of such components that each component depends on one component of the parameter vector. This is called likelihood independence of the parameters, and the likelihood is maximized, if each of its components is maximized.

This situation largely simplifies maximizing the likelihood function. The following result is straightforward but, because of its importance, is formulated as a theorem.

Theorem 4.6. *For a parameter* $\boldsymbol{p}$, *let* $\boldsymbol{p} \in H$ *be a statistical model so that the components of* $\boldsymbol{p}$ *are both likelihood and variation independent. Then the MLE* $\hat{\boldsymbol{p}}$ *is obtained by separately maximizing the components of the likelihood function in the components of* $\boldsymbol{p}$.

A direct application of Theorem 4.6 yields that

Theorem 4.7. *If* $X \sim \mathscr{P}(\boldsymbol{\lambda})$, *then*

$$\hat{\boldsymbol{\lambda}} = \boldsymbol{x}$$

is the MLE of the parameter. □

An interesting consequence of this result is that the distribution of the MLE of the parameter of a Poisson distribution is also Poisson.

Obviously, if variation independence does not hold, the parameter obtained by component-wise maximization may not be a possible parameter. The components of the probability vector of a multinomial distribution are also likelihood independent, as the likelihood of $\mathbf{X}$ is proportional to

$$\prod_{i=1}^{c} p_i^{X_i},$$

but the components are not variation independent. Variation independence and likelihood independence will be discussed further in Chap. 5.

The next theorem relates maximum likelihood estimates under multinomial and Poisson sampling.

Theorem 4.8. *If* $\boldsymbol{X} \sim \mathscr{M}(n, \boldsymbol{p})$ *and* $\boldsymbol{Y} \sim \mathscr{P}(\boldsymbol{\lambda})$ *are of the same dimension c, and* $X_i = Y_i$, *for* $i = 1, \dots c$, *then*

$$n\hat{\boldsymbol{p}} = \hat{\boldsymbol{\lambda}}.$$

Proof. The result is implied by Theorems 4.3 and 4.7. □

The previous result is not surprising, as the likelihoods under multinomial and Poisson sampling are closely related.

Theorem 4.9. *If the observations* $\mathbf{X}$ *are distributed according to* $\mathscr{M}(n,\mathbf{p})$ *or* $\mathscr{P}(t\mathbf{p})$ *for some positive* t*, then the kernels of the likelihoods for* $\mathbf{p}$ *coincide.*

Proof. When the distribution of the observations is multinomial, the kernel of the log-likelihood is

$$\sum_{i=1}^{c} x(i)\log p(i).$$

When the distribution is Poisson, the log-likelihood is

$$\sum_{i=1}^{c}(x(i)\log tp(i) - tp(i) - \log x(i)!) =$$

$$\sum_{i=1}^{c}(x(i)\log p(i) - \log x(i)!) - t + \log t\sum_{i=1}^{c} x(i),$$

thus the part that depends on $\mathbf{p}$ is

$$\sum_{i=1}^{c} x(i)\log p(i).$$

That is, the kernels of the log-likelihoods are identical. □

4.1.3 *Maximum Likelihood Estimation in Parametric Models*

The setup considered here is related to the situation discussed around Theorem 4.2. In that case, some of the cell probabilities were functions of other probabilities because of the model considered. Here, the probabilities are functions of other parameters. In the example given in (4.16), the model of interest is formulated as restrictions on the cell probabilities, $\mathbf{h}(\mathbf{p}) = \mathbf{0}$. An alternative way to formulate the same model is to assume that there exists a parameter q, and

$$p_1 = q,\ p_2 = q,\ p_3 = 1 - 2q.$$

For this parametric specification, one needs to determine the range of the parameter q, which is clearly the open interval $T = (0, 0.5)$. Write $\mathbf{p}(q)$ for the probability vector as a function of the parameter q. Then, the kernel of the log-likelihood may be written as

$$\mathbf{x}'\log\mathbf{p}(q),$$

as a function of q. It is implied by Proposition 4.2 that if the above function is maximized in q on T by the value $\hat{q}$, then $\mathbf{p}(\hat{q})$ is the maximum likelihood estimate of the probability vector.

The formulation above also included the restriction that the cell probabilities sum to 1. With many of the models to be discussed in this book, including this requirement in the model formulation is not convenient. This is because most of the models to be considered are of a multiplicative nature, and enforcing that the probabilities sum to 1, which is an additive constraint, would make the model formulation too complicated. To apply this to the above example (where the application is not really warranted, but later on there will be situations where it is), one may write

$$p_1 = q_1, \; p_2 = q_1, \; p_3 = q_2.$$

In this case, the kernel of the augmented log-likelihood function is

$$\mathbf{x}' \log \mathbf{p}(\mathbf{q}) + \lambda \left(\sum_{i=1}^{3} p_i(\mathbf{q}) - 1 \right). \tag{4.25}$$

The next theorem establishes a general method of finding maximum likelihood estimates in these cases.

Theorem 4.10. *Let the model of interest consist of all probability distributions* $\boldsymbol{p}$, *which can be written in the form of* $\boldsymbol{p} = \boldsymbol{f}(\boldsymbol{q})$ *for some differentiable function* $\boldsymbol{f}: \boldsymbol{R}^l \rightarrow \boldsymbol{R}^k$ *and some parameter* $\boldsymbol{q}$, *such that the model is an open set in* $\boldsymbol{R}^k$. *Consider the kernel of the augmented log-likelihood function*

$$\boldsymbol{x}' \log \boldsymbol{f}(\boldsymbol{q}) + \lambda \left(\boldsymbol{1}'\boldsymbol{f}(\boldsymbol{q}) - 1 \right). \tag{4.26}$$

If $(\hat{\boldsymbol{q}}, \hat{\lambda})$ *is a stationary point of (4.26), considered as a function of* $\boldsymbol{q}$ *and of* λ, *then* $\boldsymbol{p} = \boldsymbol{f}(\hat{\boldsymbol{q}})$ *is a probability distribution, and it is a stationary point of the kernel of the log-likelihood function*

$$\boldsymbol{x}' \log \boldsymbol{f}(\boldsymbol{q}). \tag{4.27}$$

Proof. Because the derivative of (4.26) according to λ is the deviation of the sum of the components of $\mathbf{p}$ from 1, for the stationary point $\hat{\mathbf{q}}$, $\mathbf{f}(\hat{\mathbf{q}})$ is in the model. The function $\mathbf{f}$ is differentiable, thus continuous, and then the set of $\mathbf{q}$ values, say T, for which $\mathbf{f}(\mathbf{q})$ is in the model, which is an open set, is also open. For every $\mathbf{q} \in T$, the functions in (4.26) and in (4.27) are equal. As T is an open set, also the derivatives of these functions according to $\mathbf{q}$ are equal. □

4.1.4 Unbiased Estimation with Unequal Selection Probabilities

As discussed in Sect. 2.1, the binomial distribution is a reasonable approximation of the sampling distribution in real surveys, when the selection probability is the same for each member of the population, and the same applies to the multinomial distribution, as well. In the practice of surveys, however, there are several sampling procedures which do not assign equal selection probabilities to each member of the population. For example, if the population of interest consists of those, who have a landline phone at home, a survey may be conducted by calling the phone number of a household and then selecting a person living in this household to be interviewed. This is a two-stage sampling procedure. Even if each phone number is selected with equal probability and the individual from a selected household is chosen in such a way that each person living in the household has the same probability of selection, the sample units are not going to have equal selection probabilities. Indeed, a person living alone will have twice the chance of being included in the sample, compared to someone living in a two-person household. This is often formulated by saying that individuals from larger households are underrepresented in the sample. See, for example, [38] for a description of such and related sampling procedures. In certain cases, stratified sampling also leads to unequal selection probabilities.

To formalize this situation, assume that the size of the population of interest is N, and it is divided into k mutually exclusive and exhaustive groups of size N_i, $i = 1, \ldots, k$. Let the sampling procedure consist of choosing a simple random sample of size n_i from group i. Thus, the selection probability for every individual in group i is $p_i = n_i/N_i$. Finally, let the total sample size be denoted by n.

The variable X has a fixed value for every member of the population, and it is random sampling of the observations which make it a random variable. Let x_{ij} be its value on the j-th selected individual in group i, $j = 1, \ldots, n_i$, $i = 1, \ldots, k$. The next theorem gives an unbiased estimator of $E(X)$, often called the Hansen-Hurwitz estimator. $E(X)$, in this case, is the population average of X.

Theorem 4.11. *Let every individual in a population of size N possess a value X, and let this population be divided into mutually exclusive and exhaustive subpopulations of respective sizes N_i, $i = 1, \ldots, k$. Let a sample of total size n be selected so that from subpopulation i, n_i individuals are selected with simple random sampling without replacement, $i = 1, \ldots, k$. Then, an unbiased estimator of $E(X)$ is*

$$\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{X_{ij}}{p_i}$$

where $p_i = n_i/N_i$ is the selection probability within the i-th group and X_{ij} is the value of the j-th individual in the i-th group.

Proof. For the sum of the X values in group i,

$$\frac{N_i}{n_i}\sum_{j=1}^{n_j} X_{ij}$$

that is, the observed average times the group size is an unbiased estimator. Thus, for the population total

$$\sum_{i=1}^{k} \frac{N_i}{n_i} \sum_{j=1}^{n_j} X_{ij}$$

is an unbiased estimator. Finally, for the population average

$$\frac{1}{N} \sum_{i=1}^{k} \frac{N_i}{n_i} \sum_{j=1}^{n_j} X_{ij} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{X_{ij}}{p_i}$$

is unbiased. □

This result is often referred to by saying, that to obtain an unbiased estimate of the average, weights inversely proportional to the selection probabilities have to be applied.

4.2 Standard Errors of Quantities Reported in Surveys

The characteristics of many of the quantities estimated and reported in surveys can be described using methods developed in the present and the previous chapters. We focus here on surveys where the population is much larger than the sample size. Examples include surveys that are conducted to predict results of (general or presidential) elections and are based on samples of a few thousand respondents or surveys that are conducted on a national level to estimate unemployment rate. We assume simple random sampling and that everybody who was selected for the sample responded and responded truthfully. These assumptions are highly unrealistic.[5]

Yet, the above assumptions make straightforward calculations possible and lead to characteristics that may be used as benchmarks. Usually, a survey is not conducted to estimate a single quantity from the population, rather several quantities and also their relationships. Very often, the information collected in a survey includes the distributions of categorical variables. For example, a respondent may be employed or not employed, and this is a binary variable. Those who are not employed are not necessarily unemployed, rather either are outside of the labor force (retired or in school or homemaker) or unemployed. This is a categorical variable with three categories. Or, if there are three candidates for the office of president, respondents may be classified as those not wishing to vote, or wishing to vote for one of the candidates, leading to a categorical variable with four categories. Or if a survey is to explore what people think about the role of the state or of the fed-

[5] Avoiding self-selection of the respondents and reducing the effects of nonresponse and other kinds of missing data are major problems of survey methodology. Also, much of the published statistical analyses of survey data disregard the peculiarities of the sampling procedure and work as if the sampling distribution was multinomial or Poisson. In reality, most of the nationwide surveys use complex sampling procedures that often include stratification and multistage selection. The goal of applying these procedures may be to reduce the data collection cost per respondent or to incorporate information about the population to reduce the standard deviations of estimates. There are many good books on survey sampling; [41] and [54] are recommended in particular.

eral government, respondents may be confronted with a number of statements (often called items), and whether or not they agree with those statements is recorded. For example, some of the statements may sound like:

> "The state should provide education to all children".
> "The state should provide health care to everybody".
> "The state should provide health care to everybody who cannot pay for themselves".
> "The state should provide old-age pension to everybody who does not have their own retirement funds".

Each of these items is characterized by the fraction of those in the population who agree with it. This fraction, say p_i for item i, is unknown, and its MLE is the relative frequency of the "yes" responses. This estimate, X_i/n, where $X_i \sim \mathscr{B}(p_i, n)$ has variance $p_1(1-p_i)/n$ and standard deviation $\sqrt{p_1(1-p_i)/n}$.[6] The standard deviations may be used to construct confidence intervals for the population quantities. For large sample sizes (certainly for sample sizes in the range of several hundreds of observations and above), the asymptotic distribution of $\sqrt{n}(X_i/n - p_i)$, $N(0, p_i(1-p_i))$ is a good approximation, so the distribution of X_i/n is taken approximately as $N(p_i, p_i(1-p_i)/n)$.

However, the above formulas would require the value of p_i to produce standard errors of its estimates or to produce confidence intervals for its true value. The estimated values of p_i may be used to estimate the standard error of the estimated value or to construct a confidence interval. Very often in practice, one would like to have an overall standard error (or a half-length of a confidence interval with specified coverage) that applies not only to one but to all probabilities estimated from the survey. This is usually constructed by noting that

$$\arg\max\{p(1-p)\,|\,0 \le p \le 1\} = 0.5,$$

and with it $p = 0.5$, the maximum of $p(1-p)$ is 0.25. Using this, one may obtain upper bounds for standard errors that depend on the sample size only. The upper bound for the standard error of p_i with sample size n is $1/(2\sqrt{n})$. For various sample sizes, the following upper bounds of standard errors are obtained:

$$\begin{aligned}
n &= 100,\ \text{upper bound of standard error} : 0.05\\
n &= 400,\ \text{upper bound of standard error} : 0.025\\
n &= 900,\ \text{upper bound of standard error} : 0.0167\\
n &= 1600,\ \text{upper bound of standard error} : 0.0125\\
n &= 2500,\ \text{upper bound of standard error} : 0.01
\end{aligned}$$

[6] Note that some authors distinguish between standard deviation, which is a parameter associated with a random variable, and standard error, which is the same parameter associated with a quantity determined from a sample. Such a strict distinction is not made in this book.

Using the normal approximation, a 95% confidence interval for the true population fraction is obtained as $\hat{p} \pm 1.96\,\mathrm{SE}$. The value of $1.96\,\mathrm{SE}$ is called the margin of error by survey practitioners. It gives an approximate probabilistic bound (margin) for the error.[7] The margin of error is used to illustrate the precision that may be achieved based on the survey.[8]

4.2.1 Standard Errors of More Complex Quantities

The foregoing considerations apply only to the estimates of fractions out of the entire population. For example, the fraction of those who wish to participate in an upcoming election, out of the population of those eligible to vote, is such a quantity. Several relevant parameters, that are routinely reported based on surveys, are of a more complex nature. For example, the results of the election do not depend on what fraction of the eligible voters vote for a candidate rather on the fraction of the votes received, out of all votes, by a candidate.[9] A natural estimate of this quantity is the ratio of the number of those who say they would vote for this candidate, to the number of those who say they would vote (for any candidate). This is the ratio of two random variables, and its behavior is different from the one discussed above.

To illustrate the behavior of some of the relevant parameters, assume that there are two candidates, and out of those eligible to vote, the fraction of those who will not vote is $p_\emptyset$, the fraction of those who will vote for candidate A is p_A, and the fraction of those who will vote for candidate B is p_B; all these fractions are positive and they sum to 1. If a sample of size n is observed, assuming, again, simple random sampling and that everybody gives a response, the vector valued variable $(T_\emptyset, T_A, T_B)'$ has a multinomial distribution. Then, with $\mathbf{X} = \mathbf{T}/n$,

$$Z_1 = \frac{X_A}{X_A + X_B} \tag{4.28}$$

is an estimate for the fraction of votes to be cast on candidate A, further

$$Z_2 = \frac{X_A}{X_A + X_B} - \frac{X_B}{X_A + X_B} \tag{4.29}$$

[7] Care should be taken not to interpret the margin of error as if it was an absolute bound on the magnitude of the possible error of the estimate.

[8] More precisely, the margin of error addresses only the size of the so-called sampling error, that is, the difference between the estimate from the survey and the value that would be obtained if the methods of the survey were used to carry out a census. In a census, data are collected from the entire population; no sampling occurs. In most cases, however, even the value obtained from the census may be different from the true value. For example, respondents may not remember or do not want to tell the truth, or any other kind of measurement error may occur. The difference between the census value and the true value is called the nonsampling error of the survey.

[9] In many countries, a party has to receive at least 5% of the votes that were actually cast to get into the parliament. For this, and other reasons, the number of seats in the parliament may not be a linear function of the fraction of votes received.

is an estimate of the lead candidate A has over candidate B, and if an identical survey was conducted in the previous month with relative frequencies $(Y_\emptyset, Y_A, Y_B)'$, then

$$Z_3 = \left(\frac{X_A}{X_A+X_B} - \frac{X_B}{X_A+X_B}\right) - \left(\frac{Y_A}{Y_A+Y_B} - \frac{Y_B}{Y_A+Y_B}\right) \tag{4.30}$$

is an estimate of the change in the advantage of A over B during the last month.

The δ-method may be applied to obtain asymptotic variances for these estimators, and these are often applied as approximations of the true variances. The relevant covariance matrices are of the structure

$$\mathbf{\Sigma}(\sqrt{n}\mathbf{X}) = \begin{pmatrix} p_\emptyset(1-p_\emptyset) & -p_\emptyset p_A & -p_\emptyset p_B \\ -p_A p_\emptyset & p_A(1-p_A) & -p_A p_B \\ -p_B p_\emptyset & -p_B p_A & p_B(1-p_B) \end{pmatrix} \tag{4.31}$$

and, if the two surveys are assumed to be independent and the probabilities in the previous month are denoted by q, then

$$\mathbf{\Sigma}(\sqrt{n}\mathbf{X}, \sqrt{n}\mathbf{Y}) = \begin{pmatrix} p_\emptyset(1-p_\emptyset) & -p_\emptyset p_A & -p_\emptyset p_B & 0 & 0 & 0 \\ -p_A p_\emptyset & p_A(1-p_A) & -p_A p_B & 0 & 0 & 0 \\ -p_B p_\emptyset & -p_B p_A & p_B(1-p_B) & 0 & 0 & 0 \\ 0 & 0 & 0 & q_\emptyset(1-q_\emptyset) & -q_\emptyset q_A & -q_\emptyset q_B \\ 0 & 0 & 0 & -q_A q_\emptyset & q_A(1-q_A) & -q_A q_B \\ 0 & 0 & 0 & -q_B q_\emptyset & -q_B q_A & q_B(1-q_B) \end{pmatrix}$$

The partial derivative vectors of (4.28), (4.29), (4.30), evaluated at the expectation, are

$$\nabla(Z_1) = \left(0, \frac{p_B}{(p_A+p_B)^2}, \frac{-p_A}{(p_A+p_B)^2}\right)'$$

$$\nabla(Z_2) = \left(0, \frac{2p_B}{(p_A+p_B)^2}, \frac{-2p_A}{(p_A+p_B)^2}\right)'$$

$$\nabla(Z_3) = \left(0, \frac{2p_B}{(p_A+p_B)^2}, \frac{-2p_A}{(p_A+p_B)^2}, 0, \frac{-2q_B}{(q_A+q_B)^2}, \frac{2q_A}{(q_A+q_B)^2}\right)'$$

The asymptotic variances are obtained as

$$V(\sqrt{n}Z_1) = \nabla(Z_1)\mathbf{\Sigma}(\sqrt{n}\mathbf{X})\nabla(Z_1)',$$

$$V(\sqrt{n}Z_2) = \nabla(Z_2)\mathbf{\Sigma}(\sqrt{n}\mathbf{X})\nabla(Z_2)',$$

$$V(\sqrt{n}Z_3) = \nabla(Z_3)\mathbf{\Sigma}(\sqrt{n}\mathbf{X}, \sqrt{n}\mathbf{Y})\nabla(Z_3)',$$

respectively. This gives for $\sqrt{n}Z_1$, as seen in Chap. 3,

$$V(\sqrt{n}Z_1) = \frac{p_A p_B}{(p_A+p_B)^3}.$$

Depending on the p_A and p_B values, the asymptotic standard deviation of $\sqrt{n}Z_1$ may be considerably higher than implied by the formula $1/(2\sqrt{n})$, often used in survey practice as an upper bound. More precisely, this upper bound for the standard deviation of the scaled relative frequency $\sqrt{n}X_A$ is $1/2$. Table 4.1 shows the asymptotic standard deviation of $\sqrt{n}Z_1$ for different values of p_A and p_B. The entries are to be compared to 0.5 and may exceed it by a factor larger than 2.

If the upper bound for the standard deviation of a scaled fraction is not used, rather the precise (but in the current context incorrect) formula of $\sqrt{p_A(1-p_A)}$ is used to obtain the standard error, then for an estimated p_A value of 0.2, this would give 0.4. As Table 4.1 shows, this is the correct value if $p_B = 0.8$, but, depending on the value of p_B, the correct standard deviation may exceed the one given by the incorrect formula, by a factor of 2. Note that the case when the two formulas give the same value is the one when everybody votes: $p_A + p_B = 1$.

Table 4.1 Asymptotic standard errors for $\sqrt{n}\hat{p}_A/(\hat{p}_A + \hat{p}_B)$

$p_A =$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$p_B = 0.1$	1.118	0.861	0.685	0.566	0.481	0.418	0.370	0.331	0.300
$P_B = 0.2$	0.861	0.791	0.693	0.609	0.540	0.484	0.438	0.400	
$P_B = 0.3$	0.685	0.693	0.645	0.591	0.541	0.497	0.458		
$P_B = 0.4$	0.566	0.609	0.591	0.559	0.524	0.490			
$P_B = 0.5$	0.481	0.540	0.541	0.524	0.500				
$P_B = 0.6$	0.418	0.484	0.497	0.490					
$P_B = 0.7$	0.370	0.438	0.458						
$P_B = 0.8$	0.331	0.400							
$P_B = 0.9$	0.300								

4.2.2 Standard Errors Under Stratified Sampling

To obtain standard errors of estimates under stratified sampling, assume that the population of size N consists of c groups with sizes N_i. The groups may be defined by a variable (e.g., men and women) but may also be defined by combinations of variables (e.g., gender, age group, education, etc.). The sample of size n consists of subsamples of sizes $n(i)$, which are simple random samples from the relevant groups. As was described in Sect. 2.2.1, $n(i)$ may be $n_i = nN_i/N$, so that n_i/n is the population fraction but, in general, may be different from it. In this subsection, the goal of the sampling procedure is to reproduce the groups of the population, so $n(i) = n_i$. To refer to group membership, let the observations be indexed by i for groups and j for the observations within groups. Thus X_{ij} is an indicator which is 1, if the j-th observation of the i-th group possesses, and is zero if it does not possess a characteristic of interest. Let p_i be the fraction of those possessing this characteristic in group i, and then $X_{ij} \sim \mathscr{B}(n_i, p_i)$ and X_{ij} and $X_{k,l}$ are independent, if $i \neq k$. Note that here the probabilities p_i are not components of a probability distribution.

The population fraction is

$$p = \frac{1}{N}\sum_{i=1}^{c} N_i p_i = \frac{1}{n}\sum_{i=1}^{c} n_i p_i \tag{4.32}$$

and

$$\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n_i} X_{ij} \tag{4.33}$$

is an unbiased estimator of p. One would expect that in comparison to the relative frequency based on a simple random sample, (4.33) should perform better, because it incorporates information about the population (namely, the sizes of the strata). On the other hand, if the fractions within each group are the same, $p_i = p$, this information becomes irrelevant.

Theorem 4.12. *The variance of the estimator in (4.33) is*

$$\frac{1}{n^2}\sum_{i=1}^{c} n_i p_i(1-p_i)$$

and it is smaller than the variance of the relative frequency estimator based on a simple random sample of size n, except for $p_i = p$, for all i, when the two variances are equal.

Proof. The variance of (4.33) is

$$E\left(\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n_i} X_{ij} - \frac{1}{n}\sum_{i=1}^{c} n_i p_i\right)^2 = \frac{1}{n^2}E\left(\sum_{i=1}^{c}\left(\sum_{j=1}^{n_i} X_{ij} - n_i p_i\right)\right)^2$$

$$= \frac{1}{n^2}\sum_{i=1}^{c} E\left(\sum_{j=1}^{n_i} X_{ij} - n_i p_i\right)^2 = \frac{1}{n^2}\sum_{i=1}^{c} n_i p_i(1-p_i).$$

Indeed, the second equality is implied the fact that the subsamples from different strata are independent, and thus

$$E\left(\left(\sum_{j=1}^{n_i} X_{ij} - n_i p_i\right)\left(\sum_{l=1}^{n_k} X_{kl} - n_k p_k\right)\right) = 0,$$

for all $i \neq k$. To see that

$$\frac{1}{n^2}\sum_{i=1}^{c} n_i p_i(1-p_i) \le \frac{1}{n}p(1-p),$$

write, using (4.32), the two sides as

$$\frac{1}{n^2}\sum_{i=1}^{c} n_i p_i(1-p_i) = \frac{1}{n}\left(\sum_{i=1}^{c}\frac{n_i p_i}{n} - \frac{1}{n}\sum_{i=1}^{c} n_i p_i^2\right) = \frac{1}{n}p - \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{c} n_i p_i^2\right)$$

and

$$\frac{1}{n}p(1-p)=\frac{1}{n}p-\frac{1}{n}p^2.$$

The inequality to be proved is equivalent to

$$\sum_{i=1}^{c}\frac{n_i}{n}p_i^2\geq p^2, \tag{4.34}$$

which is a consequence of the convexity of the square function. To see this, assume first that $c=2$. Write $\alpha_i=n_i/n$, and then $\alpha_1 p_1+\alpha_2 p_2=p$, so that p is a linear combination of p_1 and p_2. The linear combination has positive coefficients that sum to 1; thus, p lies between p_1 and p_2. The right-hand side of (4.34) is the square of the linear combination, and the left hand side is the linear combination of the squares of the p_i values with the same coefficients. Therefore, the left hand side is, indeed, greater than the right-hand side, except if $p_1=p_2$, in which case they are equal. To complete the induction argument, assume (4.34) is proved for up to $c-1$ categories. Write

$$q_{c-1}=\sum_{i=1}^{c-1}\frac{n_i}{\sum_{j=1}^{c-1}n_j}p_i$$

and with this

$$\sum_{i=1}^{c-1}\frac{n_i}{\sum_{j=1}^{c-1}n_j}p_i^2\geq q_{c-1}^2$$

by the induction assumption. Then, using the relationship for q_{c-1} and p_c,

$$\sum_{i=1}^{c}\frac{n_i}{n}p_i^2=\frac{1}{n}\sum_{i=1}^{c-1}n_i p_i^2+\frac{n_c}{n}p_c^2$$

$$=\frac{\sum_{j=1}^{c-1}n_j}{n}\sum_{i=1}^{c-1}\frac{n_i}{\sum_{j=1}^{c-1}n_j}p_i^2+\frac{n_c}{n}p_c^2\geq\frac{\sum_{j=1}^{c-1}n_j}{n}q_{c-1}^2+\frac{n_c}{n}p_c^2\geq p^2.$$

□

The results of the previous subsection may be extended to product multinomial distributions obtained from stratified sampling. This only requires using, instead of (4.31), the product multinomial covariance matrix. The derivation involving two surveys essentially did this, because estimates from two independent samples with multinomial distributions behave exactly as if they were based on the components of a product multinomial sample.

4.3 Things to Do

1. Prove that a uniform variable taking integer values on $[0,u]$ has expectation $u/2$.
2. Assume that someone has a square-shaped metal sheet of size $x \times x$ inches and wants to cut out squares of the size $y \times y$ inches at all four corners of the sheet so that with the remaining sides folded up, a box (without lid) will be created. If x is given, determine y so that the volume of the box is maximal. Solve the problem with and without using the Lagrange multiplier technique.
3. Derive the solution given in (4.21).
4. Compute the likelihood function in the above example for $(q_1, (1-q_1)/2, (1-q_1)/2)'$ with $q_1 = 0.2, 0.1, 0.01, 0.001$ for $\mathbf{x} = (0,\ 12,\ 18)'$.
5. Find the maximum of the likelihood function in (4.22) numerically.
6. Determine the partial derivatives of the likelihood function in (4.22).
7. Prove that under multinomial sampling, the maximum likelihood estimator of the probabilities is also unbiased.
8. When does a stratified sampling procedure lead to unequal selection probabilities?
9. In a survey based on 2500 observations with simple random sampling and no nonresponse, the estimated value of the probability of an event is 0.6. Estimate the standard error. Give an upper bound for the standard error.
10. Using the normal approximation, determine a 95% confidence interval for the true population fraction using information in the previous exercise.
11. Assume that 25% of the eligible voters want to vote for candidate A, and 35% for candidate B, and there are no further candidates. If a survey with sample size 2000 is used to estimate the fraction of votes that would go for candidate A, give an approximate standard error for the observed fraction of votes that would go for A. What additional assumptions need to be made?
12. In the above problem, give an approximation for the standard error of the lead B will show over A.
13. Determine the approximate standard error formula for Z_3.
14. Set up the δ-method for the following problem: There are several candidates for a post, and some of those eligible will vote, and some will not. One is interested in estimating the lead of one candidate over another one. The difference compared to Z_2 is that, in this case, some voters may vote for candidates other than these two.
15. Prove that the estimator in (4.33) is unbiased for the population fraction p.
16. Create a small population with c strata and with p_i, the fraction of those who possess a certain characteristic within stratum i. Apply simple random sampling with sample size n, by selecting every subset of the population of size n with the same probability, and apply stratified sampling similarly. Show that the subsets that are samples under stratified sampling are also samples under simple random sampling. Study the estimates obtained from stratified samples and the estimates that may be obtained under simple random sampling but not under stratified sampling.

17. Assume that a sample is stratified by the gender of the respondent and that the sampling distribution is the relevant product multinomial. Determine the asymptotic variance of the difference in the fraction of supporters of a political party between men and women.
18. Assume, just like in the item above, that the sample is stratified according to gender. Determine the asymptotic variance of the fraction of those who would vote for a party among those who say they would vote for any party.

Chapter 5
Basic Testing for Categorical Data

Abstract This chapter first discusses tests of hypotheses pertaining to the probability of a binomial distribution and gives a brief review of the fundamental concepts of tests of hypotheses. Exact, randomized, and asymptotic tests are considered. The Pearson chi-squared and likelihood ratio statistics are introduced for tests of fit. The concept of independence, which will play a central role in later chapters of the book, is also introduced for the simple case of two-way tables and tests of independence are discussed.

The chapter starts with considering the simplest testing problem with categorical data. The discussion includes a review of many of the fundamental concepts of hypothesis testing in general.

5.1 Tests of $p = p_0$

For a binomial variable, tests of the hypothesis $p = p_0$ will be presented in this section. Testing this hypothesis is simple, but the discussion gives an opportunity to present, exact, randomized, and asymptotic testing, which will then be extended to more complex cases.

The hypothesis $p = p_0$ for a fixed value of p_0, when applied to $X \sim \mathscr{B}(n, p)$ is simple, because it determines the distribution of X entirely. Under the hypothesis, any possible sample has a specified probability. A testing procedure is based on dividing the sample space[1] into a region of rejection R (also called critical region)

[1] In the case of the binomial distribution, the sample space, strictly speaking, is the set of yes-no sequences of length n. It has been seen earlier that the probability of a sequence depends only on the number of yes responses in the sequence. Therefore, it would not be reasonable to take into account in the testing procedure the order in which the yes and no responses occurred. The testing procedure will be based on the number of yes responses only and therefore the sample space is identified with the set integers from 0 to n.

and a region of acceptance A.[2] The hypothesis is rejected if and only if $X \in R$. The probability of error type I of this procedure is the probability of $X \in R$, if this probability is determined assuming the hypothesis holds.

What has been said so far would allow one to choose any subset of the sample space as the region of rejection. For example, with $X \sim \mathscr{B}(15, p)$ and hypothesis $p = 0.6$, the sample space is $\{0, 1 \ldots, 15\}$ and one may choose $R = \{12\}$. The probability of error type I is $P(X = 12)$ for $X \sim \mathscr{B}(15, 0.6)$, which is about 0.06.

What is strange with this procedure is that one rejects the hypothesis, when $x = 12$ is observed, but not when $x = 13$ or $x = 14$, although these values are less plausible observations under the hypothesis than $x = 12$. The plausibility of observations depends on their likelihoods but comparison is to other observations from the same distribution.[3] Observing, say, 13, provides one with more evidence against the hypothesis than observing 12. This is not directly because 13 is farther form the expectation under the hypothesis than 12 is, rather because observing 13, if the hypothesis is true, is less likely than observing 12. Of course, in the present case, the former fact (being farther from the expectation) implies the latter fact (being less likely), but this is not necessarily the case with other distributions that might be assumed for X. It is thus reasonable to define R as the collection of the least plausible observations under the hypothesis:

$$R_a = \{x_i \in S : P(x_i \,|\, p = p_0) \leq a\},$$

where S is the sample space of X. Then, the probability of error type I is

$$\alpha_a = \sum_{x_i \in R_a} P(x_i \,|\, p = p_0).$$

In the case of $X \sim \mathscr{B}(15, 0.6)$, the possible observations ordered according to their probabilities under the hypothesis are

[2] Region of acceptance is a misleading but commonly used name. When the observation falls in A, the hypothesis is not "accepted," rather it is "not rejected." The procedure applied is not appropriate to provide evidence that the hypothesis is true: even if the sample belongs to the group of most likely observations, it may have been generated by a population entirely different from what is described in the hypothesis. On the other hand, the procedure is appropriate to provide evidence against the hypothesis: this happens when the sample belongs to the group of least likely ones under the hypothesis.

[3] The observation with the maximum plausibility is the one that has the highest likelihood over the sample space when the parameter is kept constant. Based on this concept, a principle of estimation may be defined: choose, from among the possible parameter values, the one under which the current data would be the most plausible (when compared to other data under the same parameter). The principle of maximum likelihood is different: choose, from among the possible parameter values, the one under which the current data would be the most likely (when compared to the same data under different parameters). It is easy to see that in the current example, the estimates of p would be the same.

P(X=0)=1.07E-06, P(X=1)=2.42E-05, P(X=2)=2.54E-04,
P(X=15)=4.70E-04, P(X=3)=1.65E-03, P(X=14)=4.70E-03,
P(X=4)=7.42E-03, P(X=13)=2.19E-02, P(X=5)=2.45E-02,
P(X=6)=6.12E-02, P(X=12)=6.34E-02, P(X=7)=1.18E-01,
P(X=11)=1.27E-01, P(X=8)=1.77E-01, P(X=10)=1.86E-01,
P(X=9)=2.07E-01.

Here, 1.07E$-$06 means $1.07 \cdot 10^{-6}$. For example, $R_{0.0245} = \{0, 1, 2, 15, 3, 14, 4, 13, 5\}$ and $\alpha_{0,245} = 0.061$. This is an exact test, because it is based on an exact calculation of the relevant probabilities. This method, however, does not make it possible to design tests with arbitrary type I error probabilities. For example, if a test with type I error probability equal to 0.05 is required, there is no appropriate choice of a. With $a \geq 0.0245$, $\alpha_a \geq 0.06$, and with $a < 0.0245$, $\alpha_a \leq 0.04$ (the last value is obtained as the sum of the 8 smallest probabilities).

To design a test with an arbitrary α, one may use randomization. A randomized test is defined by dividing the sample space into subsets R, A, and Q, and the hypothesis is rejected if the observation is in R, and it is not rejected, if the observation is in A. When the observation is in Q, the hypothesis is rejected with probability q and is not rejected with probability $1-q$. Indeed, q may be determined so that the probability of type I error is equal to the desired value. In the case of a randomized test, the probability of error type I is

$$\alpha = P(X \in R) + qP(X \in Q),$$

where the probabilities are understood under the assumption that the hypothesis is true. Then

$$q = \frac{\alpha - P(X \in R)}{P(X \in Q)}.$$

In the case of the above example, if $R = \{0, 1, 2, 15, 3, 14, 4, 13\}$, $Q = \{5\}$ and $q = (0.05 - 0.036)/0.025 = 0.56$ yields a test with $\alpha = 0.05$. That is, the hypothesis of $p = 0.6$ is rejected if the observation is in R and is rejected in about every other case (with probability 0.56) when the observation is equal to 5.

When, in addition to the hypothesis (that is called null hypothesis, or simply the null, in this situation) $p = p_0$, also an alternative hypothesis is defined, the power of the procedure is also of interest. The power is 1 minus the probability that the error of type II is committed, that is, 1 minus the probability that the alternative is true but the null is not rejected. When the alternative is composite, that is, contains many distributions, probabilities, including the one relevant for the power, cannot be determined under the alternative. In such cases, instead of probabilities under the alternative, the supremum of such probabilities is considered.

When the alternative is that the null is not true, that is, that $p \neq p_0$, the power of the (nonrandomized) test for $p = p_0$ is

$$1-\sup_p P(X\in A\,|\,p\neq p_0)=1-\sup_p(1-P(X\in R\,|\,p\neq p_0))=$$
$$\inf_p P(X\in R\,|\,p\neq p_0)=P(X\in R\,|\,p=p_0), \qquad (5.1)$$

where the last equality is implied by the continuity of the binomial probability as a function of p. This means that the power is equal to the probability of type I error. If the probability of type I error is chosen to be "small," the power of the procedure will be "small," too. Note that the power in this case, after the probability of error type I was determined, does not depend on the sample size, so decisions based on larger samples are not going to be better, that is, giving higher power, than decisions based on smaller samples.

When the null and the alternative are separated, one obtains better power, and the power increases with increasing sample size and with increasing separation. If the hypotheses are set up as

$$H_0: p-p_0=0 \qquad (5.2)$$

and

$$H_1: |p-p_0|\geq b, \qquad (5.3)$$

for some positive p_0 and $0<b<\min(p_0, 1-p_0)$, then the power[4] is

$$1-\sup_p P(X\in A\,|\,p\in H_1)=$$
$$1-\max\left(P(X\in A\,|\,p=p_0+b), P(X\in A\,|\,p=p_0-b)\right) \qquad (5.4)$$

When $p_0=0.5$, the two probabilities above are equal because of the symmetry of the binomial distribution. In that case, the power is $1-P(X\in A\,|\,p=0.5+b)$.

For example, with $p_0=0.5$ and $n=15$, the test with $R=\{0,1,2,3,12,13,14,15\}$ has level 0.965, so $\alpha=0.035$. When the alternative is defined with $b=0.1$, the power is 0.093, with $b=0.2$, the power is 0.297, and with $b=0.3$, the power is 0.648. If $n=30$, $R=\{0,1,2,3,4,5,6,7,8,9,21,22,23,24,25,26,27,28,29,30\}$ defines a test with level 0.957, so $\alpha=0.043$, that is, the error type I probability is slightly higher than in the previous case. The powers for the b values 0.1, 0.2, and 0.3 are 0.176, 0.589, and 0.939, respectively, much higher than those for the smaller sample size. These examples are based on exact calculation, but for large sample sizes, the normal approximation of the binomial distribution may also be used to obtain similar results. Using that approximation, asymptotic tests may also be constructed. The increasing power of the test for larger sample sizes follows from the fact that the variance $V(X/n)=p(1-p)/n$ decreases with increasing sample sizes.

The normal approximation may also be used to generate asymptotic tests as follows. When the distribution of a $\mathscr{B}(n,p)$ variable is approximated by the distribution of a $N(np,np(1-p))$ variable, the region of acceptance to test (5.2) against (5.3) may be selected to be (u_1,u_2) such that

$$P(u_1\leq Y_0\leq u_2)=1-\alpha,$$

[4] The power of the test is the same whether or not H_1 is defined by allowing equality.

if the desired type I error probability is α, where $Y_0 \sim N(np_0, np_0(1-p_0))$. Obviously, there are infinitely many choices of u_1 and u_2. The power of this test, using (5.4), is

$$1 - \max\left(P(u_1 \le Y_{+b} \le u_2), P(u_1 \le Y_{-b} \le u_2)\right),$$

where $Y_{+b} \sim N(n(p_0+b), n(p_0+b)(1-(p_0+b)))$ and $Y_{-b} \sim N(n(p_0-b), n(p_0-b)(1-(p_0-b)))$. With given level, the power is maximized out of the infinitely many (u_1, u_2) intervals that have probability $1-\alpha$ according to the null hypothesis that is according to Y_0, when the one with minimum of the larger of the probabilities under Y_{+b} and Y_{-b} is selected.

5.2 Basic Properties of the Pearson and the Likelihood Ratio Statistics

Another idea of testing the hypothesis $p = p_0$ is the following. Under the hypothesis, one expects np_0 'yes' and $n(1-p_0)$ "no" responses. So if x "yes" and $n-x$ "no" responses were observed, $(x-np_0)+((n-x)-n(1-p_0))$ is the sum of deviations from the expected number of responses. Unfortunately, this quantity is always zero, and no test can be based on it. An alternative is to add the squares of the deviations: $(x-np_0)^2+((n-x)-n(1-p_0))^2$ which is always nonnegative, and larger deviations lead to larger values of this expression. This quantity, however, requires one more modification before it becomes a good measure of the magnitude of deviation from what was expected under the hypothesis. When $n=10$, $p=0.3$, and $x=5$, the value is the same as when $n=50$, $p=0.6$, and $x=32$. In both cases, the quantity is 8, because in both cases the deviation is 2, disregarding the fact that the 2 extra observations are there when 3 were expected in the first case and 30 in the second one. To make up for this difference, the squared deviations will be compared to the expectations:

$$\frac{(x-np_0)^2}{np_0} + \frac{((n-x)-n(1-p_0))^2}{n(1-p_0)}. \tag{5.5}$$

This quantity is called the Pearson chi-squared statistic. It can be rewritten as follows:

$$\frac{(1-p_0)(x-np_0)^2 + p_0((n-x)-n(1-p_0))^2}{np_0(1-p_0)} =$$

$$\frac{(1-p_0)(x-np_0)^2 + p_0(np_0-x)^2}{np_0(1-p_0)} =$$

$$\frac{(x-np_0)^2}{np_0(1-p_0)} = \left(\frac{x-np_0}{\sqrt{np_0(1-p_0)}}\right)^2.$$

Before the data were observed, the Pearson chi-squared statistic is a random variable that is equal to

$$\left(\frac{X - np_0}{\sqrt{np_0(1-p_0)}}\right)^2,$$

and the asymptotic distribution of this variable is equal to that of the square of a standard normal (see also Theorem 3.2), if the hypothesis holds. This is summarized in the next proposition.

Proposition 5.1. *If the hypothesis $p = p_0$ holds, the asymptotic distribution of the Pearson chi-squared statistic (5.5) is that of the square of a standard normal variable.* □

The distribution of the square of a standard normal variable is called a chi-squared distribution with 1 degree of freedom (df) and is denoted as $\mathscr{C}(1)$.

Theorem 5.1. *If $X \sim \mathscr{C}(1)$, then its density function is*

$$f(x) = \frac{1}{\sqrt{2\pi x}} \exp(-\frac{1}{2}x), \textit{ for } x > 0,$$

and the density is zero elsewhere.

Proof. If $Y \sim \mathscr{N}(0,1)$, then

$$P(Y^2 \leq x) = P(Y \leq \sqrt{x}) - P(Y \leq -\sqrt{x}) = F_Y(\sqrt{x}) - F_Y(-\sqrt{x}),$$

where F_Y is the distribution function of Y. Then, the density function of Y^2 is obtained by derivation as

$$\begin{aligned} f_{Y^2} &= \frac{1}{2}\frac{1}{\sqrt{x}} f_Y(\sqrt{x}) + \frac{1}{2}\frac{1}{\sqrt{x}} f_Y(\sqrt{x}) \\ &= \frac{1}{\sqrt{x}} f_Y(\sqrt{x}) = \frac{1}{\sqrt{2\pi x}} \exp(-\frac{1}{2}x). \end{aligned}$$

□

Proposition 5.2. *If $X \sim \mathscr{C}(1)$, then $E(X^2) = 3$.*

Proof. Using the density function from Theorem 5.1,

$$E(X^2) = \int_0^\infty x^2 \frac{1}{\sqrt{2\pi x}} \exp(-\frac{1}{2}x)dx = \frac{1}{\sqrt{2\pi}} \int_0^\infty x^{3/2} \exp(-\frac{1}{2}x)dx$$

The integral is evaluated by integration by parts, using that the derivative of

$$-2x^{3/2} \exp(-\frac{1}{2}x)$$

is

$$-3x^{1/2} \exp(-\frac{1}{2}x) + x^{3/2} \exp(-\frac{1}{2}x).$$

Thus,

$$E(X^2) = \frac{1}{\sqrt{2\pi}} \left(\left[-2x^{3/2} \exp(-\frac{1}{2}x) \right]_0^\infty + \int_0^\infty 3x^{1/2} \exp(-\frac{1}{2}x) dx \right).$$

The first term between the parentheses on the right-hand side is zero, because it is zero at $x = 0$, and for $x \to \infty$, it converges to zero. To see this, consider

$$g(x) = \frac{-2x^{3/2}}{\exp(\frac{1}{2}x)},$$

and then

$$\frac{g(x+1)}{g(x)} = \frac{-2(x+1)^{3/2}}{\exp(\frac{1}{2}(x+1))} \frac{\exp(\frac{1}{2}x)}{-2x^{3/2}} = \left(1 + \frac{1}{x}\right)^{3/2} \frac{1}{\exp(\frac{1}{2})}.$$

This converges to

$$\frac{1}{\exp(\frac{1}{2})} < 1,$$

implying, that for large enough x, $g(x+1) < ag(x)$, for a positive constant a, which is less than 1. Therefore, $g(x)$ cannot remain above any positive constant and thus converges to zero. Therefore,

$$E(X^2) = \frac{3}{\sqrt{2\pi}} \int_0^\infty x^{1/2} \exp(-\frac{1}{2}x) dx.$$

Another integration by parts uses that the derivative of

$$-2x^{1/2} \exp(-\frac{1}{2}x)$$

is

$$-x^{-1/2} \exp(-\frac{1}{2}x) + x^{1/2} \exp(-\frac{1}{2}x).$$

Therefore,

$$E(X^2) = \frac{3}{\sqrt{2\pi}} \left(\left[-2x^{1/2} \exp(-\frac{1}{2}x) \right]_0^\infty + \int_0^\infty x^{-1/2} \exp(-\frac{1}{2}x) dx \right).$$

By an argument similar to the one used after the first integration by parts, the first term on the right and side is zero, so

$$E(X^2) = \frac{3}{\sqrt{2\pi}} \int_0^\infty x^{-1/2} \exp(-\frac{1}{2}x) dx = 3 \int_0^\infty \frac{1}{\sqrt{2\pi}} x^{-1/2} \exp(-\frac{1}{2}x) dx.$$

Finally, because the integrand is the density function in Theorem 5.1,

$$E(X^2) = 3.$$

□

The expectation and variance of a variable with distribution $\mathscr{C}(1)$ is given in the next result.

Theorem 5.2. *If $X \sim \mathscr{C}(1)$, then $E(X) = 1$ and $V(X) = 2$.*

Proof. If Y is a standard normal variable, then $1 = Var(Y) = E(Y^2) - (E(Y))^2 = E(Y^2)$, so $E(X) = 1$. The variance of X, using Proposition 5.2, is $V(X) = E(X^2) - (E(X))^2 = 3 - 1$. □

A test of the hypothesis $p = p_0$, based on the Pearson chi-squared statistic, rejects the hypothesis if the observed value of (5.5) exceeds the relevant percentage point of the chi-squared distribution on 1 df.

In general, if $Y_i \sim N(0,1)$ are independent for $i = 1, \ldots, k$, then the distribution of

$$Y_1^2 + Y_2^2 + \ldots + Y_k^2$$

is called chi-squared distribution on k df and is denoted as $\mathscr{C}(k)$. The definition implies the following property.

Proposition 5.3. *If $X \sim \mathscr{C}(k)$ and $Y \sim \mathscr{C}(l)$ are independent, then*

$$X + Y \sim \mathscr{C}(k + l).$$

□

Similarly, the definition of the chi-squared distribution and Theorem 5.2 imply immediately the following:

Theorem 5.3. *If $X \sim \mathscr{C}(k)$, then $E(X) = k$ and $V(X) = 2k$.* □

An advantage of the test based on the Pearson chi-squared statistic is that it extends in a straightforward way to the multinomial distribution. To simplify notation, the index 0 is dropped and $\mathbf{p}$ will denote from now on the hypothetical true probability distribution. If $\mathbf{X} \sim \mathscr{M}(n, \mathbf{p})$ has k categories, then the statistic takes the form

$$C = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}. \tag{5.6}$$

The next result establishes the asymptotic distribution of this statistic, if the hypothesis holds.

Theorem 5.4. *If $\mathbf{X} \sim \mathscr{M}(n, \mathbf{p})$ and it has k categories, then the Pearson chi-squared statistic (5.6) is asymptotically distributed as $\mathscr{C}(k-1)$.*

Proof. The proof is by induction on k. The claim has been proved for $k = 2$, see Proposition 5.1. Suppose it has been proved for $k - 1$ categories. The idea of the induction step is to combine two categories, say the last ones, and then to consider two statistics

$$T = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} \ldots + \frac{(X_{k-1} + X_k - n(p_{k-1} + p_k))^2}{n(p_{k-1} + p_k)} \tag{5.7}$$

and

$$U = \frac{(X_{k-1} - (X_{k-1} + X_k)\frac{p_{k-1}}{p_{k-1}+p_k})^2}{np_{k-1}} + \frac{(X_k - (X_{k-1} + X_k)\frac{p_k}{p_{k-1}+p_k})^2}{np_k}. \tag{5.8}$$

First we show that

$$T + U = C.$$

Indeed, with

$$\beta_{k-1} = \frac{X_{k-1}}{np_{k-1}},\ \beta_k = \frac{X_k}{np_k}$$

and

$$\beta = \frac{X_{k-1} + X_k}{n(p_{k-1} + p_k)},$$

$$T + U - C = \frac{(X_{k-1} + X_k - n(p_{k-1} + p_k))^2}{n(p_{k-1} + p_k)}$$
$$+ \frac{(X_{k-1} - (X_{k-1} + X_k)\frac{p_{k-1}}{p_{k-1}+p_k})^2}{np_{k-1}} + \frac{(X_k - (X_{k-1} + X_k)\frac{p_k}{p_{k-1}+p_k})^2}{np_k}$$
$$- \frac{(X_{k-1} - np_{k-1})^2}{np_{k-1}} - \frac{(X_k - np_k)^2}{np_k} =$$
$$n(p_{k-1} + p_k)(\beta - 1)^2 + np_{k-1}(\beta_{k-1} - \beta)^2$$
$$+ np_k(\beta_k - \beta)^2 - np_{k-1}(\beta_{k-1} - 1)^2 - np_k(\beta_k - 1)^2 =$$
$$n(p_{k-1}((\beta - 1)^2 + (\beta_{k-1} - \beta)^2 - (\beta_{k-1} - 1)^2) + p_k((\beta - 1)^2$$
$$+ (\beta_k - \beta)^2 - (\beta_k - 1)^2)).$$

That is,

$$T + U - C = 2n((p_{k-1} + p_k)(\beta^2 - \beta) + p_{k-1}(\beta_{k-1} - \beta\beta_{k-1}) + p_k(\beta_k - \beta\beta_k)).$$

Because of the definitions of the β terms,

$$\beta = \frac{p_{k-1}\beta_{k-1} + p_k\beta_k}{p_{k-1} + p_k}$$

and thus

$$T + U - C = 0.$$

Therefore, the asymptotic distribution of $T + U$ needs to be determined to obtain the asymptotic distribution of C. The asymptotic distribution of T, by the induction assumption, is $\mathscr{C}(k-1)$. To determine the asymptotic distributions of U, write

$$q_{k-1} = \frac{p_{k-1}}{p_{k-1} + p_k},\quad X = X_{k-1} + X_k$$

and then

$$U = \frac{(X_{k-1} - Xq_{k-1})^2}{(X/\beta)q_{k-1}} + \frac{(X - X_{k-1} - X(1 - q_{k-1}))^2}{(X/\beta)(1 - q_{k-1})}$$
$$= \beta \left[\frac{X_{k-1} - Xq_{k-1}}{\sqrt{Xq_{k-1}(1 - q_{k-1})}} \right]^2,$$

using the same transformation as after (5.5).

Therefore,[5]

$$\log(U) = \log\beta + \log \left[\frac{X_{k-1} - Xq_{k-1}}{\sqrt{Xq_{k-1}(1 - q_{k-1})}} \right]^2. \tag{5.9}$$

If β is fixed, that determines the value of X and the conditional distribution of X_{k-1} is $\mathscr{B}(X, q_{k-1})$. Then, the conditional distribution of the second term in (5.9) is asymptotically the distribution of the logarithm of the square of a standard normal, see Theorem 3.2. This asymptotic distribution does not depend on the condition; therefore, it is also the unconditional asymptotic distribution.

By its definition, β is a binomial variable divided by its own expectation. It is implied by Theorem 3.2 that

$$\sqrt{n}\frac{(X_{k-1} + X_k)/(p_{k-1} + p_k) - n}{n}$$

converges in distribution to a normal with zero expectation and then by Proposition 3.4,

$$\frac{(X_{k-1} + X_k)/(p_{k-1} + p_k)}{n} - 1$$

converges in probability to zero. Therefore, $log(\beta)$ converges in probability to zero. But then, by Theorem 3.4, U converges in distribution to the square of a standard normal.

Finally we recall that the asymptotic distribution of U is the same as that of the second term on the right-hand side of (5.9), and this was shown to be independent of $X_{k-1} + X_k$ and therefore is independent of the distribution of T, so Proposition 5.3 may be applied and that completes the induction step and the proof. □

A further use of the chi-squared statistic will be considered in the last section of this chapter, and Chap. 13 also discusses briefly tests based on the chi-squared statistic and on other approaches.

A seemingly unrelated idea to test the fit of a statistical model uses the comparison of two maximized likelihoods. One is maximized without restriction on p, and the other one is maximized under the model (or hypothesis). The unrestricted likelihood is maximized by choosing $\hat{p} = x/n$. If $X \sim \mathscr{B}(n, p)$ and the hypothesis

[5] U is nonnegative and may be zero with zero probability. Thus, its logarithm is defined with probability 1 and that $\log U$ is undefined with probability zero does not effect its asymptotic behavior.

is that $p = p_0$, then, because the hypothesis is simple, the maximization under the hypothesis is straightforward. This leads to the following test statistic:

$$\frac{\max_p Cp^x(1-p)^{n-x}}{\max_{p\in H} Cp^x(1-p)^{n-x}},$$

where C is the constant (with respect to p) of the binomial distribution. This is equal to

$$\frac{(x/n)^x((n-x)/n)^{n-x}}{p_0^x(1-p_0)^{n-x}},$$

which is not less than 1, so its logarithm is nonnegative. In fact, the test statistic

$$2((x\ln(x/n)+(n-x)\ln((n-x)/n))-(x\ln p_0+(n-x)\ln(1-p_0))),$$

when the hypothesis holds is asymptotically distributed as chi-square with 1 df.

When $\mathbf{X} \sim \mathscr{M}_k(n,\mathbf{p})$ and the hypothesis is that $\mathbf{p} = \mathbf{p}_0$, let $L(\mathbf{x},\mathbf{p})$ denote the kernel of the log-likelihood. Then the likelihood ratio statistic to test the hypothesis is

$$2(L(\mathbf{x},\mathbf{x}/n)-L(\mathbf{x},\mathbf{p})) = 2\sum_{i=1}^{k} x_i log\frac{x_i/n}{p_i}. \tag{5.10}$$

Theorem 5.5. *If $\boldsymbol{X} \sim \mathscr{M}(n,\boldsymbol{p})$, and it has k categories, then the asymptotic distribution of (5.10) is chi-squared with $k-1$ degrees of freedom.*

Proof. The sketch of the proof is as follows. It will be shown that, asymptotically, the value of the likelihood ratio statistic is equal to the value of the Pearson statistic. The second-order Taylor expansion of (5.10) with variables $\mathbf{x}/n$ around $\mathbf{p}$ is used to show this. To obtain this expansion, note that the derivative of a member of the sum in (5.10), according to x_i/n, is

$$log\frac{x_i/n}{p_i}+1,$$

and this evaluated at $x_i/n = p_i$ is 1. Then, the second derivative is

$$\frac{1}{x_i/n}$$

and this evaluated at $x_i/n = p_i$ is $1/p_i$. Then, the Taylor expansion is

$$2n\sum_{i=1}^{k}\frac{x_i}{n}log\frac{x_i/n}{p_i} = 2n\sum_{i=1}^{k}0+(x_i/n-p_i)+\frac{1}{2}\frac{1}{p_i}(x_i/n-p_i)^2+O_{i,n}((x_i/n-p_i)^3),$$

where $O_{i,n}((x_i/n-p_i)^3)$ is such that there exist constants C_i, so that

$$\frac{O_{i,n}((x_i/n-p_i)^3)}{(x_i/n-p_i)^3}\to C_i, \tag{5.11}$$

if $n \to \infty$.

Given that

$$2n\sum_{i=1}^{k}(x_i/n-p_i)=0,$$

except for the term

$$2n\sum_{i=1}^{k}O_{i,n}(x_i/n-p_i)^3, \tag{5.12}$$

the Pearson statistic and the likelihood ratio statistic are equal.

Taking into account now that x_i/n is the observed value of a random variable, if one took, for $n \to \infty$, a series of multinomial random variables and would consider the series of Pearson statistics and likelihood ratio statistics, as random variables, one would obtain that they differ only in the series consisting terms in (5.12), and the latter would also be random variables. If one could show that these converge to zero in probability, then Theorem 3.4 would imply that the two statistics have the same asymptotic distribution.

In the random version, each term in (5.12) may be approximated asymptotically as twice

$$\sqrt{n}(X_i/n-p_i)\sqrt{n}(X_i/n-p_i)C_i(X_i/n-p_i).$$

Out of the three terms, the first two do have limiting distributions, thus are bounded in probability, and the last one converges to zero in probability. Therefore, (5.12) converges to zero in probability. □

In more general cases, when the hypothesis of interest is not simple, the maximum likelihood estimates under the model are used instead of $\mathbf{p}_0$ both in the Pearson chi-squared and in the likelihood ratio statistics. The number of degrees of freedom depends, in general, on the size of the table and the structure of the models. A very important case when these statistics are being used is described in the last section of this chapter, and further comments on these statistics are given in Chap. 13.

5.3 Test of $p \le p_0$

For a binomial variable, the observed relative frequency x/n is either in the hypothesis[6] of $H_0: p \le p_0$, and in this case the MLE is $\hat{p} = x/n$, or the relative frequency falls outside of the hypothesis, so that $x/n > p_0$. Of course, the likelihood function is the same as for the hypothesis $p = p_0$, but one wants its maximum not over $(0,1)$, rather over $(0,p_0)$. It was seen in the argument leading to Proposition 4.1 that the derivative of the kernel of the log-likelihood is

$$\frac{x}{p}-\frac{n-x}{1-p}=\frac{(1-p)x-p(n-x)}{p(1-p)}=\frac{x-pn}{p(1-p)},$$

[6] The following considerations apply with minimal changes if the hypothesis is defined as $p < p_0$. Usually, there is little difference between these hypotheses in applications.

which applies if $0 < p \leq p_0$ and is zero elsewhere. It is easily seen that this is positive if and only if $p < x/n$ and is negative if and only if $p > x/n$. Therefore, the log-likelihood function monotone increases from $p = 0$ to $p = x/n$ and monotone decreases from $p = x/n$ to $p = 1$. When $p_0 < x/n$, the kernel of the log-likelihood on $(0, p_0]$ is maximal at $p = p_0$. These result are summarized in the next Proposition.

Proposition 5.4. *Under the hypothesis $p \leq p_0$, the MLE of the probability of a binomial distribution is*

$$\hat{p} = min(p_0, x/n).$$

□

If one wishes to apply the Pearson chi-squared statistic to test the hypothesis, the behavior of the statistic is markedly different if $p < p_0$ or if $p = p_0$ or if $p > p_0$.

As it was shown after Proposition 3.4, X/n converges in probability to p, so if $p < p_0$, the probability that $X/n < p_0$ converges to 1, as $n \to \infty$; thus, the probability that the MLE is X/n converges to 1, because the MLE would be different only if $X/n > p_0$ occurred and the probability of this latter event converges to zero. But if the actual MLE is x/n, then the value of the Pearson chi-squared statistic is 0, and this happens with probability converging to 1, as $n \to \infty$.

When $p > p_0$, X/n converges in probability to p, so the MLE $\hat{p}$ converges in probability to p_0, because whenever $X/n > p_0$, $\hat{p} = p_0$. Therefore, the value of the Pearson statistic divided by n converges in probability to

$$\frac{(p - p_0)^2}{p_0} + \frac{(p - p_0)^2}{(1 - p_0)},$$

where p is the true probability, and thus the Pearson statistic converges in probability to infinity, that is, its value exceeds any threshold with arbitrarily large probability as $n \to \infty$.

When $p = p_0$, $(X - np_0)/\sqrt{np_0(1 - p_0)}$ converges in distribution to a standard normal variable. Therefore, asymptotically, $X/n < p_0$ occurs with probability 0.5 and $X/n > p_0$ also occurs with probability 0.5. In the former case the MLE is x/n and the value of the Pearson chi-squared statistic is zero. In the latter case, the MLE is p_0 and the Pearson chi-squared statistic has asymptotic chi-squared distribution with 1 df. In this case, the Pearson chi-squared statistic

$$T = \frac{(X - n\hat{p})^2}{n\hat{p}} + \frac{((n - X) - (n - n\hat{p}))^2}{n - n\hat{p}} \tag{5.13}$$

is, asymptotically,

$$T = 0.5Z + 0.5Y,$$

where $Z = 0$ and $Y \sim \mathscr{C}(1)$. If c_α is a critical value of the $\mathscr{C}(1)$ distribution so that

$$P(Y > c_\alpha) = \alpha,$$

then

$$P(T > c_{2\alpha}) = 0.5P(Z > c_{2\alpha}) + 0.5(Y > c_{2\alpha}) = 0 + 0.5 \cdot 2\alpha = \alpha,$$

so the critical value of the Pearson chi-squared statistic in the case of $p = p_0$ that belongs to an error type I probability of α is the critical value of a $\mathscr{C}(1)$ distribution belonging to 2α.

Of course, in a usual testing situation, one does not know whether $p < p_0$ or $p = p_0$ and only assumes that $p \le p_0$ to assess the characteristics of the testing procedure under the hypothesis. If $p < p_0$, then asymptotically $P(T > c_{2\alpha}) = 0$ and when $p = p_0$, asymptotically, $P(T > c_{2\alpha}) = \alpha$; thus, asymptotically,

$$P(T > c_{2\alpha} | p \le p_0) \le \alpha. \tag{5.14}$$

In general, when the hypothesis is composite, the probability of its wrong rejection (i.e., rejection, when true) depends on which distribution in the hypothesis is the true distribution, and one usually wants to control the smallest upper bound of the error type I probabilities. The upper bound α given in (5.14) is minimal, because in the case of $p = p_0$, equality holds.[7]

5.4 Test of Independence in Two-Way Tables

Most of the scientific problems or policy analyses where statistical methods (i.e., methods to generalize from the observation in a sample to the underlying population) are used involve several variables. For example, although unemployment rate in itself is an important indicator, to design policies to reduce the unemployment rate, one has to understand whether the unemployment is equally high among men and women, or among the old and the young, among the well and the lesser educated. Whether it is about the same in all regions of the country, and if there are ethnic groups, the members of which are more hardly hit by unemployment than members of other ethnic groups. It may be even more useful to identify groups, defined by combinations of these factors, with particularly high unemployment rates, if such groups exist. For example, it may be the case that young females belonging to an ethnic minority, with low levels of education have high levels of unemployment in all regions of the country. Then, if a program is designed to help these people to find jobs, that program should address the special needs of this demographic group and should be available in all regions of the country.

One has to be careful, however, not to jump too quickly to unwarranted conclusions. It would be wrong to think that these people are unemployed in large fractions because they are young or female or not well educated. It is quite possible, that other common characteristics make it difficult for them to get a job. For example, many of them may be single mothers. This fact may make it difficult for them to find a job, but it is even more likely that because of their lack of experience and lack of education, they could not earn enough to pay for child care while they work. In

[7] When the hypothesis is $p < p_0$, a p arbitrarily close to p_0 is in the hypothesis, and for large but in practice still finite sample sizes, one still has about one half probability of observing a relative frequency above p_0.

this case, neither the low level of education (which was listed among the original variables) nor being a single mother (which did not appear among the original variables), rather their interaction (joint effect) leads to higher unemployment rates. But it is also possible that being a single mother and having low levels of education are both related to certain behavioral patterns that these young women experience while growing up and which they follow in their lives. Or, it is also possible that all characteristics discussed so far are typical for members of certain social groups, who are disadvantaged, compared to members of other groups because of the way advantages and disadvantages are distributed within the society. In this case, members of this group, identified by gender, age, educational level, and ethnicity, do not have a single main factor behind being unemployed; rather, there is further inhomogeneity, which the variables used in the analysis do not capture.

In general, it is a very challenging task to find reasons or causes behind observed phenomena (like unemployment and its differing rates among demographic groups), even though this would be of prime importance both for scientific understanding and policy design. Although studying causality is not an aim of this book, several techniques and models will be developed that are relevant in that context, too, in Sect. 8.4. A first step is to analyze situations when the phenomenon measured (or characterized) by one variable is unrelated to the phenomenon measured (or characterized) by another variable or, briefly, when one variable is unrelated to another one. If this is the case, they cannot be causes of each other.

5.4.1 The Concept of Independence

The concept of being unrelated needs, of course, a precise definition. The most widely used definition captures the idea that the two variables are unrelated if knowing in which category of the first one a member of the population belongs does not affect it’s chances of being in any of the categories of the other one.

Table 5.1 illustrates this idea. The table contains probabilities in a hypothetical population. If one were to predict the employment status of a person selected from this population in such a way that every member of the population has the same probability of being selected, then the best prediction is that this person will be employed. Indeed, the probability that this prediction will be correct is 0.44 and the probability of either one of the other two possible predictions is correct is less than this. However, based on the data in Table 5.1, other predictions may work better for individuals belonging to different age groups. If one were to predict the employment status of a young adult, then the best prediction is that this person is unemployed, with a success probability equal to $0.2/0.3$. For middle-aged individuals, the best prediction is that they are employed, with success probability $0.35/0.4$, and for old people the best prediction in this hypothetical population is that they are outside of the labor force, and this prediction has $0.2/0.3$ as success probability. The overall probability of success, when age is known, is

$$P(\text{young adult})P(\text{unemployed}|\text{young adult}) +$$
$$P(\text{middle aged})P(\text{employed}|\text{middle aged}) +$$
$$P(\text{old})P(\text{outside of the labor force}|\text{old}) =$$
$$0.3 \times 0.2/0.3 + 0.4 \times 0.35/0.4 + 0.3 \times 0.2/0.3 = 0.75.$$

That is, knowing the age group into which the individual belongs, improves the prediction of employment status: the probability of correct prediction is 0.75 in this case, as opposed to 0.44 when the age group is not known.

Table 5.1 Cross-classification of age by employment status in a hypothetical population

	Employed	Unemployed	Outside of the labor force	Total
Young adult	0.04	0.2	0.06	0.3
Middle-aged	0.35	0.02	0.03	0.4
Old	0.05	0.05	0.2	0.3
Total	0.44	0.27	0.28	

Knowing the age group of an individual would not improve the prediction, if the conditional distribution of the employment categories was the same in every age group. In this case, the prediction would be the same in all age groups. Also in this case, the conditional probability of an employment category within any age group would be the same as overall, that is, for all individuals taken together:

$$P(j|i) = P(+, j), \text{ for all columns } j \text{ and for all rows } i. \tag{5.15}$$

When (5.15) holds, the two variables are said to be independent of each other. The concept sounds symmetric, but (5.15) seems asymmetric. In fact, (5.15) is also a symmetric property. Indeed, (5.15) implies that

$$P(i|j) = \frac{P(i,j)}{P(+,j)} = \frac{P(j|i)P(i,+)}{P(+,j)} = \frac{P(+,j)P(i,+)}{P(+,j)}$$

thus

$$P(i|j) = P(i,+), \text{ for all rows } i \text{ and for all columns } j. \tag{5.16}$$

It is straightforward that also (5.16) implies (5.15).

Another important way of looking at independence is that (5.15) and (5.16) imply that

$$P(ij) = P(i|j)P(+,j) = P(i,+)P(+,j), \text{ for all rows } i \text{ and for all columns } j. \tag{5.17}$$

Also, (5.17) implies (5.15). Indeed,

$$P(j|i) = \frac{P(ij)}{P(i,+)} = \frac{P(i,+)P(+,j)}{P(i,+)} = P(+,j), \text{ for all columns } j \text{ and for all rows } i.$$

The last variant of the definition of the concept of independence is based on the odds ratio. The odds ratio and its generalizations will be central to many of the developments later in the book. Here only the simplest case is discussed. To be able to work with odds ratios, from now on, all probabilities are assumed to be positive. The word odds means the ratio of two probabilities: how many times is one outcome more likely than another outcome. For example, with the hypothetical data in Table 5.1, the odds of being employed versus unemployed in the population is 0.44/0.27. The value of the same odds for young adults is 0.04/0.2. This latter quantity is a conditional odds: the odds of being employed versus unemployed, if only young adults are considered.

The odds ratio is defined as the ratio of two conditional odds, for example, the ratio of the conditional odds of being employed versus unemployed for young adults divided by the same conditional odds for middle-aged individuals. This odds ratio has the value of

$$\frac{0.04/0.2}{0.35/0.02} = 0.011.$$

The meaning of this number is that the odds of being employed versus unemployed for young adults is about 1% of the same odds for middle-aged people. If both variables have only two categories, there is 1 odds ratio in the table. For tables formed by variables other than binary, there is an odds ratio defined by every pair of adjacent rows and adjacent columns. The intersection of the two adjacent rows and the two adjacent columns is a 2×2 subtable of the original table and has an odds ratio. If the table has I rows and J columns, then there are $(I-1) \times (J-1)$ such odds ratios. These are called the local odds ratios. In general, a local odds ratio has the form of

$$\frac{p(i,j)/p(i,j+1)}{p(i+1,j)/p(i+1,j+1)} = \frac{p(i,j)p(i+1,j+1)}{p(i,j+1)p(i+1,j)}, \quad i=1,\ldots,I-1, \quad j=1,\ldots,J-1.$$

It is straightforward that if the two variables are independent, then all the local odds ratios are equal to 1. Indeed, (5.17) implies that

$$\frac{p(i,j)p(i+1,j+1)}{p(i,j+1)p(i+1,j)} = \frac{p(i,+)p(+,j)p(i+1,+)p(+,j+1)}{p(i,+)p(+,j+1)p(i+1,+)p(+,j)} = 1, \tag{5.18}$$

$$\text{for } i=1,\ldots,I-1, \text{ and } j=1,\ldots,J-1.$$

It will be shown now that (5.18) also implies the independence of the two variables. Perhaps the most instructive proof of this fact uses an alternative set of odds ratios, called spanning cell odds ratio. If the table is of the size 2×2, the odds ratio is also the spanning cell odds ratio. Otherwise, the spanning cell odds ratios are defined by selecting a reference category of every variable. It will be assumed that for both variables the reference category is the first one. Then cell $(1,1)$, which contains the reference categories of both variables, is the reference cell. To define a spanning cell odds ratio, one also needs a spanning cell. Any cell (i,j), with $i>1$ and $j>1$, may serve as a spanning cell. Therefore, there are $(I-1) \times (J-1)$ spanning cells. Because (i,j) is neither in the row nor in the column of $(1,1)$, they together span a

2 × 2 subtable of the original table. The odds ratio in this subtable,

$$\frac{p(1,1)p(i,j)}{p(1,j)p(i,1)},$$

is a spanning cell odds ratios. Obviously,

$$\frac{p(1,1)p(i,j)}{p(1,j)p(i,1)} = \prod_{k=1}^{i-1}\prod_{l=1}^{j-1}\frac{p(k,l)p(k+1,l+1)}{p(k,l+1)p(k+1,l)}. \tag{5.19}$$

Then (5.19) implies immediately that if (5.18) holds, then all the spanning cell odds ratios of the table are also equal to 1:

$$\frac{p(1,1)p(i,j)}{p(1,j)p(i,1)} = 1, \text{ for } i = 2,\ldots,I, \text{ and } j = 2,\ldots,J. \tag{5.20}$$

When (5.20) holds,

$$p(i,j) = \frac{p(1,j)p(i,1)}{p(1,1)}$$

and

$$p(+,j) = \frac{p(1,j)p(+,1)}{p(1,1)}.$$

Thus

$$p(i|j) = \frac{p(i,j)}{p(+,j)} = \frac{p(i,1)}{p(+,1)} = p(i|1).$$

In other words, the conditional probability of a row category, given a column category, does not depend on the column given. Therefore,

$$\begin{aligned} p(i,+) &= \sum_{j=1}^{J} p(i,j) = \sum_{j=1}^{J} p(i|j)p(+,j) \\ &= \sum_{j=1}^{J} p(i|1)p(+,j) = p(i|1)\sum_{j=1}^{J} p(+,j) \\ &= p(i|1). \end{aligned}$$

Thus, (5.16) holds, indeed.

The summary of the results of this section is that

Theorem 5.6. *For an I × J contingency table, the properties (5.15), (5.16), (5.17), (5.18) and (5.20) are equivalent.* □

5.4.2 *Maximum Likelihood Estimation Under Independence*

To compute maximum likelihood estimates under the model of independence, the Lagrange multiplier method will be used. The model structure is written in the likelihood function, so that the marginal probabilities are the parameters. The marginal probabilities are arbitrary probability distributions.[8] The condition which is imposed by Lagrange multipliers is that the probabilities sum to 1. To impose the model structure, (5.17) is used. All the cell probabilities are assumed to be positive.

The components of the vector variable containing the frequencies in the cells of the table will be denoted as $X(i,j)$. Under multinomial sampling with n observations, the kernel of the log-likelihood is

$$\sum_{i,j} X(i,j)\log(p(i,+)p(+,j)).$$

The cell probabilities are supposed to sum to 1. The kernel of the augmented log-likelihood with Lagrange multiplier λ is

$$\sum_{i,j} X(i,j)\log(p(i,+)p(+,j)) + \lambda\left(\sum_{i,j} p(i,+)p(+,j) - 1\right).$$

For a fixed row i, the partial derivative of the augmented log-likelihood function according to $p(i,+)$ is

$$\sum_j \frac{X(i,j)}{p(i,+)} + \lambda \sum_j p(+,j) = \frac{X(i,+)}{p(i,+)} + \lambda. \tag{5.21}$$

Similarly, for a fixed column j, the partial derivative according to $p(+,j)$ is

$$\sum_i \frac{X(i,j)}{p(+,j)} + \lambda \sum_i p(i,+) = \frac{X(+,j)}{p(+,j)} + \lambda. \tag{5.22}$$

To find the MLE, the above expressions, for all i and j, are set to zero, and the resulting system of equations

$$\begin{aligned} \frac{X(i,+)}{p(i,+)} + \lambda &= 0, \quad \text{for } i = 1,\ldots,I \\ \frac{X(+,j)}{p(+,j)} + \lambda &= 0 \quad \text{for } j = 1,\ldots,J \end{aligned}$$

is solved. In order to obtain a solution, every equation in the first set is multiplied by $p(i,+)$, and every equation in the second set is multiplied by $p(+,j)$ to yield

$$\begin{aligned} X(i,+) + \lambda p(i,+) &= 0, \quad \text{for } i = 1,\ldots,I \\ X(+,j) + \lambda p(+,j) &= 0 \quad \text{for } j = 1,\ldots,J \end{aligned}$$

If the first set of equations is summed for all i, and the second set of equations is summed for all j, one obtains that

[8] This formulation and the notation which follows is not entirely precise. It relies on the fact that if arbitrary probability distributions $q(i)$ and $r(j)$ are chosen, and then $p(i,j)$ is defined as $q(i)r(j)$, then $p(i,+) = q(i)$ and $p(+,j) = r(j)$.

$$n + \lambda = 0.$$

From this, $\lambda = -n$, and if this is plugged in the above equations, one obtains the solution

$$p(i,+) = \frac{X(i,+)}{n}, \quad \text{for } i = 1,\ldots,I$$
$$p(+,j) = \frac{X(+,j)}{n}, \quad \text{for } j = 1,\ldots,J$$

or

$$p(i,j) = \frac{X(i,+)}{n}\frac{X(+,j)}{n}, \text{ for all } i,j. \tag{5.23}$$

The following theorem summarizes the properties of the MLE under the model of independence for a two-way table.

Theorem 5.7. *Let $\boldsymbol{p} = (p(i,j), i = 1,\ldots,I, j = 1,\ldots,J)$ be a strictly positive probability distribution on an $I \times J$ table and let it be independent, as defined in (5.17). Let the observations $\boldsymbol{X}$ be $\mathscr{M}(n,\boldsymbol{p})$ and assume that the row marginal frequencies $x(i,+)$ and the column marginal frequencies $x(+,j)$ are all positive. Then, the maximum likelihood estimates are as follows:*

$$\hat{p}(i,j) = \frac{x(i,+)}{n}\frac{x(+,j)}{n}, \textit{ for all } i,j, \tag{5.24}$$

$$\hat{p}(i,+) = \frac{x(i,+)}{n} \textit{ and } \hat{p}(+,j) = \frac{x(+,j)}{n}, \textit{ for all } i,j, \tag{5.25}$$

$$E(X\hat{(}i,j)) = \frac{x(i,+)x(+,j)}{n}, \textit{ for all } i,j. \tag{5.26}$$

Proof. Theorem 4.10 implies that (5.24) is a stationary point of the log-likelihood function and it is also in the model. To obtain the second derivative of the log-likelihood function, one has to consider the vector of first derivatives, which has length $I+J$, so that the first I components are those given in (5.21) and the last J components are those in (5.22). The second derivative is a block diagonal matrix of size $(I+J)\times(I+J)$. Both the $I\times I$ and $J\times J$ matrices along the main diagonal are diagonal, with entries

$$-\frac{x(i,+)}{p^2(i,+)}$$

and

$$-\frac{x(+,j)}{p^2(+,j)},$$

respectively.

Thus, the second derivative matrix is negative definite and (5.24) gives the MLE. The other statements are implied by Proposition 4.2. □

As can be seen from the likelihood function, the row and column marginal observed frequencies are sufficient statistics to estimate the cell probabilities under the model of independence and they need to be positive to get an MLE in the model. Later in the book, the assumption of all cell probabilities being positive will be relaxed; see Theorem 7.6. Also, the model of independence is an example of exponential families, which will be discussed in Sect. 7.2.

When the sampling procedure is Poisson, so that independent Poisson distributions are observed in each cell and the $\lambda(i,j)$ intensities of these distributions are proportional to the cell probabilities $p(i,j)$,[9] then, as implied by Theorem 2.5, the total of the observations, as a random variable, also has Poisson distribution with intensity[10]

$$\lambda = \sum_{i=1}^{I}\sum_{j=1}^{J}\lambda(i,j).$$

Then, $\lambda(i,j) = \lambda p(i,j)$, because the sum of the probabilities is 1. The kernel of the log-likelihood is

$$\sum_{i,j} X(i,j)\log\lambda(i,j) - \lambda(i,j).$$

and if independence holds,

$$\lambda(i,j) = \lambda p(i,+)p(+,j), \text{ for all } i \text{ and } j.$$

The kernel of the augmented log-likelihood function may be written as

$$\sum_{i,j}(X(i,j)\log(\lambda p(i,+)p(+,j)) - \lambda p(i,+)p(+,j)) + \mu\left(\sum_{i,j}p(i,+)p(+,j) - 1\right)$$

$$= \sum_{i,j}X(i,j)\log\lambda + \sum_{i,j}X(i,j)\log p(i,+)p(+,j) - \lambda + \mu\left(\sum_{i,j}p(i,+)p(+,j) - 1\right)$$

$$= N\log\lambda + \sum_{i,j}X(i,j)\log p(i,+)p(+,j) - \lambda + \mu\left(\sum_{i,j}p(i,+)p(+,j) - 1\right),$$

where μ is a Lagrange multiplier and N is the sum of the observations. The partial derivative according to λ is

$$\frac{N}{\lambda} - 1,$$

according to $p(i,+)$, for a fixed i is

$$\sum_j \frac{X(i,j)}{p(i,+)} + \mu\sum_j p(+,j) = \frac{X(i,+)}{p(i,+)} + \mu$$

[9] If the assumption that $\lambda(i,j)/p(i,j)$ is constant is not made, rather these quantities are unknown, it does not seem possible to test independence.

[10] Unfortunately, λ is the usual notation both for the Lagrange multiplier and the intensity of a Poisson distribution. It is used in the latter sense here.

and according to $p(+,j)$, for a fixed j is

$$\sum_i \frac{X(i,j)}{p(+,j)} + \mu \sum_i p(i,+) = \frac{X(+,j)}{p(+,j)} + \mu.$$

By setting the first one of these partial derivatives to zero, one obtains that $\lambda = N$ and the second and third of these partial derivatives are of the same structure as the partial derivatives (5.21) and (5.22) obtained under multinomial sampling. Consequently, the same solutions will maximize the likelihood function. A summary of these results is

Theorem 5.8. *Let $\mathbf{p} = (p(i,j), i = 1,\ldots,I, j = 1,\ldots,J)$ be an independent and strictly positive probability distribution on an $I \times J$ table, as defined in (5.17). Let independent Poisson variables be observed in each cell, so that their intensities are proportional to the cell probabilities, and let $X(i,j)$ be the number of observations from cell (i,j). Assume that all $x(i,+)$ and $x(+,j)$ are positive. Then, the maximum likelihood estimates are as follows:*

$$\hat{p}(i,j) = \frac{X(i,+)}{n}\frac{X(+,j)}{n}, \text{ for all } i,j, \tag{5.27}$$

$$\hat{p}(i,+) = \frac{X(i,+)}{n} \text{ and } \hat{p}(+,j) = \frac{X(+,j)}{n}, \text{ for all } i,j, \tag{5.28}$$

$$E(X(\hat{i},j)) = \frac{X(i,+)X(+,j)}{n}, \text{ for all } i,j. \tag{5.29}$$

□

The result in Theorem 5.8 is not based on conditioning on the total number of observations. As Theorem 2.6 shows, if the sample was assumed to have a multivariate Poisson distribution, then, conditioned on the total, the distribution would become multinomial, and the same MLEs would be obtained. This latter approach is widely used in the literature (see, e.g., [1]). The approach put forward here does not condition on the total. Such a conditioning is often possible and useful but not in all cases.

5.4.3 Tests of the Hypothesis of Independence

Tests of independence are based on comparing the observed data to their expectations under the model of independence. For this comparison, the Pearson chi-squared and the likelihood ratio statistics are used most often. Chapter 13 discusses briefly other related and also some completely unrelated methods of testing independence and other models for categorical data.

The Pearson chi-squared statistic for testing independence in an $I \times J$ contingency table, with observed frequencies in $\mathbf{X}$, is

$$\sum_{i,j} \frac{(X(i,j) - E(X\hat{(i,j)}))^2}{E(X\hat{(i,j)})}, \tag{5.30}$$

where $E(X\hat{(i,j)}$ is the maximum likelihood estimate of the expected frequency under independence. This expected frequency was shown to be the same under multinomial and Poisson sampling in the previous subsection.

The likelihood ratio statistic is

$$2\sum_{i,j} X(i,j) \log \frac{X(i,j)}{E(X\hat{(i,j)})}. \tag{5.31}$$

This statistic is a comparison of two maximized log-likelihoods for the observed data: in the numerator, the maximized log-likelihood if no model is assumed, and in the denominator, the maximized log-likelihood under the model of independence. Thus, the statistic measures how restrictive the model of independence is for the actual data.

The distributions of the test statistics (5.30) and (5.31), even if the hypothesis of independence holds, depend on the sample size and the true distribution $\mathbf{p}$. Obviously, in any realistic hypothesis, testing situation $\mathbf{p}$ is not known (if it was, there would be no need to test hypotheses about it). Because $\mathbf{p}$ is not known, the distributions of the test statistics (5.30) and (5.31) are not known, and no testing can be performed.

Quite interestingly, however, the asymptotic distributions as $n \to \infty$ do not depend on $\mathbf{p}$, as long as the true distribution is independent. In fact, the following result holds.

Theorem 5.9. *Let the distribution* $\mathbf{p}$ *on an* $I \times J$ *contingency table be independent. Then, under multinomial or Poisson sampling, the asymptotic distributions of (5.30) and (5.31) are the same, namely* $\mathscr{C}((I-1)(J-1))$.

Proof. The proof of this result is quite involved. Reference [13] contains a detailed proof with all the prerequisites. □

The number of degrees of freedom of the asymptotic distribution is usually interpreted as the number of rows minus one multiplied by the number of columns minus one. This is certainly true, but a different view of the formula is more relevant in understanding why this is the number of degrees of freedom and also generalizes to larger tables and more complex models to be discussed later in the book. The model of independence specifies the values of $(I-1)(J-1)$ linearly independent parameters of the distribution $\mathbf{p}$. The meaning of a parameter here is simply a function of the probabilities $\mathbf{p}$. That this many parameters are specified by the model of independence is best seen from (5.18) and Theorem 5.6. The linear independence of the restrictions in (5.18) is best seen by considering the $(I-1)(J-1)$ spanning cell odds ratios and showing that they are linearly independent. To see this, note first that the odds ratios may be computed from any table, not only one that contains probabilities. For an arbitrary set of $(I-1)(J-1)$ spanning cell odds ratios, a (not necessarily probability) table can be easily constructed that has these spanning cell odds ratios. Choose the first row and first column arbitrarily and determine the entry in the (i, j) cells for $i > 1$ and $j > 1$, so that the odds ratio spanned by this cell is as required. Finally, the table obtained may be transformed into a probability distribution by dividing each entry by the total of the table. Obviously, this normalization does not affect the odds ratios of the table.

Therefore, the number of degrees of freedom is equal to the number of linearly independent parameters prescribed by the model. This may sound counterintuitive. However, if one considers that this is the number of degrees of freedom of the (asymptotic) distribution of the statistics that are used to test model fit, then it seems logical that more deviation between the observed and the estimated frequencies is allowed (expected), when the model is more restrictive, i.e., when it is defined by a larger number of (linearly independent) restrictions.

By Theorem 5.9, an asymptotic test of the hypothesis of independence may be performed, if the data have a multinomial or Poisson distribution, by determining the MLEs under the model of independence, then computing the value of the Pearson chi-squared or of the likelihood ratio statistic, and then comparing the value obtained to a critical value taken from the chi-squared distribution with $(I-1)(J-1)$ degrees of freedom.

Because the test is asymptotic, "practical" rules are needed to determine in any given situation whether it can be used. There is a long history of somewhat contradicting suggestions in the literature. The controversy arises because of the imprecise nature of the question to which an answer is being sought. The most conservative rule is that in each cell, the expected frequency must not be less than 5. In addition to being (in light of other suggestions) too restrictive, a drawback of this rule is that it can only be checked after the data were collected. Other rules are stated in terms of the ratio of the total sample size to the number of cells of the table, e.g., one may require this ratio to be at least 5. Under multinomial or Poisson sampling, the same bound will apply to the average expected frequency. Based on the criterion that in a series of simulations, 95% of the simulated confidence intervals with 95% coverage actually contain the asymptotic critical value, [70] found that the asymptotic test based on the Pearson chi-squared statistic may be used with sample sizes equal to 2–3 times the number of cells, but the likelihood ratio statistic seems to require somewhat larger sample sizes. The simulation results pertained, beyond independence in two-way tables, to certain models for three-way tables but were restricted to variables with few categories.

With the increasing availability of Monte Carlo testing procedures (see Chap. 13) to approximate exact p-values, the asymptotic testing procedures in cases when their applicability is questionable are becoming less frequent, and, thus, the relevance of the rules discussed in the paragraph above diminishes.

5.5 Things to Do

1. Let X be $\mathscr{B}(n, 0.5)$. Show that if $|x_1 - np| > |x_2 - np|$, then $P(X = x_1) < P(X = x_2)$ Is this true when $p \neq 0.5$?
2. For $X \sim \mathscr{B}(15, 0.6)$ and $H_0 : p = 0.6$, $H_1 : |p - 0.6| > 0.2$ determine the region of rejection so that the error type I probability is closest to 0.1 and determine the power of the test using (5.4).
3. Repeat the previous exercise with $H_1 : |p - 0.6| > 0.3$
4. Repeat the previous two exercises for $n = 30$.
5. Repeat the previous exercise using the normal approximation.
6. For $X \sim \mathscr{B}(15, 0.6)$ and observation $x = 9$, determine the maximum plausibility estimate of p.
7. Develop a test of $H_0 : p = 0.6$ against the one sided alternative that $H_1 : p - 0.6 > 0.2$.
8. Determine the power of the above test when $n = 30$.
9. Give a rigorous proof of Theorem 5.5.
10. Suppose $X \sim \mathscr{B}(100, p)$ and let the hypothesis be that $p = 0.7$. Let $x = 60$. Determine the value of the Pearson chi-squared statistic and find the critical value for a decision with probability of error type I equal to 0.05. Decide about the hypothesis. Determine the achieved probability level of the test.
11. Repeat the previous exercise with the likelihood ratio statistics. Rely on the asymptotic distribution of the test statistic.
12. Test the hypothesis of $p \leq 0.3$ using a binomial observation of 4 with $n = 10$. Use 0.05 as the minimal error type I probability.
13. In a survey, there were 1288 male and 1359 female respondents. The numbers of those employed, unemployed, and out of the labor force were 701, 115, and 472 for the male respondents and 723, 97, and 549 for the female respondents. Compute the maximum likelihood estimate of the probability that a man is unemployed, under the hypothesis that the two variables are independent. Compute, under the same hypothesis, the MLE of the probability that someone is unemployed. Compute the MLE of the expected number of employed men with and without the hypothesis of independence. Explain why this expected number is the MLE of the relevant probability multiplied by the sample size. How does this expectation depend on the data collection procedure?
14. Find the assumptions with respect to the data collection procedure in the problem above, which are needed for the calculations to be valid.
15. Determine the values of the Pearson chi-squared and the likelihood ratio statistics for the hypothesis of independence and the employment data above. Carry out the test of the hypothesis.
16. Assume that you wish to test the hypothesis of independence, but $x(1,+) = 0$ but the other row and all column marginals are positive. Restrict the model of independence to the rows where the observed marginal is not zero. Obtain the MLE under this model.
17. Test the hypothesis of independence in a 2×2 table, using the observations $x(1,1) = 10$, $x(1,2) = 20$, $x(2,1) = 30$, $x(2,2) = 40$.

18. Repeat the previous exercise with $x(1,1) = 100$, $x(1,2) = 200$, $x(2,1) = 300$, $x(2,2) = 400$. Is your decision the same or is it different? Why?
19. Suppose someone says that in the two exercises above, the same probability distribution was observed, so the decision should be the same. Explain why is this a false argument.
20. The spanning cell odds ratios of an $I \times J$ table may be arranged into an $(I - 1) \times (J - 1)$ table. Generate a 3×4 probability distribution with the spanning cell odds ratios 2, 5, and 0.2 in the first row and 0.5, 1, and 2 in the second row.

Chapter 6
Association

Abstract The statistical analysis of associations is a central theme in this book. This chapter starts with a description of the properties of the odds ratio, including its maximum likelihood estimation. Because of its variation independence from the marginal distributions, it is argued the odds ratio is the most useful measure of association. The structure of $I \times J$ tables, as described by the systems of local or spanning cell odds ratios, which are generalizations of the simple odds ratio defined for 2×2 tables, is described and analyzed by association models. The odds ratio is generalized to higher-dimensional tables by introducing a hierarchical structure of conditional odds ratios. Independence may be seen as lack of association, but a related simple structure, conditional independence, is found more often in real data, and properties of the maximum likelihood estimates under conditional independence are studied.

The concept of association is studied in this chapter by introducing general variants of the odds ratio and by discussing conditional independence, which is perhaps the most often found simple structure in categorical data.

6.1 The Odds Ratio

The odds ratio, as the ratio of two conditional odds, was introduced in Sect. 5.4.1. For the case of $I \times J$ tables, the odds ratio was generalized to the systems of local or spanning cell odds ratios. As implied by Theorem 5.6, these two systems are equal to 1[1] at the same time, namely, if the two categorical variables forming the table are independent. Also, (5.19) means that whether or not the local odds ratios are equal to 1, if one knows their values, the spanning cell odds ratios can be determined. Conversely,

[1] That is, each of their elements is equal to 1.

$$\frac{p(i,j)p(i+1,j+1)}{p(i,j+1)p(i+1,j)} = \frac{\frac{p(i,j)p(1,1)}{p(i,1)p(1,j)} \frac{p(i+1,j+1)p(1,1)}{p(i+1,1)p(1,j+1)}}{\frac{p(i,j+1)p(1,1)}{p(i,1)p(1,j+1)} \frac{p(i+1,j)p(1,1)}{p(i+1,1)p(1,j)}}, \tag{6.1}$$

for $i = 1, \ldots, I-1$ and $j = 1, \ldots, J-1$. Therefore, the two sets of odd ratios contain the same information, and the topic of this section is the meaning and interpretation of this information.

In a 2×2 table, if the two variables are independent, then the odds ratio is 1. Values of the odds ratio other than 1 mean the lack of independence, called the presence of association between the variables. If the conditional odds of being unemployed, versus being employed, are not equal in the case of men and women, rather, say, the value of the odds is higher for men than for women, one can say that being unemployed is associated with being a man. If, to the contrary, the value of the conditional odds is higher for women than for men, then being unemployed is associated with being a woman.

The odds ratio quantifies the strength of the association. When the odds ratio comparing the odds of men to the odds of women to be unemployed versus being employed is, say, 4, then the odds of being unemployed versus employed is four times higher for men than for women. This means a stronger association between being a man and being unemployed than if the value of the odds ratio was, say, 2. As measured by the odds ratio, the association is twice as strong in the former than in the latter case. The strength of the association is the same, when the odds ratio is α or $1/\alpha$, but the directions of association are the opposite. Indeed, when either the rows or the columns of the 2×2 table are swapped, the value of the odds ratio changes to its reciprocal.

Thus, the odds ratio is a measure of (the strength of) association. When its value is 1, the two variables are independent; hence there is no association. In this sense, independence (no association) is a special case of association, rather than being opposite to it. The difference between the two views is that if one considers independence as the opposite of association, the two concepts appear to be far and distinct. If one considers independence as a special case of association, then it is clear that sometimes very little distinguishes association from independence. For example, if the odds ratio is 1.03, one may be wondering, whether this should be interpreted as association. The answer to this question depends strongly on whether the 1.03 is a population value or is a value found in a sample.

If the value of the odds ratio in the population of interest is, say, 1.03, deciding whether the weak association seen should be interpreted as essentially independence or as essentially association is not a statistical question.[2] This is a question of substantive significance. Whether taking into account aspects like the validity of the data collection procedure, or the possible measurement errors, or the expected effect or importance of a value of 1.03, as opposed to a value of 1, these values should or should not be interpreted as different—is a question the substantive scientists should answer. The answer often depends on traditions, on the usual effect sizes seen in that

[2] Remember, statistics is about inference from sample to population and one is assumed to have the population value in this case.

particular field, and also on personal experiences. However, many statisticians experience a reluctance on the side of the substantive scientists to take responsibility for such decisions, and they often try to determine the statistical significance of the deviation of 1.03 from 1, which is entirely inappropriate in this case.[3]

On the other hand, when the value of 1.03 is observed in a sample, the issue of statistical significance[4] is relevant. To answer this question, the MLE of the odds ratio will be derived, and then the asymptotic standard error of the MLE of the odds ratio will be determined using the δ-method.

6.1.1 Maximum Likelihood Estimation of the Odds Ratio

To derive the MLE of the odds ratio,[5] the distribution on the 2×2 table will be parameterized in such a way that the odds ratio is one of the parameters. Parameterization such that the odds ratio (or its generalizations) is one of the parameters will be of central importance in later developments in the book. Define the following quantities:

$$t^4 = p(1,1)p(1,2)p(2,1)p(2,2) \tag{6.2}$$

$$u^4 = \frac{p(1,1)p(1,2)}{p(2,1)p(2,2)} \tag{6.3}$$

$$v^4 = \frac{p(1,1)p(2,1)}{p(1,2)p(2,2)} \tag{6.4}$$

$$w^4 = \frac{p(1,1)p(2,2)}{p(1,2)p(2,1)}. \tag{6.5}$$

Clearly, w is the 4th root of the odds ratio. Then,

$$\begin{aligned} p(1,1) &= t\,u\,v\,w \\ p(1,2) &= t\,u\,v^{-1}\,w^{-1} \\ p(2,1) &= t\,u^{-1}\,v\,w^{-1} \\ p(2,2) &= t\,u^{-1}\,v^{-1}\,w. \end{aligned} \tag{6.6}$$

If the observations are $\mathbf{X} \sim \mathscr{M}(n, \mathbf{p})$, the kernel of the log-likelihood, augmented with a Lagrangian term, is

[3] Some textbooks go as far as suggesting classifications as to what is a weak or a strong effect, for example, in terms of odds ratios (or the correlation coefficient in a different context). Such suggestions, without taking into account the circumstances of data collection and the actual research question or policy implications, go beyond the scope of statistical analysis. Unfortunately, some substantive scientists are happy to rely on such suggestions.

[4] That is, whether or not the observed value warrants us to think that the population value is not 1.

[5] The odds ratio is not a one-to-one function of the cell probabilities, and thus Proposition 4.2 cannot be used directly.

$$\begin{aligned} n\log t + (X(1,+) - X(2,+))\log u + (X(+,1) - X(+,2))\log v + \\ (X(1,1) + X(2,2) - X(1,2) - X(2,1))\log w + \\ \lambda(tuvw + tuv^{-1}w^{-1} + tu^{-1}vw^{-1} + tu^{-1}v^{-1}w - 1). \end{aligned} \tag{6.7}$$

We are interested in finding out whether w^4 has a value which uniquely maximizes (6.7). Only positive values of w are considered, and as $(.)^4$ is a strictly monotone and thus one-to-one function on the positive real numbers, this is the same value for which w uniquely maximizes (6.7). Obviously,

$$w^4 = \frac{\hat{p}(1,1)\hat{p}(2,2)}{\hat{p}(1,2)\hat{p}(2,1)} = \frac{x(1,1)x(2,2)}{x(1,2)x(2,1)}, \tag{6.8}$$

and similar choices for t, u, v maximize (6.7) because with these values, the likelihood is the same as with the MLE for the cell probabilities. We will show now that no other choice of w maximizes (6.7). Because of (6.8), this is implied by showing that no choice of $\mathbf{p}$ other than $\hat{\mathbf{p}}$ maximizes (6.7). To see this, consider the partial derivatives of (6.7), according to t, u, v, w, respectively, simplified using (6.6)

$$\begin{aligned} \frac{n}{t} + \frac{\lambda}{t}, \\ \frac{X(1,+) - X(2,+)}{u} + \lambda\frac{p(1,+) - p(2,+)}{u}, \\ \frac{X(+,1) - X(+,2)}{v} + \lambda\frac{(p(+,1) - p(+,2)}{v}, \\ \frac{X(1,1) + X(2,2) - X(1,2) - X(2,1)}{w} + \\ \lambda\frac{p(1,1) + p(2,2) - p(1,2) - p(2,1)}{w}. \end{aligned} \tag{6.9}$$

By setting (6.9) equal to zero, one obtains that $\lambda = -n$ and the other equations become

$$\begin{aligned} X(1,+) - X(2,+) - n(p(1,+) - p(2,+)) = 0, \\ X(+,1) - X(+,2) - n(p(+,1) - p(+,2)) = 0, \\ X(1,1) + X(2,2) - X(1,2) - X(2,1) - \\ n(p(1,1) + p(2,2) - p(1,2) - p(2,1)) = 0. \end{aligned} \tag{6.10}$$

Together with

$$p(1,1) + p(1,2) + p(2,1) + p(2,2)) = 1, \tag{6.11}$$

the equations in (6.10) are a system of linear equations with a full rank matrix of coefficients

$$\begin{pmatrix} -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

and thus a unique solution, which is $\mathbf{p} = \hat{\mathbf{p}}$. The next theorem summarizes this result.

Theorem 6.1. *If observations $\boldsymbol{X} \sim \mathscr{M}(n,\boldsymbol{p})$ are available on a 2×2 contingency table, the MLE of the odds ratio θ is*

$$\hat{\theta} = \frac{X(1,1)X(2,2)}{X(1,2)X(2,1)} = \frac{\hat{p}(1,1)\hat{p}(2,2)}{\hat{p}(1,2)\hat{p}(2,1)}. \tag{6.12}$$

□

That is, the MLE of the odds ratio is the odds ratio computed from the MLEs of the cell probabilities.

Theorem 6.2. *Under the conditions of Theorem 6.1, the asymptotic standard deviation*[6] *of the MLE of the odds ratio is*

$$\frac{\theta}{\sqrt{n}}\sqrt{\frac{1}{p(1,1)} + \frac{1}{p(1,2)} + \frac{1}{p(2,1)} + \frac{1}{p(2,2)}}. \tag{6.13}$$

Proof. The proof uses the δ-method discussed in Sect. 3.4. The formula in (6.12) is a transformation of an $\mathbf{X}$ which has a multinomial distribution, and its partial derivative is a 1×4 vector:

$$\left(\frac{\hat{\theta}}{X(1,1)}, -\frac{\hat{\theta}}{X(1,2)}, -\frac{\hat{\theta}}{X(2,1)}, \frac{\hat{\theta}}{X(2,2)}\right). \tag{6.14}$$

The partial derivative evaluated at $E(\mathbf{X}) = n\mathbf{p}$ is

$$\frac{\theta}{n}\left(\frac{1}{p(1,1)}, -\frac{1}{p(1,2)}, -\frac{1}{p(2,1)}, \frac{1}{p(2,2)}\right).$$

To obtain the asymptotic variance, this needs to be multiplied by the covariance matrix of $\mathbf{X}$

$$\begin{pmatrix} \mathrm{np(1,1)(1-p(1,1))} & \mathrm{-np(1,1)p(1,2)} & \mathrm{-np(1,1)p(2,1)} & \mathrm{-np(1,1)p(2,2)} \\ \mathrm{-np(1,2)p(1,1)} & \mathrm{np(1,2)(1-p(1,2))} & \mathrm{-np(1,2)p(2,1)} & \mathrm{-np(1,2)p(2,2)} \\ \mathrm{-np(2,1)p(1,1)} & \mathrm{-np(2,1)p(1,2)} & \mathrm{np(2,1)(1-p(2,1))} & \mathrm{-np(2,1)p(2,2)} \\ \mathrm{-np(2,2)p(1,1)} & \mathrm{-np(2,2)p(1,2)} & \mathrm{-np(2,2)p(2,1)} & \mathrm{np(2,2)(1-p(2,2))} \end{pmatrix}$$

yielding

$$\theta(1,-1,-1,1)$$

and this multiplied by the transpose of the partial derivative is

$$\frac{\theta^2}{n}\left(\frac{1}{p(1,1)} + \frac{1}{p(1,2)} + \frac{1}{p(2,1)} + \frac{1}{p(2,2)}\right).$$

□

[6] The $\sqrt{n}$ is removed from the formula, if the asymptotic standard error of $\sqrt{n}$ times the estimate of the odds ratio is considered.

The MLE of the odds ratio is asymptotically normal, with variance given as the square of (6.13). In practice, the population quantities in (6.12) are replaced by their respective MLEs. This yields an approximate test for the hypothesis that the odds ratio has a particular value, e.g., 1. Similarly, asymptotic confidence intervals may be constructed.

Instead of the odds ratio, often its logarithm is used. Many users feel that the log odds ratio is a more readily interpretable measure of the strength of the association than the odds ratio itself. The log odds ratio gives positive or negative values (instead of values larger or smaller than 1) and the same strength, but opposite direction of association is shown by values of the same magnitude but different signs (instead of reciprocals).

As the logarithm is a one-to-one function on its domain, Proposition 4.2 may be used to obtain that the MLE

$$\hat{\log \theta} = \log X(1,1) - \log X(1,2) - \log X(2,1) + \log X(2,2). \tag{6.15}$$

The vector of partial derivatives of (6.15) is

$$\left(\frac{1}{X(1,1)}, -\frac{1}{X(1,2)}, -\frac{1}{X(2,1)}, \frac{1}{X(2,2)} \right). \tag{6.16}$$

A comparison of (6.16) with (6.14) shows how the asymptotic standard deviation will differ from the one given in Theorem 6.13, and it is of the form given next.

Proposition 6.1. *Under multinomial sampling, the asymptotic standard deviation of the MLE of the log odds ratio is*

$$\frac{1}{\sqrt{n}} \sqrt{\frac{1}{p(1,1)} + \frac{1}{p(1,2)} + \frac{1}{p(2,1)} + \frac{1}{p(2,2)}}.$$

□

This result is important, because it shows that the asymptotic standard deviation of the MLE of the log odds ratio does not depend on the true value of the odds ratio.

6.1.2 *Variation Independence of the Odds Ratio and the Marginal Distributions*

This subsection presents the main argument in favor of using the odds ratio as a measure of association. The argument starts with the simple observation that if the joint distribution of the variables A and B is known, then their marginal distributions are also known. But if the marginal distributions of A and of B are given, their joint distribution cannot be derived without additional information.

That is, the joint distribution of two variables contains information in addition to their individual distributions. This additional information is called the association (or interaction) of the two variables. We argue in this subsection that the information that is contained in the joint distribution but not in the marginal distributions is best operationalized by the odds ratio.

In fact, the additional information is not operationalized that easily. For instance, consider $p(1,1)$ as a candidate to be identified with the additional information. When, in addition to $p(1,+)$ and $p(+,1)$, which contain the same information as the marginal distributions, also $p(1,1)$ is given, the joint distribution may be determined as

$$\begin{aligned} p(1,1) &= p(1,1) \\ p(1,2) &= p(1,+) - p(1,1) \\ p(2,1) &= p(+,1) - p(1,1) \\ p(2,2) = 1 - p(1,1) &- p(1,2) - p(2,1). \end{aligned}$$

Therefore, knowing $p(1,1)$, in addition to the one-way marginals, is sufficient to reconstruct the joint distribution; thus the information is sufficient. However, giving the value of $p(1,1)$ cannot be identified with providing the information which is additional in the joint distribution, as compared to the marginal distributions. To illustrate this, note that the possible range of $p(1,1)$ is the interval $(0,1)$, so are the ranges of the marginal probabilities. Therefore, the choices $P(1,+) = 0.3$ and $p(+,1) = 0.6$ are possible choices. But these values of the marginal probabilities change the possible range of $p(1,1)$, from $(0,1)$ to $(0,0.3)$. Thus, $p(1,1)$ is not entirely unaffected by the marginal distributions, and, therefore, the information in $p(1,1)$ is not entirely additional to the information contained in the marginals.

A more formal way of describing the above relationship is that the joint range of the parameters $p(1,+)$ and $p(+,1)$ is the Cartesian product of their individual ranges, $(0,1) \times (0,1)$, but the joint range of $p(1,+)$ and $p(+,1)$, on the one hand, and of $p(1,1)$, on the other hand, is not the Cartesian product of the respective ranges, which is $(0,1) \times (0,1)) \times (0,1)$. In general, let $\mathscr{P}$ contain the probability distributions[7] $\mathbf{p}$ with c categories, and let $\theta_i : \mathscr{P} \to \mathbb{R}^{d_i}$, $i = 1, \ldots, k$, arbitrary functions. The functions θ_i are called parameters of the distributions in $\mathscr{P}$. Let $R(\theta_i)$ denote the range of θ_i, i.e., if $t \in R(\theta_i)$, then there exists $\mathbf{p} \in \mathscr{P}$, such that $\theta_i(\mathbf{p}) = t$. Further, a vector consisting of such parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$', is also a parameter.

The parameters $\theta_1, \ldots, \theta_k$ are called variation independent, if

$$R(\boldsymbol{\theta}) = \times_{i=1}^{k} R(\theta_i).$$

This concept was introduced and studied in simple settings in Sect. 4.1.2.

[7] The subsequent concepts of parameter and parameterization are also used for frequency distributions.

A parameter $\boldsymbol{\theta}$ is a parameterization, if it contains all information that is needed to determine $\mathbf{p}$, i.e., if it is invertible on its range $R(\boldsymbol{\theta})$.

Obviously, the simplest parameterization of $\mathbf{p}$ is $\boldsymbol{\theta} = \mathbf{p}$. But there may be other parameterizations, such that some of the parameters used represent characteristics of $\mathbf{p}$ that are relevant for the statistical problem at hand. Such a characteristic may be, in the case of a 2×2 table, the lack or presence of independence or the strength of association among the two variables. Important properties of parameterizations containing the odds ratio are given in the next theorem.

Theorem 6.3. *For positive distributions* $\mathbf{p}$ *on a* 2×2 *table:*

1. *The parameters* $\theta_1 = (p(1,+), p(+,1))'$ *and* $\theta_2 = \frac{p(1,1)p(2,2)}{p(1,2)p(2,1)}$ *are variation independent.*
2. (θ_1, θ_2) *is a parameterization of* $\mathbf{p} \in \mathscr{P}$.
3. *Let* (θ_1, θ) *be a parameterization of* $\mathbf{p}$ *with variation independent components. Then* θ *is a one-to-one function of the odds ratio.*

Proof. 1. The range of $(p(1,+), p(+1))'$ is $(0,1)^2$, and the range of the odds ratio is all positive numbers. The proof shows that for any choices of the parameters, there is a distribution with those values of the parameters. More precisely, it is shown that for any value of $(p(1,+), p(+1))$, there is a choice of $p(1,1)$ so that they lead to the selected value of the odds ratio. It can be assumed without loss of generality that $p(1,+) \leq p(+,1)$ and thus the range of $p(1,1)$ is $(0, p(1,+))$.

In order to sum to 1, $\mathbf{p} = (p(1,1), p(1,+) - p(1,1), p(+,1) - p(1,1), 1 - p(1,+) - p(+,1) + p(1,1))'$. Thus, the numerator of the odds ratio is

$$p(1,1)(1 - p(1,+) - p(+,1) + p(1,1)),$$

which is a parabola in $p(1,1)$, that is zero at $p(1,1) = 0$ and at $p(1,1) = p(1,+) + p(+,1) - 1$. The parabola is positive outside of this interval.

The denominator of the odds ratio is

$$(p(1,+) - p(1,1))(p(+,1) - p(1,1)) = (p(1,1) - p(1,+))(p(1,1) - p(+,1)),$$

which is also a parabola in $p(1,1)$, that is, zero at $p(1,1) = p(1,+)$ and at $p(1,1) = p(+,1)$ and is positive outside of this interval. It follows that over the range $(0, p(1,+))$ of $p(1,1)$, the denominator is positive and monotone decreasing.

To complete the proof, it is shown that as $p(1,1)$ moves over its range, the odds ratio takes on all its values, and each of them exactly once. Two cases are distinguished.

Assume first that $p(1,+) + p(+,1) - 1$ is less than or equal to zero. Then, the interval on which the numerator is negative is outside of the range of $p(1,1)$. At $p(1,1) = 0$, the numerator is zero, the denominator is positive, and at $p(1,1) = p(1,+)$, the numerator is positive, the denominator is zero. Further, the numerator monotonically increases, the denominator monotonically decreases as $p(1,1)$ moves from 0 to $p(1,+)$. Therefore, the odds ratio monotonically increases from zero to infinity[8] and, by continuity, takes on every nonnegative value exactly once.

If, instead, $p(1,+)+p(+,1)-1$ is more than zero, then in order for $p(2,2)$ to be positive, the range of $p(1,1)$ is $(p(1,+)+p(+,1)-1), p(1,+)$. On this range, the odds ratio is zero at the starting point, and monotone increases to infinity at the endpoint, and the claim is implied just like in the previous case.

In either case, there is exactly one choice of $p(1,1)$, under which $\mathbf{p}$ has the required marginals and the required odds ratios.

2. The argument above shows that for every choice of the parameters, there is exactly one solution for $p(1,1)$ on the interval

$$(max(0, p(1,+)+p(+,1)-1), p(1,+)).$$

If $max(0, p(1,+)+p(+,1)-1)$ is not zero, then on the interval between zero and $p(1,+)+p(+,1)-1$, the numerator of the odds ratio is negative, the denominator is positive, so no solution is possible. Also, $p(1,1)$ cannot exceed $p(1,+)$, so there is exactly one choice of $p(1,1)$ on the interval $(0,1)$ and, therefore, the parameter function is invertible.

3.This follows from the second part of the theorem. □

Theorem 6.3 implies that in the case of a 2×2 table, the odds ratio is essentially the only parameter, which, together with the marginal distributions, provides a parameterization of the distribution and is also variation independent from them. Because it provides a parameterization, it contains all information in the joint distribution that is additional to the information in the marginal distributions. Because it is variation independent from the marginals, it contains only information that is additional to the marginals.

Therefore, the association among the variables forming the table, as the information in the joint distribution that is additional to the information in the marginals, is represented by the odds ratio.

The values of parameters may be used to describe distributions and to compare various distributions. These distributions may pertain to different populations or to the same population in different time periods. For example, consider the 2×2 distributions in Table 6.1. Using the marginal distributions and the odds ratio as parameters, one sees that the marginals of the first two distributions are the same (i.e., the individual distributions of the two variables are identical), but the odds ratio is higher in the second distribution than in the first one (i.e.,, the association among the two variables is stronger in the second table). In the third table, the strength of the association is (about) the same as in the second table, but the marginals are different. This parameterization also shows that in the second table, $p(1,+)=p(+,1)$, that is, the joint distribution is symmetric. On the other hand, this parameterization does not make it obvious that the conditional probabilities $p(1,1)/p(+,1)$ are the same in the

[8] More precisely: increases without and upper bound.

last two distributions. Different parameterizations may be useful when one is interested in studying different properties of distributions. Of course, the more complex is the contingency table, the more important is the choice of the parameterization.

Table 6.1 Three 2×2 distributions with different parameter values

0.30	0.10	0.35	0.05	0.165	0.165
0.40	0.20	0.35	0.15	0.165	0.505

Variation independent parameters have the advantage, that they do not influence each other's range and, thus, may be interpreted without respect to the values of other parameters. To illustrate this, consider a parameterization of the distribution on the 2×2 table which uses, in addition to the marginal probabilities, the ratio of the value of $p(1,1)$ to the value it would have under independence[9]:

$$\frac{p(1,1)}{p(1,+)p(+,1)}.$$

Obviously, the two marginal probabilities and the probability ratio above provide a parameterization of the distribution, just like $p(1,+)$, $p(+,1)$, $p(1,1)$. Further, the probability ratio could be used as a measure of association. The value of 1 means no association (independence), and values different from 1 mean association. The question is the strength of association one can infer from different values of the probability ratio. Table 6.2 shows two probability distributions given in [72]. The probability ratio is 2 for both tables, that is, the probability of cell $(1,1)$ is twice as much as it would be under independence, given the marginals. Does association in the two distributions have the same strength? Given the marginal probabilities, the upper bound of the value of $p(1,1)$ in the first distribution is 0.2 and in the second one 0.3. Therefore, the upper bound of the probability ratio is about 2.2 in the first table and is about 3.3 for the second table. Given the different ranges implied by the different marginal distributions, association, as measured by the probability ratio, is much closer to its maximal value in the case of the first distribution, than in the case of the second one, and, therefore, may be considered stronger, in spite of the same value.

Table 6.2 Two 2×2 distributions with the same value of the probability ratio

0.18	0.27	0.18	0.12
0.02	0.53	0.12	0.58

In general, if a parameter is not variation independent of other parameters, it cannot be interpreted without taking those parameters into account, because its range may depend on the values of those other parameters and thus lacks calibration.

[9] This ratio was used historically in social mobility research, see, e.g., [88].

To sum up, in the case of 2 × 2 tables, the odds ratio contains all information in the joint distribution, which is there in addition to the marginals, and contains only additional information, because it is variation independent from them. Therefore, the odds ratio measures the strength of association between the two variables that form the table. The usage and interpretation of parameters that are not variation independent form each other should be avoided.

In the case of $I \times J$ tables, the association cannot be described by a single number, if at least one of I or J is greater than 2. In that case, the association structure is described by the $(I-1)(J-1)$ local or spanning cell odds ratios. Either set of odds ratios is variation independent from the marginal distributions and, at the same time, contains enough information, in addition to the marginals, so that the distribution can be reproduced. This fact is implied by a general result to be given in Sect. 12.1. In this sense, the set of local or spanning cell odds ratios really describes the association structure in the $I \times J$ table.

We close this subsection with important results about variation and likelihood independence. The results are formulated for a two-way table but apply more generally if the variables forming the contingency table are combined into two groups.

Theorem 6.4. *Suppose that a model for the distribution* $\boldsymbol{p}$ *on a two-way table is such that it contains two groups of restrictions, one on the marginal distribution* $\boldsymbol{p}_{i,+} = \{p_{i,+} : i = 1,\dots,I\}$[10]

$$f(\boldsymbol{p}_{i,+}) = 0,$$

and one on the conditional distribution $\boldsymbol{p}_{j|i} = \{p_{j|i} : i = 1,\dots,I, j = 1,\dots,J\}$

$$g(\boldsymbol{p}_{j|i}) = 0.$$

Then, under this model, $\boldsymbol{p}_{i+}$ *and* $\boldsymbol{p}_{j|i}$*, as parameters of the distribution* $\boldsymbol{p}$*, are variation independent and likelihood independent in the case of multinomial sampling.*

Proof. The kernel of the logarithm of the multinomial likelihood, when no model is assumed, only the probabilities are required to sum to 1, may be written as

$$\begin{aligned}
&\sum_i \sum_j X_{i+} X_{j|i} \log(p_{i+} p_{j|i}) + \\
&\lambda(\sum_i p_{i+} - 1) + \sum_i \kappa_i (\sum_j p_{j|i} - 1) = \\
&\sum_i \sum_j X_{i+} X_{j|i} \log p_{i+} + \lambda(\sum_i p_{i+} - 1) + \\
&\sum_i \sum_j X_{i+} X_{j|i} \log p_{j|i} + \sum_i \kappa_i (\sum_j p_{j|i} - 1) = \\
&\sum_i X_{i+} \log p_{i+} \sum_j X_{j|i} + \lambda(\sum_i p_{i+} - 1) + \\
&\sum_i X_{i+} \sum_j X_{j|i} \log p_{j|i} + \sum_i \kappa_i (\sum_j p_{j|i} - 1) =
\end{aligned}$$

[10] The notations $p(i,j)$ and p_{ij} mean the same and are used interchangeably for better readability of the text.

$$\sum_i X_{i+} \log p_{i+} + \lambda(\sum_i p_{i+} - 1) +$$
$$\sum_i X_{i+} \sum_j X_{j|i} \log p_{j|i} + \sum_i \kappa_i (\sum_j p_{j|i} - 1),$$

where the conditional distributions from the observed frequencies are defined as $X_{j|i} = X_{ij}/X_{i+}$, with $0/0 = 0$. The model under consideration may be imposed by adding two more Lagrangians:

$$\sum_i X_{i+} \log p_{i+} + \lambda(\sum_i p_{i+} - 1) + \mu f(\mathbf{p}_{i+}) +$$
$$\sum_i X_{i+} \sum_j X_{j|i} \log p_{j|i} + \sum_i \kappa_i (\sum_j p_{j|i} - 1) + \nu g(\mathbf{p}_{j|i}).$$

Therefore, the kernel of the likelihood factorizes into two components, each depending on one component of the parameter only; thus likelihood independence holds.

The joint domain of $(\mathbf{p}_{i+}, \mathbf{p}_{j|i})$ under the model is

$$\{\mathbf{p}_{i+} : f(\mathbf{p}_i) = 0\} \times \{\mathbf{p}_{j|i} : g(\mathbf{p}_{j|i}) = 0, \},$$

which is the Cartesian product of the individual ranges of $\mathbf{p}_{i+}$ and $\mathbf{p}_{j|i}$; thus variation independence holds, too. □

An immediate consequence of Theorem 6.4 and of the component-wise maximization procedure, mentioned in Sect. 4.1.2 (see in particular Theorem 4.6) which is possible in this case, is that if any of f and g is constant zero, that is, it does not restrict the relevant parameters, then the MLE of this parameter is equal to its observed value (its unrestricted MLE).

Am important case, when the conditions of Theorem 6.4 hold, is when the model restricts the (local or spanning cell) odds ratios of a two-way table.

Theorem 6.5. *If a model restricts the (local or spanning cell) odds ratios of a two-way table, as*

$$g(OR) = 0,$$

using the notation of Theorem 6.4, then the restriction can be written as

$$g^{\sim}(\boldsymbol{p}_{j|i}) = 0,$$

for some function $g^{\sim}$. If the MLE under this model exists, then its one-way marginals are equal to the observed marginal distributions.

Proof. A local odds ratio may be written as

$$\frac{p(i,j)p(i+1,j+1)}{p(i,j+1)p(i+1,j)} = \frac{p(i,+)p(j|i)p(i+1,+)p(j+1|i+1)}{p(i,+)p(j+1|i)p(i+1,+)p(j|i+1)} = \frac{p(j|i)p(j+1|i+1)}{p(j+1|i)p(j|i+1)},$$

and a spanning cell odds ratio may be written as

$$\frac{p(1,1)p(i,j)}{p(1,j)p(i,1)} = \frac{p(1,+)p(1|1)p(i,+)p(j|i)}{p(1,+)p(j|1)p(i,+)p(1|i)} = \frac{p(1|1)p(j|i)}{p(j|1)p(1|i)},$$

thus any function g of the odds ratios may be written as a function of the conditional distributions $\mathbf{p}_{j|i}$.

Theorem 6.4 implies that the $\mathbf{p}_{i+}$ marginal of the MLE is equal to the observed marginal distribution.

The argument above may be presented with the roles of i and j interchanged, implying that the $\mathbf{p}_{+j}$ marginal of the MLE is equal to the observed marginal distribution.

Finally, because of the unicity of the MLE, it is not possible that one distribution maximizes the likelihood with only one marginal equal to the observed, and another one also maximizes the likelihood in which only the other marginal is as observed. These distributions are the same; thus both one-way marginals in the MLE are also observed. □

6.1.3 Association Models for Two-Way Tables

In the case of 2×2 tables, the most frequently used model for the association structure is the model of independence. This model may be formulated by assuming that the odds ratio is equal to 1 and was discussed in detail in Sect. 5.4. In the case of $I \times J$ tables, independence is identical to all the local or all the spanning cell odds ratios being equal to 1. In particular, when I or J or both are large, independence is very restrictive, as it specifies the values of $(I-1)(J-1)$ parameters. The so-called association models assume some kind of a simple but less restrictive structure of the local or spanning cell odds ratios.

The models presented here very briefly were introduced in [34], and [15] gave a detailed account of them. To relate the association models to the more general class of log-linear models discussed later in the book, we first state a reformulation of the model of independence in a two-way table.

Proposition 6.2. *For a probability distribution* $\mathbf{p}$ *on an* $I \times J$ *table,*

$$p(i,j) = p(i,+)p(+,j),\ i = 1,\ldots,I,\ j = 1,\ldots,J, \tag{6.17}$$

that is, independence holds if and only if there exist numbers α, β_i, γ_j, *such that*

$$p_{i,j} = \alpha\beta_i\gamma_j,\ i = 1,\ldots,I,\ j = 1,\ldots,J. \tag{6.18}$$

□

Proof. Obviously, (6.17) implies (6.18). To see the converse, sum both sides of (6.18) in i and j, to obtain that $1 = \alpha\beta_+\gamma_+$, implying that

$$\alpha = \frac{1}{\beta_+ \gamma_+}.$$

Thus

$$p(i,+) = \frac{1}{\beta_+ \gamma_+} \beta_i \gamma_+ = \frac{\beta_i}{\beta_+},$$

and similarly for $p(+,j)$, implying (6.17).

□

It may seem that the model (6.18) is more general than independence, but, in fact, it is not. Multiplicative models similar to (6.18) will be discussed repeatedly in this book. For the time being, it is enough to say that the parameter α is the same in every cell and may be called the overall effect. The parameter β_i depends on the row the cell (i,j) is in, and the parameter γ_j only depends on the column. They are called the row and column parameters, respectively. Lack of association, that is, independence, is implied because there is no term that would depend on both i and j; rather the effects of the row and of the column the cell is in, on the probability of the cell, may be separated.

The most restrictive association model assumes that the local or the spanning cell odds ratios have a constant value, but this value may be different from 1. This is called the uniform association or U model. While independence has the same formulation for local and spanning cell odds ratios, the U model is different, if stated for local or if stated for spanning cell odds ratios. If in an $I \times J$ table the local odds ratios are all equal to C, then

$$\frac{p(i,j)p(i+1,j+1)}{p(i,j+1)p(i+1,j)} = C,$$

which is different from the spanning cell odds ratios being all equal to C:

$$\frac{p(i,j)p(1,1)}{p(i,1)p(1,j)} = C.$$

For an $I \times J$ table, there are $IJ-1$ parameters that need to be estimated based on the data. The U model assumes that the $(I-1)(J-1)$ odds ratios are equal to each other; thus it requires estimating the $I+j-2$ marginal probabilities and the common value of the (local or spanning cell) odds ratios, so, in total, $I+J-1$ parameters need to be estimated, and the values of $(I-1)(J-1)-1$ odds ratios are implied by the model. Indeed,

$$I+J-1+(I-1)(J-1)-1 = I+J-1+IJ-I-J+1-1 = IJ-1.$$

The row marginal distribution, the column marginal distribution, and the common value of the odds ratio to be estimated are variation independent from each other, which justifies taking the remaining parameters as implied. This applies to all similar calculations in this subsection.

The (local or spanning cell) odds ratios of an $I \times J$ table may be arranged in an $(I-1) \times (J-1)$ array, called the table of odds ratios. The U model may be thought of as assuming that the entries in the table of odds ratios are all the same.

Obviously, whether the odds ratios or their logarithms are assumed to have a common value,[11] does not make a difference. The U model for the logarithms of the local odds ratios, assuming that they are all equal to u, is implied by the following multiplicative model:

$$p_{ij} = \alpha\beta_i\gamma_j \exp(iju). \tag{6.19}$$

When the logarithm of the local odds ratio is computed, the first three parameters in (6.19) cancel out, and one obtains

$$\log \frac{\exp(iju)\exp((i+1)(j+1)u)}{\exp(i(j+1)u)\exp((i+1)ju)} = u.$$

The row effects association model, usually denoted as the R model, assumes that, in the table of the logarithms of the odds ratios, the entries in each row are identical; thus there is an additive row effect on the logarithms of the odds ratios. This model has $I+J-2$ parameters for the marginal distributions and $I-1$ odds ratios that are estimated from the data, and the remaining $IJ-1-(I+J-2)-(I-1) = (I-1)(J-2)$ parameters are implied by the model. A form of the R model for local odds ratios similar to (6.19) is

$$p_{ij} = \alpha\beta_i\gamma_j \exp(jv_i), \tag{6.20}$$

and the logarithm of the local odds ratio is equal to

$$v_{i+1} - v_i,$$

thus it depends on the row only.

Similarly, the column effects association model (the C model) assumes an additive column effect on the logarithm of the table of odds ratios.

There are two variants of the models assuming both a row and a column effect on the table of logarithms of odds ratios. Let r_i denote the effect in the ith row and c_j the effect in the jth column of the table of odds ratios. In an R+C model, the logarithm of the odds ratio in the ith row and jth column is assumed to be equal to $r_i + c_j$, while in an RC model, the logarithm of the odds ratio is assumed to be $r_i c_j$. For example, if the model is formulated for local odd ratios, the R+C model assumes the existence of r_i and c_j values such that

$$\log \frac{p(i,j)p(i+1,j+1)}{p(i,j+1)p(i+1,j)} = r_i + c_j,\ i = 1,\dots,I-1,\ j = 1,\dots,J-1.$$

Thus in both the R+C and the RC model, one estimates $I-1+J-1+I-1+J-1$ parameters for the row marginal probabilities, the column marginal probabilities,

[11] Positivity of the probabilities is assumed here.

the row effects on the logarithms of the odds ratios, and the column effects on them, respectively. This is $2I+2J-4$, and the number of parameters implied is $IJ-1-(2I+2J-4) = IJ-2I-2J+3 = (I-2)(J-2)-1$.

The R+C model for the logarithms of the local odds ratios is implied by

$$p_{ij} = \alpha\beta_i\gamma_j \exp(jv_i + iw_j). \tag{6.21}$$

In this case, the logarithm of the local odds ratio simplifies to

$$\log \frac{\exp(jv_i + iw_j)\exp((j+1)v_{i+1} + (i+1)w_{j+1})}{\exp((j+1)v_i + iw_{j+1})\exp(jv_{i+1} + (i+1)w_j} = (v_{i+1} - v_i) + (w_{j+1} - w_j).$$

In the case of the RC model, the probabilities may be written as

$$p_{ij} = \alpha\beta_i\gamma_j \exp(v_i w_j) \tag{6.22}$$

and the logarithms of the local odds ratios in this case are

$$\log \frac{\exp(v_i w_j)\exp(v_{i+1} w_{j+1})}{\exp(v_{i+1} w_j)\exp(v_i w_{j+1})} = (v_{i+1} - v_i)(w_{j+1} - w_j).$$

Inspection of the formulas (6.19), (6.20, (6.21), and (6.22) reveals that only the last term (in the exponential function) influences the local odds ratios, and the overall and row and column effects cancel out. Therefore, the last term is related to the interaction or association (lack of independence) between the variables. The contribution of being in the ith row is i in the U and C models and depends on v_i in the R, R+C, and RC models. When the variables are ordinal, these parameters are sometimes interpreted as a score, that is, the appropriate numerical value that can be associated with category i, if the respective model is assumed. In this sense, the score of category i is itself under the U and C models and is v_i under the R, R+C, and RC models. In the latter cases, the scores are estimated from the data. Similar scores may also be determined for the categories j of the column variable. The details of this interpretation (see, e.g., [15]) will not be discussed in this book.

The effort of finding such scores is related to the view that categorical variables are poorly observed ratio level variables, a view that may or may not be appropriate in the context of a particular research problem.

There are many contributions dealing with various extensions of association models; we only mention [6] and [7].

6.2 Conditional and Higher-Order Odds Ratios

Generalizations of the odds ratio play a central role in understanding and modeling the structure of higher-dimensional contingency tables. Such generalizations are obtained via conditioning. Table 2.3 illustrated the structure of a 2 × 4 × 3 contingency table formed by the variables A, B, and C. It is obvious that for every fixed category of C, one has a two-way table of the joint distribution if A and B. This joint distribution is conditional on a category of C. Conditioning on a category of C means considering only those who belong to this category. If the variables have I, J, and K categories, respectively, and the cell probabilities[12] are denoted as $p(i,j,k)$, then the probabilities in the conditional distribution are $p(i,j,k)/p(+,+,k)$ and the odds ratios in the conditional table have the form

$$\frac{p(1,1,k)/p(+,+,k)p(2,2,k)/p(+,+,k)}{p(1,2,k)/p(+,+,k)p(2,1,k)/p(+,+,k)} = \frac{p(1,1,k)p(2,2,k)}{p(1,2,k)p(2,1,k)}. \tag{6.23}$$

In fact, (6.23) is the conditional odds ratio in the upper left hand corner of the $C=k$ conditional table. If any of I and J is greater than 2, local or spanning cell odds ratios may be defined in the conditional table, just like in the case of two-way tables. In the rest of this subsection, all variables will be assumed to be binary. The results hold in binary subtables of arbitrary contingency tables and may also be generalized to local or spanning cell conditional odds ratios.

The quantity in (6.23) is called the conditional odds ratio of A and B given $C=k$ and will be referred to as

$$COR(A,B|C=k).$$

It is clear from (6.23) that the conditional odds ratio may be determined directly from the cell probabilities; no conditional probabilities need to be computed.

The conditional odds ratio $COR(A,B|C=k)$ measures the strength of association between variables A and B for those who are in category k of variable C. The strength of association may be different for those in different categories of C. A comparison of $COR(A,B|C=1)$ to $COR(A,B|C=2)$, as the ratio

$$\frac{COR(A,B|C=1)}{COR(A,B|C=2)} \tag{6.24}$$

measures how strong is the effect of C on the association between A and B. Similarly to the odds ratio, the value 1 of (6.24) means no effect (of C on the association between A and B), and values farther away from 1 indicate stronger effects. Also, a value and its reciprocal mean the same strength of effect but in different directions. So if (6.24) is, say, 2, then the association among variables A and B is twice stronger (as measured by the odds ratio) for those who are in $C=1$, than for those who are in $C=2$. In this interpretation, effect has direction (the effect of C on the strength of association), but association is symmetric.

Similarly, $COR(A,C|B=1)$ is a measure of the strength of association among A and C for those in category 1 of B, and

$$\frac{COR(A,C|B=1)}{COR(A,C|B=2)} \tag{6.25}$$

[12] The same development is possible for cell frequencies, too.

measures the strength of effect of B on the association between A and C. There is a third related quantity,

$$\frac{COR(B,C|A=1)}{COR(B,C|A=2)}, \tag{6.26}$$

which is a measure of the strength of effect of A on the association between B and C.

Depending on the substantive problem at hand, any or all of these may be relevant. It might be the case, for instance, that with the particular data, education has a strong effect on the association between gender and income (i.e., gender and income are unrelated for people with a college degree, but gender and income are strongly associated for people without a college degree), but gender has only a weak effect on the association between educational level and income (e.g., the association between education and income has about the same strength for men and women). Somewhat surprisingly, this cannot be the case.

Proposition 6.3. *The ratios of conditional odds ratios given in (6.24), (6.25), and (6.26) are always equal.*

Proof. Using (6.23), one obtains that

$$\frac{COR(A,B|C=1)}{COR(A,B|C=2)} = \frac{p(1,1,1)p(2,2,1)}{p(1,2,1)p(2,1,1)} / \frac{p(1,1,2)p(2,2,2)}{p(1,2,2)p(2,1,2)}$$

$$\frac{COR(A,C|B=1)}{COR(A,C|B=2)} = \frac{p(1,1,1)p(2,1,2)}{p(1,1,2)p(2,1,1)} / \frac{p(1,2,1)p(2,2,2)}{p(1,2,2)p(2,2,1)}$$

$$\frac{COR(B,C|A=1)}{COR(B,C|A=2)} = \frac{p(1,1,1)p(1,2,2)}{p(1,1,2)p(1,2,1)} / \frac{p(2,1,1)p(2,2,2)}{p(2,1,2)p(2,2,1)},$$

and all these are equal to

$$\frac{p(1,1,1)p(1,2,2)p(2,1,2)p(2,2,1)}{p(1,1,2)p(1,2,1)p(2,1,1)p(2,2,2)}. \tag{6.27}$$

□

The meaning of Proposition 6.3 is that the effect of one variable on the strength of association between the other two variables does not depend on the particular choice of the variable the effect of which is being looked at. It is not a property of a particular splitting of the three variables; rather it is a characteristic of the joint distribution of the three variables. The common value, given in (6.27), is the strength of association among the three variables, called the second-order odds ratio, and is denoted as $OR(A,B,C)$.

The second-order odds ratio measures association that occurs among the three variables and cannot be attributed to any two of them. In fact, $OR(A,B,C)$, on the one hand, and the three two-way marginal distributions, on the other hand, are variation independent. This property will be further discussed in Sect. 10.1.

When one has a four-dimensional contingency table, formed by the variables A, B, C, and D, conditioning on one of the variables, say on D, yields three-dimensional conditional distributions, and the odds ratios of these distributions are conditional odds ratios in the four-way table. One has

$$COR(A,B,C|D=1) = \frac{p(1,1,1,1)p(1,2,2,1)p(2,1,2,1)p(2,2,1,1)}{p(1,1,2,1)p(1,2,1,1)p(2,1,1,1)p(2,2,2,1)}$$

and

$$COR(A,B,C|D=2) = \frac{p(1,1,1,2)p(1,2,2,2)p(2,1,2,2)p(2,2,1,2)}{p(1,1,2,2)p(1,2,1,2)p(2,1,1,2)p(2,2,2,2)}.$$

Again,

$$\frac{COR(A,B,C|D=1)}{COR(A,B,C|D=2)}$$

is a measure of how strong is the effect of variable D on the second-order association among A, B, and C, and, similarly to the previous case, it is easily seen that

$$\frac{COR(A,B,C|D=1)}{COR(A,B,C|D=2)} = \frac{COR(A,B,D|C=1)}{COR(A,B,D|C=2)} =$$
$$\frac{COR(A,C,D|B=1)}{COR(A,C,D|B=2)} = \frac{COR(B,C,D|A=1)}{COR(B,C,D|A=2)},$$

and that the common value is

$$OR(A,B,C,D) =$$

$$\frac{p(1,1,1,1)p(1,1,2,2)p(1,2,1,2)p(1,2,2,1)p(2,1,1,2)p(2,1,2,1)p(2,2,1,1)p(2,2,2,2)}{p(1,1,1,2)p(1,1,2,1)p(1,2,1,1)p(2,1,1,1)p(1,2,2,2)p(2,1,2,2)p(2,2,1,2)p(2,2,2,1)}.$$

$OR(A,B,C,D)$ is the third-order odds ratio among the four variables, and it is a measure of the strength of association among them, which cannot be attributed to any subset of them.

In general, if one has k variables, then the $k-1$st-order odds ratio has the product of 2^{k-1} cell probabilities in the numerator and the product of the same number of cell probabilities in the denominator. The probabilities in the denominator belong to the cells where the parity of 1's is the same as that of k, and the other cells are in the denominator.

Conditional odds ratios may be defined for conditional tables that are obtained by conditioning on several variables. For example, if the contingency table is formed by the variables A, B, C, D, E, all binary, then

$$COR(B,C,E|A=2,D=1) = \frac{p(2,1,1,1,1)p(2,1,2,2,1)p(2,2,1,2,1)p(2,2,2,1,1)}{p(2,1,1,2,1)p(2,1,2,1,1)p(2,2,1,1,1)p(2,2,2,2,1)}.$$

The general form of conditional odds ratios is given next.

Proposition 6.4. *If the variables forming the table are divided into disjoint groups $\mathscr{A}$ and $\mathscr{B}$ and* **b** *is a joint category of the variables in $\mathscr{B}$, then*

$$COR(\mathscr{A}|\mathscr{B}=\boldsymbol{b}) = \frac{\prod_{\boldsymbol{a}\in s} p(\boldsymbol{a},\boldsymbol{b})}{\prod_{\boldsymbol{a}\in d} p(\boldsymbol{a},\boldsymbol{b})},$$

where s is the set of joint indices **a** *of the variables in $\mathscr{A}$, for which the number of indices* 1 *has the same parity as the number of variables in $\mathscr{A}$, and d is the set of those joint indices, where the two parities are different.*

Proof. As the conditional odds ratio is the odds ratio in a conditional table, it is enough to see that the odds ratio of the variables in $\mathscr{A}$, in a table formed by the same variables, is

$$\frac{\prod_{\mathbf{a}\in s} p(\mathbf{a})}{\prod_{\mathbf{a}\in d} p(\mathbf{a})}. \tag{6.28}$$

For an induction proof, check first that (6.28) holds when $\mathscr{A} = \{A, B\}$. Indeed, in this case

$$s = \{(1,1), (2,2)\} \text{ and } d = \{(1,2), (2,1)\},$$

and (6.28) holds. Now suppose (6.28) holds for all cases when $\mathscr{A}$ contains fewer than k variables. Let $\mathscr{A}$ contain $k-1$ variables, and let Z be an additional variable. Then

$$OR(\mathscr{A} \cup \{Z\}) = \frac{COR(\mathscr{A}|Z=1)}{COR(\mathscr{A}|Z=2)} =$$

$$\frac{\frac{\prod_{\mathbf{a}\in s} p(\mathbf{a},1)}{\prod_{\mathbf{a}\in d} p(\mathbf{a},1)}}{\frac{\prod_{\mathbf{a}\in s} p(\mathbf{a},2)}{\prod_{\mathbf{a}\in d} p(\mathbf{a},2)}} = \frac{\prod_{\mathbf{a}\in s} p(\mathbf{a},1) \prod_{\mathbf{a}\in d} p(\mathbf{a},2)}{\prod_{\mathbf{a}\in s} p(\mathbf{a},2) \prod_{\mathbf{a}\in d} p(\mathbf{a},1)}.$$

The parities of k and of $k-1$ are different, and the s and d sets refer to $k-1$. In the numerator, one has $\mathbf{a} \in s$, with an added 1, so the same parity as k, and the indices $\mathbf{a} \in d$ with an added 2, so the same parity as k. Similarly, in the numerator, one has the indices with the parity of the number of indices 1 different from the parity of k. □

The definition of conditional odds ratios immediately implies the following property:

$$\frac{COR(A,B|C=1,D=1)}{COR(A,B|C=2,D=1)} = COR(A,B,C|D=1)$$

The next proposition gives this property in a general form.

Proposition 6.5. *If the variables forming the contingency table are divided into the disjoint groups $\mathscr{A}$, $\mathscr{B}$, and $\mathscr{C}$, and*

$$COR(\mathscr{A}|\mathscr{B},\mathscr{C})$$

are given for all joint categories of the variables in $\mathscr{B}$ and $\mathscr{C}$, then

$$COR(\mathscr{A},\mathscr{B}|\mathscr{C})$$

is also given for all categories of the variables in $\mathscr{C}$. □

Maximum likelihood estimates for conditional odds ratios may be obtained by plugging in the MLEs of the cell probabilities into the relevant formulas, and the δ-method may be used to derive asymptotic standard deviations of the estimates.

6.3 Independence and Conditional Independence

If two variables are independent, their joint distribution contains no information in addition to their univariate distributions. Independence is much simpler than association. This simplicity can be expressed in terms of the number of parameters that need to be estimated from the data to describe the joint distribution. If the variables are not independent and they have I and J categories, respectively, then the joint probability distribution has $IJ-1$ parameters. If the two variables are independent, the joint distribution can be described by $I+J-2$ parameters. If, say, both variables have 5 categories, under independence, one has eight parameters but with no independence, 24 parameters. For a given sample size, it is better to estimate fewer parameters. The standard error of estimating a probability p, under multinomial sampling, is $\sqrt{p(1-p)/n}$; thus the relative size of the standard error is

$$\frac{\sqrt{p(1-p)/n}}{p} = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{p}-1}.$$

If there are k probabilities to estimate, there will be at least one, which is less than or equal to $1/k$, and its relative standard error is more than $(\sqrt{k}-1)/\sqrt{n}$. If, as in the example above, there are 24 parameters to estimate, then with a sample size of $n=1000$, the lower bound for the largest relative standard error is about 0.123, but if there are only 8 parameters to be estimated, the lower bound for the largest relative standard error is about 0.058.[13]

The conceptual importance of independence, however, far exceeds its importance in providing more reliable estimates through the reduction of the number of parameters to be estimated. If A is a potential explanatory variable of B in a regression type problem, and they happen to be independent, then A has no effect on B and this is a great simplification of reality. In general, when one investigates several variables at the same time, one has to assume initially that all are related and if some

[13] In practice, both errors and relative errors are often expressed as percentages. More precisely, relative errors, including the relative standard error, are expressed in percent. For example, in the latter case, the lower bound for the largest relative standard error is about 6 percent. On the other hand, errors in absolute terms, including standard errors, should be expressed in percentage points. For example, with a sample size of $n=2500$ and simple random sampling, the standard error of the estimate of a probability is not more than 1 percentage point. Unfortunately, both percent and percentage point are denoted as %.

prove to be independent from each other,[14] one has found a simple structure that may have important substantive implications. In statistics, it is always simple structures that are being sought for. Of course, the structures considered have to be rich and flexible enough to give a good description of the substantially relevant aspects of the data and of the population behind them, but always the simplest such structures that are of interest. Simplification of the structure may be achieved by finding independences among variables.[15]

Unfortunately, one very rarely finds variables that are independent and yield the simplification described above. There is, however, a generalization of independence, which occurs more often in practice (at least, approximately) and offers similar simplification of the structure of the variables. Conditional independence of two variables given a third one occurs more often than independence because if the variables A and B are not independent, one may ask whether conditioning on any of the other variables observed makes them (conditionally) independent.

Conditional independence is defined as independence in the joint conditional distribution of two variables given a third one. If in the joint distribution of variables A, B, and C the probabilities are denoted as $p(i,j,k)$, then the joint conditional distribution of the variables A and B, given $C=k$, consists of the conditional probabilities

$$\frac{p(i,j,k)}{p(+,+,k)}$$

and conditional independence is

$$\frac{p(i,j,k)}{p(+,+,k)} = \frac{p(i,+,k)}{p(+,+,k)}\frac{p(+,j,k)}{p(+,+,k)},$$

for all i and j. Obviously, there are three different variants of conditional independence in a three-way table, which have the same structure but, in an actual data analytic situation mean, three entirely different assumptions concerning the association structure of the variables. Theorem 5.6 implies that

Proposition 6.6. *In a three-way table, conditional independence of the variables A and B, given variable C, holds if and only if for all categories k of C,* $COR(A,B|C=k)=1$. □

[14] In Sect. 10.2, generalizations of independence for more variables will be discussed that play a role very similar to independence in terms of simplifying structures. Further, several of the models discussed later in the book formulate simplifying properties which may be considered as generalizations of independence.

[15] There are many uses of statistical methods, where researchers are interested, instead of establishing that certain, theoretically possible effects do not exist, in establishing the existence of an effect unknown thus far. This may be relevant from a substantive point of view, but statistics prefers simple descriptions of reality over complex ones. Apart from the technical reasons referred to in the previous paragraph, the main reason is that assuming all variables are related – with no association considered as a special case of association – is certainly a correct starting point. If certain effects do not exist, one has a simpler yet true description of reality. On the other hand, assuming no relationships among the variables considered is certainly false. Realizing the existence of a relationship is very unlikely to turn this incorrect assumption into a correct one.

The importance of conditional independence comes from its uses in analyzing relationships among variables. The following example illustrates this role with hypothetical data. Table 6.3 shows the cross-classification of individuals according to their gender and level of income. In the hypothetical data shown, these variables are not independent. The odds ratio is $(120^*150)/(100^*110)$, which is more than 1, showing that men have a higher chance of a high income than women do. Indeed, the odds of a high versus low income is $120/110$ for men and is $100/150$ for women.[16] Based on this, one might conclude that gender and income are associated in a way that men tend to have higher incomes than women do.

Table 6.4 shows the same data split according to a third variable: educational level. In the three-way cross-classification, gender and income are independent in both conditional tables that is, gender and income are conditionally independent, given educational level. The conditional odds ratios are $(100^*30)/(60^*50)$ and $(20 * 120)/(40 * 60)$, both equal to 1.[17] Gender and income are associated, but their association disappears if they are conditioned on educational level.

The structure illustrated above can be given different interpretations, depending on the researcher's substantive knowledge concerning the variables involved. One interpretation is to say that the association between gender and income can be explained away by taking educational level into account. This means that taking education into account removes association or that association between gender and income is there because the different levels of education were not taken into account in the marginal analysis. A much stronger interpretation is that the association between gender and income is caused by differences in education. This means that if everybody had the same level of education, gender and income would be independent, just like they are when conditioned on educational level. This is a stronger interpretation than the previous one, because it tells what the researcher thinks the situation would be, if everybody had the same level of education, although one sees in the data that this is not the case. Namely, the assumption is that if there were no differences in educational levels, the data would have the independence property, just like they do within each category of educational level.

Table 6.3 Cross-classification of gender by income

	Income	
	High	Low
Men	120	110
Women	100	150

Hypothetical data

[16] The argument here is about structure, so questions related to statistical significance are avoided by assuming that the data contain the entire population of interest.

[17] The value of the conditional odds ratio remains the same, if computed from frequencies or from probabilities or from conditional probabilities.

Table 6.4 Cross-classification of gender by income, split by educational level

College			No college		
	Income			Income	
	High	Low		High	Low
Men	100	50	Men	20	60
Women	60	30	Women	40	120

Hypothetical data

Some peculiarities of inference and interpretation based on conditional and marginal tables, as above, will be considered in Chap. 9.

Another meaning of conditional independence in a three-way $I \times J \times K$ contingency table is seen by writing it in the form

$$p(i,j,k) = \frac{p(i,+,k)p(+,j,k)}{p(+,+,k)}, \text{ for } i = 1,\dots,I,\ j = 1,\dots,J,\ k = 1,\dots,K. \quad (6.29)$$

It is immediate from (6.29) that conditional independence of the first and second variables, given the third, is equivalent to

$$P(A = i|B = j, C = k) = P(A = i|C = k) \quad (6.30)$$

for all indices i, j, k. Indeed,

$$p(i,j,k) = P(A = i, B = j, C = K) =$$
$$P(A = i|B = j, C = k)P(B = j, C = k) = \frac{p(i,j,k)}{p(+,j,k)} p(+,j,k),$$

which is equal to the right-hand side of (6.29) if and only if

$$\frac{p(i,j,k)}{p(+,j,k)} = P(A = i|B = j, C = k) = P(A = i|C = k) = \frac{p(i,+,k)}{p(+,+,k)}.$$

This view of conditional independence is often interpreted by saying that the conditional distribution of A, given B and C are the same as if only C was given, so if C is known, no additional information is provided by B with respect to A. When effects are to be understood based on the analysis, the interpretation says that B may only have an effect on A through C, but it does not have a direct effect. This interpretation is of central importance in the various graphical[18] models used with causal interpretation; see later in the book.

Obviously, conditional independence if symmetric in A and B may be equivalently characterized by saying that

[18] The name graphical refers to the fact that when the variables of interest are supposed to have a joint normal distribution, many of these models may be represented by a graph. When the variables are categorical, the situation is more complex. See Sect. 1.3 for some related material and Sect. 8.4.

$$P(B=j|A=i,C=k)=P(B=j|C=k).$$

6.3.1 Maximum Likelihood Estimation Under Conditional Independence

Maximum likelihood estimates under the model of independence in a two-way table were determined in Sect. 5.4.2. In the present subsection, MLEs are obtained under the model of conditional independence. Just like in the case of independence, the MLEs will be the same, whether the observations have a multinomial or a Poisson distribution.

Theorem 6.6. *Let the observations* $\mathbf{X}$ *on an* $I \times J \times K$ *contingency table be distributed according to* $\mathscr{M}(n,\mathbf{p})$ *or* $\mathscr{P}(t\mathbf{p})$ *for some positive* t*. Then, the maximum likelihood estimate of* $\mathbf{p}$ *under the model (6.29) is*

$$\hat{p}(i,j,k)=\frac{x(i,+,k)x(+,j,k)}{nx(+,+,k)},$$

where, in the case of Poisson sampling, n *is the observed total.*

Proof. As was seen in Theorem 4.9, the kernels of the multinomial and Poisson likelihoods for $\mathbf{p}$ are identical.

One way to prove the claim is to note that conditional independence, as defined in (6.30), is a model of the type discussed in Theorem 6.4. There is no restriction on the $p(+,j,k)$ probabilities, and the restriction on the $p(i|j,k)$ conditional probabilities is that they are equal to the $p(i|k)$ conditional probabilities. Using the likelihood independence,

$$\hat{p}(+,j,k)=\frac{x(+,j,k)}{n}, j=1,\dots,J,\ k=1,\dots K.$$

As seen from the proof of Theorem 6.4, the augmented kernel of the log-likelihood that is to be maximized to find the MLE of the conditional distribution is

$$\sum_{i,j,k} x(i,j,k)\log p(i|j,k)+\sum_{j,k}\kappa_{j,k}(\sum_i p(i|j,k)-1)+\nu g(\mathbf{p}(i|j,k)),$$

with

$$g(\mathbf{p}(i|j,k))=\sum_k\sum_i\sum_j(p(i|j,k)-p(i|k))^2,$$

where $p(i|k)$ stand for $p(i,+,k)/p(+,+,k)$. This will have the same solution as

$$\sum_{i,j,k} x(i,j,k)\log p(i|k)+\sum_k \nu_k(\sum_i p(i|k)-1),$$

or

$$\sum_{i,k} x(i+k) \log p(i|k) + \sum_k \nu_k (\sum_i p(i|k) - 1).$$

For fixed i and k, the partial derivative according to $p(i|k)$ is

$$\frac{x(i,+,k)}{p(i|k)} + \nu_k.$$

To find a solution, this expression is set to zero, then both sides are multiplied by $p(i|k)$ and the resulting equations are added up for all values of i, yielding that

$$x(+,+,k) + \nu_k = 0.$$

The value of ν_k is used now to obtain the solution

$$\hat{p}(i|k) = \frac{x(i,+,k)}{x(+,+,k)},$$

which is easily seen to yield the maximum.

Another proof is based on Theorem 4.10. Write the kernel of the log-likelihood as

$$\sum_{i,j,k} x(i,j,k) \log \frac{p(i,+,k)p(+,j,k)}{p(+,+,k)} + \lambda (\sum_{i,k} p(i,+,k) - 1) + \nu (\sum_{j,k} p(+,j,k) - 1)$$
$$= \sum_{i,k} x(i,+,k) \log p(i,+,k) + \sum_{j,k} x(+,j,k) \log p(+,j,k) - \sum_k x(+,+,k) \log p(+,+,k)$$
$$+ \lambda (\sum_{i,k} p(i,+,k) - 1) + \nu (\sum_{j,k} p(+,j,k) - 1).$$

The partial derivative according to $p(i,+,k)$ for fixed i and k is

$$\frac{x(i,+,k)}{p(i,+,k)} + \lambda,$$

which, by the usual argument, implies that $\lambda = -n$, giving

$$\hat{p}(i,+,k) = \frac{x(i,+,k)}{n}$$

and similarly for the other parameters. □

Theorem 6.6 implies that the maximum likelihood estimate under the model of conditional independence of two variables given the third preserves two two-way marginals of the observed distribution $\mathbf{x}/n$. Both of these contain the conditioning variable. These two-way marginals are combined as in (6.29) to obtain a conditionally independent distribution with the same marginals.[19]

[19] This property is directly generalized by MLEs under log-linear models to be discussed later in the book.

Conditional independence may also be defined for more than two variables. For example, the conditional independence of variables A, B, C, D, given E, means that the joint conditional probabilities of $ABCD$, given E, are the products of the individual conditional probabilities:

$$P(A=i,B=j,C=k,D=l|E=m) = \tag{6.31}$$

$$P(A=i|E=m)P(B=j|E=m)P(C=k|E=m)P(D=l|E=m).$$

The following result is easy to see.

Proposition 6.7. *Let $\boldsymbol{X} \sim \mathscr{M}(n,\boldsymbol{p})$ be a k-dimensional random variable. Denote with $\hat{\boldsymbol{p}}$ the MLE of the distribution on the contingency table under the model that $X_1, X_2, \ldots, X_{k-1}$ are conditionally independent given X_k. Then, for every $(i_1, i_2, \ldots, i_k)$ joint category of $\boldsymbol{X}$,*

$$n\hat{p}(i_1, i_2, \ldots, i_k) = x(+,\ldots,+,i_k) \prod_{j=1}^{k-1} \frac{x(+,\ldots,+,i_j,+,\ldots,+,i_k)}{x(+,\ldots,+,i_k)}$$

□

6.4 Things to Do

1. Study the implications of (6.13). Generate two observed tables with the same sample size and with the same estimated odds ratio, so that in one of them, the estimated odds ratio differs significantly from 1, but not in the other.
2. Let the observed frequencies in a 2 × 2 table be 100, 200, 300, 400, in the usual order[20] of the cells. Compute the MLE of the odds ratio. Perform an asymptotic test of the hypothesis that the odds ratio is equal to 1. Construct an asymptotic 95% confidence interval for the true value of the odds ratio.
3. Repeat the previous exercise for the log of the odds ratio.
4. Is it possible that, based on the same set of data, a test for the hypothesis that the odds ratio is 1 and a test for the hypothesis that the log odds ratio is zero lead to different conclusions?
5. Develop results similar to those in Theorems 6.1 and 6.2 and Proposition 6.1 in the case of Poisson sampling.
6. Assume in a 2 × 2 table, $p(1,+) = 0.3$ and $p(+,1) = 0.8$. Determine the cell probabilities if the odds ratio is 0.1, 0.5, 1, 3.
7. Generate an example where neither the condition nor the claim of Theorem 6.4 hold.
8. Write frequencies into a 5 × 7 table so that the local odds ratios are all equal to 3. Write frequencies in the table so that the spanning cell odds ratios are all equal to 3.

[20] The order of the cells used in this book is called lexicographic order.

9. Develop a formula similar to (6.19) for the R model for spanning cell odds ratios.
10. The table of odds ratios derived from a 5×7 table has four rows. Fill in the 5×7 table so that the entries are 2, 1, 4, 2, respectively, in the rows of the table of local odds ratios.
11. Develop a formula similar to (6.20) for the C model for local odds ratios.
12. Prove that the formulas (6.19), (6.20), (6.21), and (6.22) are equivalent to the U, R, R+C, and RC models, respectively.
13. Use the data in Table 2.6 to determine the conditional odds ratio between categories 1 and 2 of A and 2 and 3 of B, given $C = 2$. Determine the second-order odds ratio between the variables when all are restricted to their first two categories.
14. Determine $COR(A,B,C|D=2)$, if $COR(A,B|C=1,D=1) = 2$, $COR(A,B|C=1,D=2) = 3$, $COR(A,B|C=2,D=1) = 4$, $COR(A,B|C=2,D=2) = 6$.
15. Prove Proposition 6.5.
16. Define conditional and higher-order odds ratios for non-binary variables by working with the local or with the spanning cell odds ratios.
17. Determine the asymptotic standard error of the MLE of $COR(A,B|C=1)$ and of $OR(A,B,C)$.
18. Someone wants the relative error of the estimates of probabilities between 0.2 and 0.8 to be less than 2%. What is the largest acceptable error?
19. The probabilities in a $4 \times 2 \times 5$ contingency table are to be estimated. The goal is to estimate those probabilities that are between 0.2 and 0.8, with relative standard errors less than 2%. What is the minimum sample size required?
20. Suppose that in a regression-type analysis, A, B, and C are explanatory variables and T is response. It turns out that A is independent of B and of T. Which one of these is more useful? Why?
21. Prove Proposition 6.6.
22. Complete the proof that was given for Theorem 6.6 using Theorem 6.4.
23. Complete the proof that was given for Theorem 6.6 by maximizing the likelihood function using Lagrange multipliers.
24. Find a real data set which is in the form of a three-dimensional contingency table. Determine the MLEs under the three possible conditional independence models.
25. Using the data in Table 6.4, determine the MLEs of the cell frequencies under the model of conditional independence of gender and income, given educational level.
26. Prove Proposition 6.7.

Chapter 7
Latent Classes and Exponential Families

Abstract In an exploratory setting, the latent class model assumes that conditioning on an unobserved class membership leads to conditional independence among the observed variables. The EM algorithm is often used to find MLEs under the latent class model, and a detailed proof of the convergence of the EM algorithm is given. As a more general theoretical framework, exponential families of probability distributions are introduced and some of their basic properties are proved. Exponential families not only give a background for many of the concepts and results in this book, e.g., marginal distributions and odds ratios are special cases of fundamental parameters in an exponential family, but also constitute the basis of most of the further developments in the book.

The latent class approach generalizes the concept of conditional independence, so that conditioning does not only take place with respect to some of the variables but also with respect to membership in constructed groups.

7.1 The Latent Class Approach

A frequently used approach to utilize conditional independence to simplify the association structure is the latent class idea; see, e.g., [51]. In this method, groups are defined in the data, so that the two (or more) variables considered are independent within each group. These groups are called latent classes, to emphasize the fact that the groups that are sought for are not necessarily defined by categories of a variable observed and not even by combinations of categories of observed variables. If this was the case, the classes would be manifest. Rather, the groups within which the two variables are independent are usually not characterized by any straightforward property, and further research may be needed to interpret the latent classes.

When a latent class model holds, the observed distribution is a mixture of the conditional distributions, and the same idea is used in other contexts with the name of finite mixture modeling. To be more specific, let A and B be two categorical variables. Their joint distribution is observed from a sample, and one would like to see whether the data could have been reasonably produced by a population distribution with k latent classes, such that within each latent class, A and B are independent. Membership in the latent classes may be described by a variable C[1] which has k categories, each indicating membership in one of the latent classes. In this formulation, the latent class problem is a missing or incomplete data problem, because the variable C was not (and usually could not be) observed in the data.

To obtain MLEs under the model of k latent classes, the so-called expectation-maximization (EM) algorithm may be used. The standard reference is [22], but variants of the algorithm had been considered before. In the context of the current application, [37] is particularly relevant.

The EM algorithm is illustrated here for multinomial sampling with the hypothetical data in Table 6.3 and $k = 2$.

The question is which 3-dimensional population distribution, with variables A, B, and C, such that A and B are conditionally independent given C, would maximize the likelihood of observing the data in Table 6.3 as its $A \times B$ marginal.[2] In the following example, the $A \times B$ marginal will be reproduced exactly.

The EM algorithm starts with an arbitrary conditionally independent probability distribution $\mathbf{q}_0$ on the $A \times B \times C$ table, for instance, with the one given in Table 7.1. This has the property that

$$q_0(i, j, k) = \frac{q_0(i, +, k)q_0(+, j, k)}{q_0(+, +, k)}.$$

In the first step, the AB marginal is set equal to the probability distribution observed in Table 6.3, say $\mathbf{p}$, so that

$$q_1(i, j, +) = p(i, j)$$

and the conditional distribution of C, given the other two variables, is preserved from $\mathbf{q}_0$, as shown in Table 7.2:

$$q_1(i, j, k) = p(i, j)q_0(i, j, k)/q_0(i, j, +).$$

The second step imposes conditional independence on q_1, which computes the MLE of a conditionally independent distribution, with q_1 playing the role of the observed distribution:

[1] It has to be emphasized that C is not a variable in the sense that it would describe any meaningful characteristic of the observations other than class membership. In this sense, a latent class problem is not a latent variable problem. A latent variable is one that has substantive meaning, and its existence can be conceived before the data were observed, but for some reason information on this variable was not collected. The variable indicating latent class membership is, to the contrary, derived from the observed variables.

[2] The question could be formulated a bit differently: which 3-dimensional (complete) data could be observed from a conditionally independent distribution with the highest likelihood, which have Table 6.3 as their $A \times B$ marginal. The two formulations yield the same maximized likelihood, see Sect. 7.2.

$$q_2(i,j,k) = \frac{q_1(i,+,k)q_1(+,j,k)}{q_1(+,+,k)},$$

These two steps are iterated during the EM algorithm. The E step[3] is

$$q_{2l+1}(i,j,k) = p(i,j)q_{2l}(i,j,k)/q_{2l}(i,j,+),$$

for any l, and the M step is

$$q_{2l+2}(i,j,k) = \frac{q_{2l+1}(i,+,k)q_{2l+1}(+,j,k)}{q_{2l+1}(+,+,k)},$$

for any l.

Table 7.1 The starting distribution $\mathbf{q}_0$ of the EM algorithm for the data in Table 6.3

Latent class 1			Latent class 2			Marginal		
	B = 1	B = 2		B = 1	B = 2		B = 1	B = 2
A = 1	0.05	0.1	A = 1	0.1	0.1	A = 1	0.15	0.2
A = 2	0.15	0.3	A = 2	0.1	0.1	A = 2	0.25	0.4

Table 7.2 The result of the first step of the EM algorithm, $\mathbf{q}_1$, for the data in Table 6.3

Latent class 1			Latent class 2			Marginal		
	B = 1	B = 2		B = 1	B = 2		B = 1	B = 2
A = 1	0.0833	0.1146	A = 1	0.1667	0.1146	A = 1	0.25	0.2292
A = 2	0.125	0.2344	A = 2	0.0833	0.07813	A = 2	0.2083	0.3125

Table 7.3 contains $\mathbf{q}_2$. The $A \times B$ marginal of $\mathbf{q}_2$ is closer to the observed distribution (which is the same as the $A \times B$ marginal of $\mathbf{q}_1$) than the $A \times B$ marginal of $\mathbf{q}_0$. The procedure converges if it seizes to change the distribution during iteration, up to the level of precision given. In the present case, the distribution does not change in the first four digits after the decimal point after about 50 iterations. The distribution $\mathbf{q}_{52}$ is given in Table 7.4. One sees that the algorithm converged to the distribution in Table 6.3 as its $A \times B$ marginal. However, neither the marginal distribution of the latent class membership nor the conditional distributions within the latent classes are equal to what was seen with the educational level variable. In fact, the result of the application of the EM algorithm is not unique and may depend on the starting distribution. This is illustrated using the starting distribution given in Table 7.5. The algorithm converges to the distribution given in Table 7.6.

Because there are usually no variables in the data set that would produce the same conditional distribution as the result of the EM algorithm, the interpretation of the latent classes is usually based on the marginal probabilities of the two variables in

[3] The names of the two steps of the EM algorithm will be clarified in the next subsection.

Table 7.3 The result of the second step of the EM algorithm, $\mathbf{q}_2$, for the data in Table 6.3

Latent class 1			Latent class 2			Marginal		
	B = 1	B = 2		B = 1	B = 2		B = 1	B = 2
A = 1	0.0740	0.1239	A = 1	0.1588	0.1224	A = 1	0.1881	0.2464
A = 2	0.1343	0.2250	A = 2	0.0912	0.0703	A = 2	0.2255	0.2953

Table 7.4 The EM algorithm converged up to four digits after the decimal point: $\mathbf{q}_{52}$

Latent class 1			Latent class 2			Marginal		
	B = 1	B = 2		B = 1	B = 2		B = 1	B = 2
A = 1	0.0540	0.1181	A = 1	0.1959	0.1112	A = 1	0.2499	0.2292
A = 2	0.1201	0.2622	A = 2	0.0883	0.0501	A = 2	0.2084	0.3124

Table 7.5 Another starting distribution of the EM algorithm

Latent class 1			Latent class 2			Marginal		
	B = 1	B = 2		B = 1	B = 2		B = 1	B = 2
A = 1	0.1845	0.1845	A = 1	0.0312	0.0312	A = 1	0.2187	0.2187
A = 2	0.1845	0.1845	A = 2	0.0933	0.0933	A = 2	0.2808	0.2808

Table 7.6 The limiting distribution yielded by the EM algorithm with the starting values in Table 7.5

Latent class 1			Latent class 2			Marginal		
	B = 1	B = 2		B = 1	B = 2		B = 1	B = 2
A = 1	0.2411	0.1923	A = 1	0.0089	0.0369	A = 1	0.25	0.2292
A = 2	0.1647	0.1314	A = 2	0.0436	0.1811	A = 2	0.2083	0.3125

the latent classes. To see the differences of the solutions given in Tables 7.4 and 7.6, Table 7.7 summarizes the properties of the latent classes found. The first solution has two latent classes of about the same sizes, and in both, the two categories of the variables A and B have probabilities close to one third and two thirds. In the first latent class, the second categories of both variables are about twice as likely as the first categories. In the second latent class, the first categories are about twice as likely as the second categories. Using the meanings of the variables and categories, the first latent class contains twice as many women as men and twice as many individuals with low income as with high income. On the other hand, the second latent class has twice as many men as women and twice as many individuals with high as with low income. In addition to the two variables being independent within the two latent classes, the first latent class is dominated by women with low income and the second latent class by men with high income. This solution may be interpreted as finding two classes, out of which in the first one, individuals tend to have low income, and in the second one, individuals tend to have high income. Further, the individuals in the first class tend to be women, and they tend to be men in the second one. In the second solution, the first latent class is much larger than the second one and contains about 50% more men than women and about 25% more individuals with high than with low income. The second, smaller class consists mostly of women with low income. The second solution may be interpreted as splitting the data into a dominant part, which does not contain some of the women and where there is a somewhat higher chance of a high than of a low income, and into a small part containing mostly women, where low income is typical.

Table 7.7 Summary of the latent classes in Tables 7.4 and 7.6

	Solution in Table 7.4		Solution in Table 7.6	
Conditional probability	LC 1	LC 2	LC 1	LC 2
P(A = 1)	0.3124	0.6891	0.5941	0.1692
P(A = 2)	0.6896	0.3109	0.4059	0.8308
P(B = 1)	0.3140	0.6379	0.5564	0.1940
P(B = 2)	0.6860	0.3621	0.4436	0.8060
LC probability	0.5545	0.4455	0.7295	0.2705

Very often, a central question of the latent class analysis is to find out how many latent classes need to be assumed. Of course, this question refers to the population underlying the data, not to the data themselves. When an exact solution with k latent classes, as seen above, cannot be achieved, the hypothesis of k latent classes is tested by maximizing the likelihood in the sense described in the next subsection, and comparing the observed marginal distribution to its MLE. Details of the estimation procedure are given in the next subsection.

The idea of latent class analysis may be generalized to any model that assumes some kind of simple structure. While the whole population may not be described by the model of interest, it may be divided into latent classes so that the model holds for each latent class.

7.1.1 Convergence of the EM Algorithm

In this subsection, the EM algorithm will be investigated in a slightly more general setting. The EM algorithm is used in certain missing data situations to obtain estimates under various models. The estimates obtained, as it will be seen, have some likelihood maximization property, but whether or not they are MLEs depends on the model, the sample space, and the missing data pattern. A detailed discussion of the convergence properties of the EM algorithm is given in [92].

Let us assume that the sample space is of the structure of an $I \times J$ contingency table, with observations for variables T and U, respectively. Both T and U may be combinations of other variables. The missing data structure under which the EM algorithm may be applied is that T is observed but U is not (U stands for unobserved). This is equivalent to observing a marginal distribution of the contingency table. For simplicity, it is assumed that the sampling distribution of $\mathbf{X}_{T,U}$, if complete data were available, is multinomial, and thus the actual observations $\mathbf{X}_T$ also have a multinomial distribution. The goal of the analysis is to obtain estimates of the parameters of a model M for the distribution on the entire table and test the model by comparing the T marginal of the MLE with the observed T marginal. In the specific situation in Sect. 7.1, T is $A \times B$, U is C, and M is conditional independence of A and B, given C.

The properties of the algorithm, of course, do not depend on the justification one associates with it, but there is one interpretation that makes the understanding of the algorithm and its properties relatively easy. The algorithm may be seen as maximizing a function derived from the likelihood function. First, this function is developed, and then its maximization is described.

The likelihood function is usually interpreted as a function of (parameters of) a distribution, with the observed data fixed. If one observed the data $\mathbf{x}_{T,U}$, then for any distribution $\mathbf{r}$ on the contingency table, the kernel of the multinomial log-likelihood would be

$$L(\mathbf{x}_{T,U}, \mathbf{r}_{T,U}) = \sum_{t,u} x_{t,u} \log r_{t,u},$$

where the summation goes for all possible categories of T and of U, that is, for the cells of the table. In the present case, $\mathbf{x}_{T,U}$ is not observed, only the $\mathbf{x}_T$ marginal frequencies. If the true distribution that generates the data is $\mathbf{s}$, and

$$\mathbf{X}_{T,U} \sim \mathscr{M}(n, \mathbf{s}),$$

then Theorem 2.3 implies that $\mathbf{X}_{+,U}|\mathbf{x}_T$ has a product multinomial distribution, and with the conditional probabilities

$$s_{u|t} = \frac{s(t,u)}{s(t,+)},$$

the conditional expectations of the complete data frequencies, given that the marginal frequencies $\mathbf{x}_T$ were observed, are

$$EX_{t,u}|\mathbf{x}_T = s_{u|t} x_t, \tag{7.1}$$

for every t and u.

The kernel of the log-likelihood L will be evaluated at $\mathbf{x}_{t,u} = E\mathbf{X}_{t,u}|\mathbf{x}_T$. To emphasize the dependence of this function on $\mathbf{s}$ and $\mathbf{r}$, the following notation will be used:

$$L(\mathbf{s}, \mathbf{r}) = \sum_{t,u} EX_{t,u}|\mathbf{x}_T \log r_{t,u} = \sum_{t,u} x_t s_{u|t} \log r_{t,u}. \tag{7.2}$$

The goal of the EM algorithm is to maximize (7.2). If (7.2) is maximized by $\hat{\mathbf{r}} \in M$ and $\hat{\mathbf{s}}$, so that

$$L(\hat{\mathbf{s}}, \hat{\mathbf{r}}) > L(\mathbf{s}, \mathbf{r}), \text{ for all } \mathbf{r} \in M, \mathbf{r} \neq \hat{\mathbf{r}}, \mathbf{s} \neq \hat{\mathbf{s}}$$

then the unrestricted MLE of the probability of the true distribution in cell (t,u) is

$$\frac{1}{n} x_t \hat{s}_{u|t}(t,u)$$

and the MLE of the probability assuming model M in cell (t,u) is

$$\hat{r}_{t,u}(t,u).$$

In other words, the procedure estimates not only the distribution in the model that would have generated the data with the highest likelihood but also the unobserved part of the data.

The EM algorithm is initiated with arbitrary[4] $\mathbf{r}^{(0)} \in M$. The E step of the algorithm is

$$\mathbf{s}^{(l)} = \mathbf{x}_T \mathbf{r}^{(l)}_{U|T}, \tag{7.3}$$

which determines the expectation in (7.1), assuming that $\mathbf{s} = \mathbf{r}$, and the M step of the algorithm[5] is

$$\mathbf{r}^{(l+1)} \text{ is the MLE of } \mathbf{s}^{(l)} \text{ in } M, \tag{7.4}$$

which is a maximization of the likelihood.

Theorem 7.1. *The kernel of the log-likelihood in (7.2) monotone increases during the EM algorithm in the following sense:*

$$L(\boldsymbol{s}^{(l)}, \boldsymbol{r}^{(l)}) \geq L(\boldsymbol{s}^{(l-1)}, \boldsymbol{r}^{(l)}),$$

and

$$L(\boldsymbol{s}^{(l)}, \boldsymbol{r}^{(l+1)}) \geq L(\boldsymbol{s}^{(l)}, \boldsymbol{r}^{(l)})$$

for all nonnegative integers l.

Proof. The first inequality is implied by the fact that for a given distribution $\mathbf{p}$, the likelihood of a data set is increased if the conditional distribution of some variables, given the others in the data, is replaced by the relevant conditional distribution implied by $\mathbf{p}$. To see this, write the likelihood for some data set $\mathbf{y}$ as

$$\sum_{t,u} y(t,u) \log p(t,u) = \sum_{t,u} y(t,u) \log p(t,+) + \sum_{t,u} y(t,u) \log p(u|t)$$

$$= \sum_{t} y(t,+) \log p(t,+) + \sum_{t} y(t,+) \left(\sum_{u} y(u|t) \log p(u|t) \right).$$

Then, the last term is maximized (and the likelihood is increased) if, for every t and u,

$$y(u|t) = np(u|t).$$

In the present case, $\mathbf{s}^{(l)}$ has the same $U|T$ conditional distribution for every $T = t$, as $\mathbf{r}^{(l)}$ does, as seen from (7.3), thus maximizing the kernel of the log-likelihood L from among distributions of the form $\mathbf{x}_T \mathbf{s}_{U|T}$, and $\mathbf{x}_T \mathbf{s}^{(l-1)}_{U|T}$ is one of these distributions.

[4] In this subsection upper indices are used.

[5] In the latent class context discussed in the previous section, this step imposes the conditional independence assumed by the model.

To see the second inequality, note that $\mathbf{r}$ is always in M, and the M step yielding $\mathbf{r}^{(l+1)}$ maximizes the likelihood for (the expectation of the data given by) $\mathbf{s}^{(l)}$ from among all distributions in M, including $\mathbf{r}^{(l)}$. □

Theorem 7.1 implies the following result:

Theorem 7.2. *The sequence*

$$L(\boldsymbol{s}^{(l)}, \boldsymbol{r}^{(l)})$$

is monotone increasing and converges as $l \to \infty$.

Proof. The likelihood is a probability, and its value is not more than 1. The kernel is obtained by omitting from the likelihood a multiplier, which, in the case of multinomial likelihoods, is greater than 1. Thus, the kernel of the log-likelihood is nonpositive, that is, $L(\mathbf{s}^{(l)}, \mathbf{r}^{(l)})$ is bounded from above. It is implied by Theorem 7.1 that

$$L(\mathbf{s}^{(l+1)}, \mathbf{r}^{(l+1)}) \geq L(\mathbf{s}^{(l)}, \mathbf{r}^{(l+1)}) \geq L(\mathbf{s}^{(l)}, \mathbf{r}^{(l)}),$$

thus the sequence is also monotone increasing and converges. □

In general, the limit of the EM algorithm may depend on the starting distribution $\mathbf{r}^{(0)}$ used. For practical purposes, if $\mathbf{r}$ and $\mathbf{s}$ get close enough (i.e., closer than a pre-specified threshold), the algorithm may be considered as having converged, and either distribution may be taken as estimates including the unobserved data. When no convergence in this practical sense occurs, one only knows that

$$L(\mathbf{s}^{(l)}, \mathbf{r}^{(l)}) \leq \lim_{l \to \infty} L(\mathbf{s}^{(l)}, \mathbf{r}^{(l)}) \tag{7.5}$$

and for an l, that is, large enough so that the left hand side of (7.5) is close enough to its right-hand side,[6] the T marginal of $n\mathbf{r}^{(l)}$, $n\mathbf{r}_T^{(l)}$, may be used for a comparison with x_T, using the likelihood ratio statistic. If the hypothesis that x_T was observed from the distribution $n\mathbf{r}_T^{(l)}$ is not rejected, one may consider the model M as fitting, and $\mathbf{r}^{(l)}$ may be considered as an estimate of the true distribution.

7.2 Exponential Families of Probability Distributions

Many of the results presented in the book are better understood in the context of exponential families of probability distributions. To define such families, let $\mathscr{S}$ be a sample space. In the most important applications, $\mathscr{S}$ is either a contingency table or the (multidimensional) Euclidean space or the Cartesian product of such spaces. Let[7]

$$\mathbf{t} : \mathscr{S} \to \mathbb{R}^k$$

[6] Of course, finding such and l is always heuristic, in the sense that it cannot be based on a mathematical fact.

[7] The notation used in this section is unrelated to that of the previous one.

be a measurable function.[8] Let also $\boldsymbol{\Theta} \subseteq \mathbb{R}^k$ be a subset, which is often assumed to be open, and $\mathbf{q}$ a distribution[9] on $\mathscr{S}$. The distribution $\mathbf{q}$ may be a probability distribution, but it does not have to be. When $\mathscr{S}$ is a contingency table, $\mathbf{q}$ is usually the uniform distribution, either normed to be a probability distribution or just taking the value of 1 in every cell. When $\mathscr{S}$ is the Euclidean space, $\mathbf{q}$ is most often the Lebesque measure.[10] Consider

$$p(\mathbf{s}) = q(\mathbf{s}) \frac{\exp(\mathbf{t}(\mathbf{s}), \boldsymbol{\theta})}{\int_{\mathscr{S}} \exp(\mathbf{t}(\mathbf{s}), \boldsymbol{\theta}) d\mathbf{q}(\mathbf{s})} = q(\mathbf{s}) a(\boldsymbol{\theta}) \exp(\mathbf{t}(\mathbf{s}), \boldsymbol{\theta}), \tag{7.6}$$

where $(.,.)$ denotes the inner product of two vectors and the integral, for a discrete sample space, is a sum weighted by $\mathbf{q}(\mathbf{s})$. Then, (7.6) may be interpreted as a probability distribution on $\mathscr{S}$. To be more specific, $\mathbf{p}$ is a probability distribution when $\mathscr{S}$ is a contingency table or, more generally, when the sample space is discrete, and it is a density function with respect to $\mathbf{q}$, when $\mathscr{S}$ is a Euclidean space. In the representation (7.6), $\mathbf{t}$ is called the canonical statistic, $\boldsymbol{\theta}$ the canonical parameter, and $\mathbf{q}$ the dominating measure. The latter name is justified by the fact that if for an $\mathbf{s} \in \mathscr{S}$, $q(\mathbf{s}) = 0$, then also $p(\mathbf{s}) = 0$.[11]

As an example, let $\mathscr{S}$ be a 2×2 contingency table, and let $\mathbf{q}$ be the uniform distribution (not normalized, $q(\mathbf{s}) = 1$) and $k = 9$. Define $\mathbf{t}$ as is given in Table 7.8. The components of $\mathbf{t}$ are indicators in this case. For example, t_2 is the indicator of the first row, t_5 is the indicator of the second column, and t_8 is the indicator of cell $(2,1)$. Any positive probability distribution $p(i,j)$ on the table may be written in the form of (7.6), by choosing the canonical parameter as follows:

$$\theta_1 = \frac{1}{4} \sum_{i,j} \log p(i,j) \tag{7.7}$$

$$\theta_2 = \frac{1}{2} \sum_{j} \log p(1,j) - \theta_1$$

$$\theta_3 = \frac{1}{2} \sum_{j} \log p(2,j) - \theta_1$$

$$\theta_4 = \frac{1}{2} \sum_{i} \log p(i,1) - \theta_1$$

[8] Measurability holds when $\mathscr{S}$ is a contingency table. When $\mathscr{S}$ is the multidimensional Euclidean space, measurability means that for any open set in $\mathbb{R}^k$, the points in $\mathscr{S}$ whose images are in this set also form an open set. This condition will hold for all examples in the book and will not be mentioned in the sequel.

[9] In a more precise presentation, a probability distribution would be distinguished from the vector of the probabilities it is made up from, in the discrete case, and from its density function, in the continuous case, but this distinction is not going to be made here.

[10] Intuitively, this may be interpreted as the mathematical construct equivalent to the everyday concept of volume. In this case, $\mathbf{q}$ is not a vector of probabilities, of course, rather a possibly multidimensional density function.

[11] In general, $\mathbf{q}$ dominates $\mathbf{p}$ in the measure theoretical sense.

$$\theta_5 = \frac{1}{2}\sum_i \log p(i,2) - \theta_1$$
$$\theta_6 = \log p(1,1) - \theta_1 - \theta_2 - \theta_4$$
$$\theta_7 = \log p(1,2) - \theta_1 - \theta_2 - \theta_5$$
$$\theta_8 = \log p(2,1) - \theta_1 - \theta_3 - \theta_4$$
$$\theta_9 = \log p(2,2) - \theta_1 - \theta_3 - \theta_5.$$

In this representation,

$$\log p(i,j) = (\mathbf{t}, \boldsymbol{\theta}),$$

thus, the normalizing constant $a(\boldsymbol{\theta})$ is 1. The parameters used are the average log probability (θ_1), the difference of the average of the log probabilities in the first row and of the average log probability (θ_2), and similarly for $\theta_3, \theta_4, \theta_5$. These parameters express row and column effects on the log probabilities, similar to those used in a two-way analysis of variance. The further parameters express how much the log probability in a cell differs from the overall mean plus the relevant row effect plus the relevant column effect. In fact,

$$\theta_6 = \log p(1,1) - \theta_1 - \theta_2 - \theta_4 =$$

$$\log p(1,1) - \frac{1}{4}\sum_{i,j}\log p(i,j) - \frac{1}{2}\sum_j \log p(1,j) + \theta_1 - \frac{1}{2}\sum_i \log p(i,1) + \theta_1 =$$

$$\frac{1}{4}(\log p(1,1) - p(1,2) - p(2,1) + p(2,2)) = \frac{1}{4}\log\frac{p(1,1)p(2,2)}{p(1,2)p(2,1)},$$

thus θ_6 is one quarter of the logarithm of the odds ratio and is zero if and only if the two variables forming the table are independent. It is easy to see that

$$\theta_2 = -\theta_3,\ \theta_4 = -\theta_5,\ \theta_6 = -\theta_7 = -\theta_8 = \theta_9. \tag{7.8}$$

Note that (7.7) gives formulas to determine $\boldsymbol{\theta}$ from the distribution $\mathbf{p}$ but does not address the question whether the choice of $\boldsymbol{\theta}$ is unique. This is also not immediate from (7.6). On the other hand, (7.6) shows that if $\boldsymbol{\theta}$ was a parameter, it would also be a parameterization, that is, it is invertible.

We have seen that all positive probability distributions on a 2 × 2 table form an exponential family. However, (7.8) suggests that the choice of the dimension k is not unique.

Table 7.8 Definition of the canonical statistic for a 2 × 2 table

$\mathbf{t}(1,1) = (1,1,0,1,0,1,0,0,0)'$	$\mathbf{t}(1,2) = (1,1,0,0,1,0,1,0,0)'$
$\mathbf{t}(2,1) = (1,0,1,1,0,0,0,1,0)'$	$\mathbf{t}(2,2) = (1,0,1,0,1,0,0,0,1)'$

Indeed, another exponential family representation of the form (7.6) is possible, with a 4-dimensional parameter. The canonical statistic of this representation is given in Table 7.9. In this case, the parameter is related to the probability distribution as follows:

$$\begin{aligned} \theta_1 &= \frac{1}{4}\sum_{i,j} \log p(i,j) \qquad (7.9) \\ \theta_2 &= \frac{1}{2}\sum_{j} \log p(1,j) - \theta_1 \\ \theta_3 &= \frac{1}{2}\sum_{i} \log p(i,1) - \theta_1 \\ \theta_4 &= \log p(1,1) - \theta_1 - \theta_2 - \theta_4 \end{aligned}$$

Again, the two variables forming the table are independent if and only if $\theta_4 = 0$. In this case,

$$\theta_4 = \frac{1}{4} \log \frac{p(1,1)p(2,2)}{p(1,2)p(2,1)}.$$

In either representation, the odds ratio, which was discussed in the previous chapter as a measure of association, is (essentially) one of the canonical parameters. This implies that the two variables forming the 2×2 table are independent, if and only if their joint distribution has a representation of the form (7.6) with the statistic given in Table 7.10. The parameters are as the first 3 in (7.9). In general, from an exponential family, another one may be obtained by fixing the values of some of the canonical parameters.

Another example of exponential families is all normal distributions with fixed, say 0, expectation. Here, $k = 1$, the canonical statistic is

$$t(x) = \frac{-x^2}{2},\ x \in \mathbb{R}$$

Table 7.9 Another definition of the canonical statistic for a 2×2 table

$\mathbf{t}(1,1) = (1,1,1,1)'$	$\mathbf{t}(1,2) = (1,1,-1,-1)'$
$\mathbf{t}(2,1) = (1,-1,1,-1)'$	$\mathbf{t}(2,2) = (1,-1,-1,1)'$

Table 7.10 Definition of the canonical statistic for an independent 2×2 table

$\mathbf{t}(1,1) = (1,1,1)'$	$\mathbf{t}(1,2) = (1,1,-1)'$
$\mathbf{t}(2,1) = (1,-1,1)'$	$\mathbf{t}(2,2) = (1,-1,-1)'$

and with this, the density of a normal distribution at $x \in \mathbb{R}$ is

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x^2}{2\sigma^2}\right). \tag{7.10}$$

Here, the canonical parameter is $1/\sigma^2$.

Sometimes, a definition more general than the one given in (7.6) is used. In this definition, the parameter $\boldsymbol{\theta}$ is replaced by a function of it, say $\boldsymbol{\tau}(\boldsymbol{\theta})$:

$$p(\mathbf{s}) = q(\mathbf{s})\frac{\exp(\mathbf{t}(\mathbf{s}), \boldsymbol{\tau}(\boldsymbol{\theta}))}{\int_{\mathscr{S}} \exp(\mathbf{t}(\mathbf{s}), \boldsymbol{\tau}(\boldsymbol{\theta}))d\mathbf{q}(\mathbf{s})} = q(\mathbf{s})a(\boldsymbol{\theta})\exp(\mathbf{t}(\mathbf{s}), \boldsymbol{\tau}(\boldsymbol{\theta})). \tag{7.11}$$

With this more general definition, the family is obtained by considering all values of $\boldsymbol{\theta}$, for which (7.11) exists. For example, in (7.10), the choices $\theta = 1/\sigma^2$ and $\tau(\theta) = \theta$, or $\theta = 1/\sigma$ and $\tau(\theta) = \theta^2$, are equally possible.

When the definition in (7.11) is used, the form in (7.6) is called canonical (of course, $\boldsymbol{\theta}$ is different in the two representations). Obviously, every exponential family, even if given in the form (7.11), may be written in canonical form, by considering $\boldsymbol{\tau}$ as the parameter, instead of $\boldsymbol{\theta}$.

Exponential families have many interesting and useful properties. Already the examples seen so far illustrate that the parameters of exponential families may be selected to reveal the structure of the distributions in the family. Many of the models used in this book, in particular log-linear models, are exponential families, and as these models will be developed, their properties as exponential families will be worked out in detail. A general reference for exponential families is the book [5].

An exponential family is regular, if the parameter space $\boldsymbol{\Theta}$ is an open set. This will be assumed to hold in the rest of this section. Also, in the rest of this section, we will only consider exponential families on finite sample spaces, although the results apply generally.

As has been seen before, the same exponential family may be represented using parameters of different dimensions, using, of course, different canonical statistics. Out of the dimensions, in which an exponential family may be represented, there is a minimal one. This is called the dimension of the exponential family. Representing the same family with different parameters is called reparameterization and is best understood in terms of the design matrix of the exponential family. The columns of the design matrix are the components of $\mathbf{t}$. For example, with the statistic in Table 7.9, the design matrix is in lexicographic order of the cells of the table.

$$\mathbf{T} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

With this design matrix,[12] the members of the exponential family may be written as

$$\log \mathbf{p} = \mathbf{T}\boldsymbol{\theta} + \log \mathbf{q} + \log a(\boldsymbol{\theta})\mathbf{1}, \tag{7.12}$$

where $\mathbf{1}$ is a vector of 1s.

The design matrix based on the canonical statistic in Table 7.8 is the following:

$$\mathbf{T}_{1'} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Obviously, the rows of $\mathbf{T}_{1'}$ are not linearly independent. Therefore, (7.12) does not have a unique solution in $\boldsymbol{\theta}$ for given $\mathbf{p}$, if $\mathbf{T}_1$ is used instead of $\mathbf{T}$. In such a case, some authors still call $\boldsymbol{\theta}$ a parameter, though unidentifiable. In this book, only identifiable parameters are called parameters.

It is clear that the column spaces of the matrices $\mathbf{T}$ and $\mathbf{T}_1$ are the same, so if $\boldsymbol{\theta}_1$ denotes the parameter in the representation based on Table 7.8, then for each $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1$, there should be a $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, such that

$$\mathbf{T}\boldsymbol{\theta} = \mathbf{T}_1\boldsymbol{\theta}_1,$$

which is possible if $\mathbf{T}$ has an inverse. In that case,

$$\boldsymbol{\theta} = \mathbf{T}^{-1}\mathbf{T}_1\boldsymbol{\theta}_1, \tag{7.13}$$

is the result of the reparameterization. The next theorem shows that the reparameterization in (7.13) to the minimal dimension is always possible.

Theorem 7.3. *If the exponential family is in the form (7.6), so that the dimension of the parameter is the dimension of the exponential family, then the components of the canonical statistic are linearly independent, and the design matrix is invertible.*

Proof. Assume, to the contrary, that the components of the canonical statistic are linearly dependent, so that, with, say, c components

$$t_c(\mathbf{s}) = \sum_{i=1}^{c-1} e_i t_i(\mathbf{s}),$$

[12] The design matrix is not necessarily symmetric, as it is in the present example.

with some coefficients e_i, for all $\mathbf{s} \in \mathbf{S}$. Then

$$(\mathbf{t}(\mathbf{s}), \boldsymbol{\theta}) = \sum_{i=1}^{c} t_i(\mathbf{s})\theta_i = \sum_{i=1}^{c-1} t_i(\mathbf{s})\theta_i + e_i t_i(\mathbf{s})\theta_c = \sum_{i=1}^{c-1} t_i(\mathbf{s})\eta_i,$$

for some parameter $\boldsymbol{\eta}$. This means that under the assumption, the space spanned the first $c-1$ components of $\mathbf{t}$, is the same as the one spanned by all its components, and the parameterization is not of minimal dimension.

The second claim of the theorem is implied by the columns of the design matrix, which are the components of the canonical statistic, being linearly independent.

□

The rest of the section concentrates on the properties of maximum likelihood estimates under exponential families.

Theorem 7.4. *If the statistical model is a regular exponential family of the form (7.6), so that the dimension of the parameter is the dimension of the exponential family, and the observations have a multinomial or Poisson distribution, further the derivative of the likelihood function is zero for a parameter $\hat{\boldsymbol{\theta}}$, then the distribution $\hat{\mathbf{p}} = \mathbf{p}(\hat{\boldsymbol{\theta}})$ is the MLE of the true distribution.*

Proof. An exponential family is a parametric model, as discussed in Sect. 4.1.3, and in view of Theorem 4.10, it is sufficient to show that $\hat{\boldsymbol{\theta}}$ does maximize the likelihood, which is implied if the second derivative matrix of the log-likelihood function is negative definite at $\hat{\boldsymbol{\theta}}$.

For convenience, write the probability function as

$$q(\mathbf{s})a(\boldsymbol{\theta})\exp(\mathbf{t}(\mathbf{s}), \boldsymbol{\theta}) = q(\mathbf{s})\exp\left((\mathbf{t}(\mathbf{s}), \boldsymbol{\theta}) - \log\frac{1}{a(\boldsymbol{\theta})}\right)$$

and then the kernel of the log-likelihood,[13] which is the same under the two sampling schemes (see Theorem 4.9) is

$$\sum_{\mathbf{s}} x(\mathbf{s})\left((\mathbf{t}(\mathbf{s}), \boldsymbol{\theta}) - \log\frac{1}{a(\boldsymbol{\theta})}\right). \tag{7.14}$$

The partial derivative of $\log\frac{1}{a(\boldsymbol{\theta})}$ according to θ_i, for a fixed i, is obtained as

$$\frac{\partial \log a(\boldsymbol{\theta})}{\partial \theta_i} = a(\boldsymbol{\theta})\frac{\partial \frac{1}{a(\boldsymbol{\theta})}}{\partial \theta_i}$$

$$= a(\boldsymbol{\theta})\frac{\partial \sum_{\mathbf{s}} q(\mathbf{s})\exp(\mathbf{t}(\mathbf{s}), \boldsymbol{\theta})}{\partial \theta_i} = a(\boldsymbol{\theta})\sum_{\mathbf{s}} q(\mathbf{s})\exp(\mathbf{t}(\mathbf{s}), \boldsymbol{\theta})t_i(\mathbf{s})$$

$$= E_{\mathbf{p}(\boldsymbol{\theta})} t_i.$$

[13] The kernel of the log likelihood does not need to be augmented with a Lagrangian term in this case, because of the presence of the normalizing constant.

With this, the derivative of the kernel of the log-likelihood is

$$\sum_{\mathbf{s}} x(\mathbf{s})(t_i(\mathbf{s}) - E_{\mathbf{p}(\boldsymbol{\theta})} t_i) = E_{\mathbf{x}} t_i - n E_{\mathbf{p}(\boldsymbol{\theta})} t_i, \tag{7.15}$$

with a bit of abuse of the notation for the expected value.[14]

An element of the second derivative matrix is obtained by considering an element of the first partial derivative vector, say the one obtained as the derivative according to θ_i, and taking its partial derivative according to θ_j, where j may be equal to i or different from it. One obtains that the (i, j) element of the second partial derivative matrix of the kernel of the log-likelihood is, using instead of (7.15), the detailed form

$$\frac{\partial \sum_{\mathbf{u}} x(\mathbf{u})(t_i(\mathbf{u}) - a(\boldsymbol{\theta}) \sum_{\mathbf{s}} q(\mathbf{s}) \exp(\mathbf{t}(\mathbf{s}), \boldsymbol{\theta}) t_i(\mathbf{s}))}{\partial \theta_j},$$

where $\mathbf{u}$ is also a cell of the table (an element of the sample space). This is equal to

$$-\frac{\partial \sum_{\mathbf{u}} x(\mathbf{u}) a(\boldsymbol{\theta}) \sum_{\mathbf{s}} q(\mathbf{s}) \exp(\mathbf{t}(\mathbf{s}), \boldsymbol{\theta}) t_i(\mathbf{s})}{\partial \theta_j} = -\sum_{\mathbf{u}} x(\mathbf{u}) \frac{\partial a(\boldsymbol{\theta}) \sum_{\mathbf{s}} q(\mathbf{s}) \exp(\mathbf{t}(\mathbf{s}), \boldsymbol{\theta}) t_i(\mathbf{s})}{\partial \theta_j} =$$

$$-\sum_{\mathbf{u}} x(\mathbf{u}) \sum_{\mathbf{s}} q(\mathbf{s}) \exp(\mathbf{t}(\mathbf{s}), \boldsymbol{\theta}) t_i(\mathbf{s}) \frac{\partial a(\boldsymbol{\theta})}{\partial \theta_j} - \sum_{\mathbf{u}} x(\mathbf{u}) a(\boldsymbol{\theta}) \frac{\partial \sum_{\mathbf{s}} q(\mathbf{s}) \exp(\mathbf{t}(\mathbf{s}), \boldsymbol{\theta}) t_i(\mathbf{s})}{\partial \theta_j} =$$

$$\sum_{\mathbf{u}} x(\mathbf{u}) \sum_{\mathbf{s}} q(\mathbf{s}) \exp(\mathbf{t}(\mathbf{s}), \boldsymbol{\theta}) t_i(\mathbf{s}) a^2(\boldsymbol{\theta}) \sum_{\mathbf{s}} q(\mathbf{s}) \exp(\mathbf{t}(\mathbf{s}), \boldsymbol{\theta}) t_j(\mathbf{s})$$

$$-\sum_{\mathbf{u}} x(\mathbf{u}) a(\boldsymbol{\theta}) \sum_{\mathbf{s}} q(\mathbf{s}) \exp(\mathbf{t}(\mathbf{s}), \boldsymbol{\theta}) t_i(\mathbf{s}) t_j(\mathbf{s})$$

$$= \sum_{\mathbf{u}} x(\mathbf{u}) (E_{\mathbf{p}(\boldsymbol{\theta})} t_i E_{\mathbf{p}(\boldsymbol{\theta})} t_j - E_{\mathbf{p}(\boldsymbol{\theta})} t_i t_j) = n(E_{\mathbf{p}(\boldsymbol{\theta})} t_i E_{\mathbf{p}(\boldsymbol{\theta})} t_j - E_{\mathbf{p}(\boldsymbol{\theta})} t_i t_j).$$

One sees that the second derivative is the negative of the covariance matrix $\mathbf{C}$ of $\mathbf{t}$, which, because of the linear independence of the components of the canonical statistic, is negative definite. Indeed, as every covariance matrix is positive semidefinite, one only needs to see that there is no nonzero vector $\mathbf{a}$, such that $\mathbf{a}'\mathbf{C}\mathbf{a} = 0$. If there was one, then $E(\mathbf{a}'(\mathbf{t} - E(\mathbf{t}))^2$ would be zero, which contradicts to the components of $\mathbf{t}$ being linearly independent, because then $\mathbf{a}'\mathbf{t} = \mathbf{a}'E(\mathbf{t})$ would hold, meaning that a linear combination with nonzero coefficients is constant. □

The importance of this result is that when a regular exponential family is in a minimal representation (representation with the possibly lowest dimension), then, if the likelihood does have a stationary point, one can conclude without further investigation of the second derivative that it maximizes the likelihood. Also, (7.15) shows an important way to characterize the stationary point of the likelihood function, which, by the previous theorem, gives the MLE.

[14] See the comment after Theorem 7.5.

Theorem 7.5. *If a regular exponential family is in the minimal representation, the observations* $\boldsymbol{X}$ *have a multinomial or a Poisson distribution, and there exists a distribution* $\hat{\mathbf{p}}$ *in the exponential family, such that*

$$\sum_{s} x(\boldsymbol{s})t_i(\boldsymbol{s}) = n\sum_{s} \hat{p}(\boldsymbol{s})t_i(\boldsymbol{s}),$$

then $\hat{\mathbf{p}}$ *is the MLE under the exponential family.* □

This result means that the expected value of the canonical statistic is inherited by the MLE from the observed distribution. A bit more precisely, if $\mathbf{x}/n$ is considered as a probability distribution, then the expectation of $\mathbf{t}$ according to it is the same as according to $\hat{\mathbf{p}}$. Moreover, if such a $\hat{\mathbf{p}}$ exists, it is the MLE.

In fact, the expected value of the canonical statistic is also used to parameterize the distributions in the exponential family and is called mean value parameter:

$$\int \mathbf{t}(\mathbf{s})d\mathbf{p}(\mathbf{s}),$$

or in the discrete case,

$$\mathbf{T}'\mathbf{p}.$$

The mean value parameters provide a parameterization of the distributions in the exponential family.

Proposition 7.1. *If* $\boldsymbol{p}_1$ *and* $\boldsymbol{p}_2$ *are both in an exponential family of the form (7.6) in a minimal representation, and their mean value parameters are equal:*

$$\boldsymbol{T}'\boldsymbol{p}_1 = \boldsymbol{T}'\boldsymbol{p}_2$$

then the distributions are also equal:

$$\boldsymbol{p}_1 = \boldsymbol{p}_2.$$

Proof. Because $\mathbf{T}$ is of full rank,

$$\mathbf{T}'(\mathbf{p}_1 - \mathbf{p}_2) = 0$$

is only possible if

$$\mathbf{p}_1 - \mathbf{p}_2 = 0.$$

□

In the setting of the Theorem, when the mean value parameters $\mathbf{T}'\mathbf{p}$ are given, using the fact that the design matrix is invertible (see Theorem 7.3), the cell probabilities may be obtained as

$$\mathbf{p} = (\mathbf{T}')^{-1}(\mathbf{T}'\mathbf{p})$$

For example, with the exponential family based on the canonical statistic in Table 7.9, the values of the mean value parameters are

$$(p(+,+), p(1,+)-p(2,+), p(+,1)-p(+,2), p(1,1)-p(1,2)-p(2,1)+p(2,2))'.$$

The results about the variation independence of the univariate marginal distributions and of the odds ratio of a 2×2 table imply that the first 3 components of the mean value parameter and the fourth component of the canonical parameter also constitute a parameterization, and these two components are variation independent. Such a parameterization is called a mixed parameterization. Mixed parameterizations in the case of arbitrary contingency tables will be discussed and used in Sect. 10.1.

It turns out the value of the mean value parameter, if computed from $\mathbf{x}/n$[15], has another important role in maximum likelihood estimation in exponential families. The kernel of the log-likelihood under multinomial or Poisson sampling was given in (7.14). This implies the following result:

Proposition 7.2. *When the exponential family is in the form (7.6), and the observations $\boldsymbol{X}$ have a multinomial distribution,*

$$\sum_{\boldsymbol{s}} x(\boldsymbol{s})\boldsymbol{t}(\boldsymbol{s})$$

is a sufficient statistic. When the observations have a Poisson distribution,

$$\sum_{\boldsymbol{s}} x(\boldsymbol{s})\boldsymbol{t}(\boldsymbol{s}) \text{ and } \sum_{\boldsymbol{s}} x(\boldsymbol{s})$$

are sufficient statistic.

Proof. Indeed, the kernel of the log-likelihood in (7.14) may be written as

$$\sum_{\mathbf{s}} x(\mathbf{s}) \left((\mathbf{t}(\mathbf{s}), \boldsymbol{\theta}) - \log \frac{1}{a(\boldsymbol{\theta})} \right) = \sum_{\mathbf{s}} x(\mathbf{s})\mathbf{t}'(\mathbf{s})\boldsymbol{\theta} - \sum_{\mathbf{s}} x(\mathbf{s}) n \log \frac{1}{a(\boldsymbol{\theta})},$$

which, by the associativity of multiplication, shows that the likelihood may be written so that it depends on the data through

$$\sum_{\mathbf{s}} x(\mathbf{s})\mathbf{t}(\mathbf{s}) \text{ and } \sum_{\mathbf{s}} x(\mathbf{s}).$$

only. The total of the observations is fixed in advance in the case of multinomial sampling but needs to be observed in the case of Poisson sampling. □

It may be the case that the structure of the canonical statistics is such that the total of the observations is implied by the values of the components of the mean value parameter, but this is not necessarily the case. This will be briefly discussed in Chap. 13.

[15] Of course, $\mathbf{x}/n$ is usually not a member of the exponential family; thus the concept of the mean value parameter is used in an extended sense.

The result of Theorem 7.5 may be reformulated by considering the exponential family

$$\mathscr{E}(\mathbf{t},\mathbf{q}) = \{\mathbf{p} : p(\mathbf{s}) = q(\mathbf{s})a(\boldsymbol{\theta})\exp(\mathbf{t}(\mathbf{s}),\boldsymbol{\theta})\}, \tag{7.16}$$

as defined in (7.6) and another family of probability distributions defined as

$$\mathscr{L}(\mathbf{t},\mathbf{r}) = \{\mathbf{p} : \int \mathbf{t}d\mathbf{p} = \int \mathbf{t}d\mathbf{r}\}. \tag{7.17}$$

The family $\mathscr{L}$ contains the distributions $\mathbf{p}$, according to which the integral of the statistic $\mathbf{t}$ is the same as according to $\mathbf{r}$. These integrals are most often marginal distributions or contrasts[16] among the cell probabilities. $\mathscr{L}$ is a linear family, in the sense that if $\mathbf{p}_1$ and $\mathbf{p}_2 \in \mathscr{L}$, then any linear combination of them that is a probability distribution is also in $\mathscr{L}$. Then the foregoing results may be summarized as follows:

Theorem 7.6. *For the regular exponential family in minimal representation (7.16) and the linear family (7.17),*

$$|\mathscr{E}(\boldsymbol{t},\boldsymbol{q}) \cap \mathscr{L}(\boldsymbol{t},\boldsymbol{r})| \leq 1.$$

If

$$|\mathscr{E}(\boldsymbol{t},\boldsymbol{q}) \cap \mathscr{L}(\boldsymbol{t},\boldsymbol{r})| \neq \emptyset,$$

then the only distribution in the intersection is the MLE in the model $\mathscr{E}(\boldsymbol{t},\boldsymbol{q})$ for any observed distribution $\boldsymbol{p} \in \mathscr{L}(\boldsymbol{t},\boldsymbol{r})$.

Proof. The first claim is implied by Proposition 7.1 and the second by Theorem 7.5. □

For example, as was described above, the distributions on a two-way table with independence of the two variables constitute an exponential family, specified by the interaction term being equal to 1. The MLE is determined by this characteristic and the two one-way marginal distributions being as observed. This is the same MLE as the one given in (5.24).

Maximum likelihood estimates in exponential families on contingency tables that are defined by specifying values of some of the canonical parameters may be determined by the Iterative Proportional Fitting Procedure (IPFP)[17] to be discussed in Sect. 12.2. IPFP uses the result of Theorem 7.6 and the variation independence of canonical parameters and the mean value parameters (see Sect. 10.1) to find the MLE.

The second result of Theorem 7.6 may be seen as a sufficient condition for the existence of the MLE: if there is a distribution in the exponential family with mean value parameters equal to the values of the sufficient statistics, then there is an MLE (which is exactly this distribution).

[16] A contrast is a linear combination, where the sum of the coefficients is zero.

[17] The same procedure is also called Iterative Proportional Scaling, or IPS.

The last comment of this section applies to the application of the EM algorithm to models when the true distribution on the contingency table is assumed to be such that each conditional distribution belongs to an exponential family. In the case discussed in Sect. 7.1, the exponential families in the two conditional tables are defined by independence of the two variables. In the more general setup considered here, conditional distributions are assumed, such that within each, the distributions belong to an exponential family, and these families may be different. The E step of the EM algorithm, given in (7.4), determines MLEs of the joint distribution in the model. If the model assumes that the conditional distributions are members of exponential families, either closed form maximum likelihood estimates exist (like in the case when the exponential family was defined by independence) or IPFP can be used for each conditional distribution. However, in the case studied in Sect. 7.1, conditional independence itself is an exponential family model and also admits a closed-form MLE.

7.3 Things to Do

1. Consider three variables A, B, C and the model assuming that A and B are conditionally independent given C. Formulate a latent class model assuming that this conditional independence holds in all groups.
2. Prove the claim made in footnote 2. More precisely, show that the maximum of the likelihood is the same, whether the complete data likelihood is maximized subject to the marginal being as observed or the likelihood of the marginal is maximized.
3. Write a computer code that applies the EM algorithm to $I \times J$ tables to find latent classes. Ask the user to specify the number of latent classes and the starting distribution. Use the numerical results given in Sect. 7.1 to check the code.
4. Show that all strictly positive probability distributions on an $I \times J$ contingency table form an exponential family.
5. Show that all l-dimensional normal distributions with zero expectation form an exponential family.
6. Show that the binomial distributions with fixed n and unspecified p are an exponential family. Determine the sample space, the canonical statistic, and the canonical parameter.
7. Show that all distributions on a three-way table, in which the first and the second variables are conditionally independent, given the third, form an exponential family. Which marginal distributions of the observed data are preserved by the MLE?
8. Compare the result obtained above to those given in Sect. 6.3.1.

Chapter 8
Effects and Associations

Abstract This chapter is somewhat different from most of the other chapters in the book. The main concern here is not statistical inference, that is, generalization from the sample to the underlying population, rather inference about expected effects of a potential treatment. Effects can be causal, evidential, and attributable, so even the formulation of the question requires some thought. As will be illustrated, such an inference would not be straightforward, even if the analyst had access to full population data. The conventional wisdom is that from data one can infer associations, but association is not necessarily causation. First, a closer look is taken at this statement, in particular, the meanings of causation and how they relate to the statistical analysis. Then, various concepts of association are discussed, out of which the one measured by the odds ratio is only one possibility. The position taken is that the different published ideas as to how to measure association are, in fact, measures of different concepts of associations. Measuring effects is a related but distinct task. The traditional theory of testing the existence and measuring the size of an effect is based on a particular way of collecting data about the responses to different treatments. This data collection procedure, the so-called designed experiment, is described, and it is explained why it is considered appropriate to establish the existence of effects. The other traditional data collection design, called observational study, is also described, and the fundamental difference with respect to measuring effects is highlighted. Next, we give a very short overview of some of the contemporary approaches to establishing effects, even causal effects, based on observational data. Such a thing was considered traditionally impossible, and the ideas summarized here, although perhaps not yet universally accepted, already have had a great impact on statistical thinking.

Although association is not necessarily causation, there is more to this issue than simply denying the relationship. The first section of this chapter takes a closer look at the relationship between association and causation.

8.1 Association and Causation

This chapter deals with inferential problems, more complex than statistical inference. Statistics, in its traditional interpretation, is concerned with making inference from the observed data in the sample to the underlying population. The structure of association among the variables is often in the focus of interest, and causation often manifests itself in association: if rain causes the pedestrians on the street to open their umbrellas, then observing rain and observing open umbrellas will be associated, that is, they will tend to occur together. But association itself does not imply causation: if it did, how would one decide if falling rain implies open umbrellas or opening umbrellas will make rain to start falling. To answer this question, one needs knowledge, which is often not part of the data, rather may be called substantive. Even if data containing the entire population, say data about every minute for a whole day covering all individuals on a square in a city in the above example, were known, the question of which causes the other one, falling rain open umbrellas or open umbrellas falling rain, would be difficult to answer, without knowing that people tend to open umbrellas to protect themselves from rain, which is not something one knows from the data. Such substantive relationships among the variables are often called the data generating mechanism.

One may think that causation cannot go back in time, so if one sees rain without open umbrellas but not open umbrellas without rain, then rain causes umbrellas to be opened. Unfortunately, this argument fails, if there are many people, who open their umbrellas in anticipation of rain, just a few seconds before it actually starts raining. In fact, the observable data do not exclude the possibility that a movie is being shot on the square, and the rain machine is turned on and the background actors playing the pedestrians open their umbrellas when the assistant director yells "action". If this is the truth (the data generating mechanism), then falling rain and open umbrellas have a common cause, and neither one is the cause of the other one. The direction of effects, sometimes called the causal structure, is often difficult to determine based on the data only.[1]

In the previous example, the meaning of cause was quite obvious, and the question was only the direction of the causal effect. The next examples illustrate that the meaning of causality may be much more complex than simply assigning the roles of cause and effect.

Suppose, one sees a car slowing down on the road and asks: Why? What causes the car to slow down? Is it the force of friction between the tires and the road surface? But why is friction there? Is it because the brakes slow down the wheels? Why is it happening? Because (jumping a few components of the chain) the driver pushed the brake pedal? Why is it? Because she saw children playing next to the road? Why does seeing children playing next to the road make the driver to push the brake pedal? And so on. In this situation, it is difficult to give a single and well-

[1] It is not a simple omission that no formal definition of causality is given. Causality is not a mathematical concept, unless specific models are defined, as in Sect. 8.4. There have been many efforts over the past centuries to define causality in philosophy, but none of these seems universally accepted.

defined cause for the car slowing down, rather one sees a long list of candidates for being the cause, which all seem necessary but none of them is sufficient to make the car slow down. In another example, the pilot of a passenger airplane sees a warning light coming up. In this case (we all hope) there is a finite and possibly short list of potential causes, which he needs to check and then fix the problem to maintain the security of the passengers and the crew.

The two situations differ from the perspective of why the cause was asked. In the first example, the general question asked does not have a simple answer. In the second example, the pilot was looking for the appropriate action (often called intervention in this context). He did not ask why the light came on in a general sense. He was not interested in the internal working of the control panel where the light came on. Rather, he wanted to find out what needed to be fixed out of a predefined list. More generally, one may say that data analysis is often applied to support policy decisions, and these applications require the statistician to deal with the concept of causality. A policy decision is a decision about an intervention in order to achieve a desired effect, just like the pilot fixing the problem.

To be able to deal with causality, one often needs to restrict the domain in which the cause is sought for. If we assume (restriction) that the driver does not want to hit children, she sees that the children are close to the road, so that there may be a danger, and that the car is in good mechanical condition, we may identify the cause for the car slowing down as the driver's decision to push the brake pedal.

Such a restriction, which is needed to be able to deal with causality, is not always immediate. It is present in strongly engineered contexts, like the airplane or a computer code (if the program does not work as expected, there is often a single programming error in the background), but it is certainly not present if one asks why a disease developed in an individual, or why a disease develops more frequently in one group of people, than in another one.

However, in closer analysis, it turns out that the problem of inferring causality from observed data also has some statistical components: some aspects of data collection affect one's ability to make causal inferences from data.[2] This effort requires one to be specific about the concept of causality one wants to use. The basic idea is that if the statistician could make rain and then observe whether or not pedestrians open their umbrellas, and if the statistician also could make people open their umbrellas and then observe whether or not rain starts to fall, and all this could be done free from other influence, then the question of which causes which could be answered.

To sum up, statistical analysis is usually concerned with inference, based on the data in a sample, to the underlying population. If a research question could not be answered even if data from the entire population were known, statisticians usually cannot help. This is why statisticians for a very long period of time abstained from trying to reach causal conclusions. Based on the situation illustrated above, they warned that association is not causation and said that causality was a substantive question, which could not be the subject of statistical analysis, except for a particular

[2] This is elaborated in Sect. 8.3.

method of data collection, to be described in Sect. 8.3.1. There are relatively recent results, to be summarized in Sect. 8.4, which led many statisticians to think that this position needs to be revised.

The complexity of the meaning of causality illustrated above is paralleled by a multitude of interpretations of the concept of association. Some of these approaches are discussed in the next section.

8.2 Different Concepts of Association

This section describes ways of thinking about association. There exists no universal definition, and different approaches may be useful in different settings.

Two events, A and B, may be said to be associated if most of the time when A occurs, B also occurs and, further, most of the times when B occurs, A also occurs. In other words, most of the times both A and B or neither A nor B occurs. This requirement is the same as demanding that one of the events without the other one does not occur often. When one event without the other one never occurs, i.e., when either both or none of A and B can be observed, the association may be said to be perfect. In this case, occurrence of any of A or B implies the occurrence of the other one. The association concept discussed so far is usually called positive association. However, in some contexts, when most of the time either A or B (but not both of them) occurs, the events are said to be negatively associated or dissociated. Indeed, when the two events are either positively or negatively associated, based on the lack or presence of one event, one can predict the lack or presence of the other one. For example, if A was observed, then under positive association one can also expect B to have occurred, and under negative association, one can expect B not to have occurred. The strength of association is a measure of how often such expectations turn out to be correct.

The concept of association among variables is a generalization of the concept of association between events. If the random variables A and B are binary, then (positive) association between them means that in the contingency table representing their joint distribution, most of the probability is on the main diagonal, and the off-diagonal cells have little probability. If also negative association is taken into account, which means concentration of the probability off the diagonal, the strength of association depends on how strong is the concentration of the probability, in either the diagonal or in the off-diagonal cells. Table 8.1 shows distributions with different patterns of association.

Table 8.2 illustrates that the above considerations are not always sufficient to describe the lack or presence of association. In the first distribution, probability is concentrated, to a certain extent, on the main diagonal, but the two variables are also independent, so it seems hard to decide, based on the concept of concentration, whether one sees positive association or independence. Recall that in Sect. 5.4.1, independence was interpreted as knowing the category of one variable does not help to guess the category of the other, which is the lack of association. In the second distribution, the amount of probability (concentration) on the main diagonal is the same, but one sees no independence. A comparison of these two distributions suggests that association cannot be decided entirely based on the concentration of the probability on the main diagonal (or in the off-diagonal cells). To underline this problem, note that if in a series of 2×2 distributions, both $P(A)$ and $P(B)$ converge to 1, then even if the two variables are independent, the probability on the main diagonal $P(A)P(B) + (1 - P(A)(1 - P(B)) = 1 + 2P(A)P(B) - P(A) - P(B)$ will converge to 1, as if association was very strong in the sense of concentration on the main diagonal.

Table 8.1 Different patterns of association for two binary variables

	$B : yes$	$B : no$	$B : yes$	$B : no$
$A : yes$	0.35	0.05	0.05	0.65
$A : no$	0.15	0.45	0.25	0.05
	Positive association		Negative association	

Table 8.2 Different patterns of no association for two binary variables

	$B : yes$	$B : no$	$B : yes$	$B : no$
$A : yes$	0.06	0.14	0.5	0.03
$A : no$	0.24	0.56	0.35	0.12
	Positive association or independence?		Positive association?	

The previous examples illustrate that the amount and direction of association may have to be defined in comparison with independence.[3] Positive association then may mean that the concentration of probabilities on the main diagonal is stronger than what it would be under independence (given the marginal distributions), and negative association would mean that the concentration of probabilities on the main diagonal is weaker than what it would be under independence. The following approach implements this idea but using a specific way of measuring concentration.

Association will now be defined based on a comparison of conditional odds. If

$$\frac{P(A : yes, B : yes)}{P(A : no, B : yes)} = \frac{P(A : yes|B : yes)}{P(A : no|B : yes)} > \frac{P(A : yes|B : no)}{P(A : no|B : no)} = \frac{P(A : yes, B : no)}{P(A : no, B : no)},$$

that is, when the odds ratio is greater than 1, the association is positive, if

[3] In Chap. 6, association was defined as the information in the joint distribution, which is additional to the information in the marginal distributions, and independence was interpreted as no such additional information in the joint distribution.

$$\frac{P(A:yes,B:yes)}{P(A:no,B:yes)} = \frac{P(A:yes|B:yes)}{P(A:no|B:yes)} < \frac{P(A:yes|B:no)}{P(A:no|B:no)} = \frac{P(A:yes,B:no)}{P(A:no,B:no)},$$

that is, when the odds ratio is less than 1, the association is negative, and if the odds ratio is 1, the two variables are independent, that is, there is no association. Independence is not the opposite of association, rather a special case of it: no association.

If the odds ratio is used to define the strength and direction of association, as described above, positive association means that the ratio of concentration of the probability in the cells on the main diagonal, compared to the concentration in the off-diagonal cells, is stronger than what it would be under independence, if the amount of concentration is measured by the product of the two probabilities. Whether or not this is a meaningful way to measure concentration, may, of course, be debated. In summary, association is a characteristic of the joint distribution of two variables, and it may be quantified by the odds ratio, as was discussed in Chap. 6.[4]

There is a large literature on measures of association, that is, functions of the joint distribution proposed to measure the strength of association among the two variables. It is, however, more precise to say that these do not measure association in different ways, rather measure different concepts of association. These different concepts are not even monotone functions of each other, so a distribution showing stronger association than another one may have less association, if a different measure (rather, concept) is used. A review of many of the early developments is given in [35], and [77, 78, 79], illustrated the multitude of approaches. Some of the aspects which are relevant when choosing a measure of association but also a measure of effect will be discussed in Sect. 9.3. Except for the current chapter, however, the methods discussed in this book rely mostly on the odds ratio to measure association.

The next sections present different ways of data collection, and the aspects discussed will be relevant when one wishes to make inference about effects, not only associations.

8.3 Experiments and Observational Studies

Traditionally in statistics, two fundamentally different methods of data collection are distinguished, but real data collection procedures often have characteristics from both. This distinction is very important from the perspective of establishing effects that can be attributed to a treatment or intervention. The variables considered, in a basic setup, are treatment and response, but in more complex designs, other variables which are both treatment and response are also taken into account. In this context, the goal is to collect data, based on which one can decide, whether in the underlying population, treatment has an effect on response.[5] The concept of attributable effect is somewhat weaker than that of the causal effect and is used whenever the

[4] When the two variables are not binary, there is no single number that would describe the association structure.

[5] If there is an effect, one also wants to estimate its size.

existence of an effect may be established, but the causal mechanism behind it is not understood. Some authors distinguish (see, e.g., [57]) between evidential and causal decision-making, where the first one refers to cases when evidence is available that an intervention (treatment) is followed by the desired outcome, and the second one refers to cases when the intervention (treatment) causes the desired outcome.

The two different data collection methods are called designed experiments and observational studies. The fundamental difference is that in a designed experiment, the analyst decides which subject receives which treatment, while in an observational study, the analyst only observes the choices of the subjects as to which treatment to receive.

8.3.1 Designed Experiments

Experiments are designed to maximize the ability to draw conclusions about effects. If a group of surgeons invents a new way to perform an operation, they may want to compare the recovery times after the operation for the old and the new methods. In this experiment, treatment is the operation (old or new), response is the recovery time (e.g., less than a week or more than a week, but the variable may also be defined as the number of days till a patient achieves a certain status), and the subjects of the experiment are the patients who need the surgery in the hospital where the surgeons work. The essence of the experiment is comparison of responses to different treatments. If the surgeons decide to apply the new method to less severe cases, and the old one to more severe cases, then they are likely to experience faster recovery times for those who have undergone the new method of operation, but one cannot conclude that this is because the new method is better than the old one. This may be the case, but it is just as well possible that less severe cases recover faster after surgery, irrespective of the mode of operation. Similarly, if the surgeons apply the new method to younger patients, or to ones who feel more trust toward the operating surgeon, the difference in recovery times (distribution or average) cannot be attributed to the different modes of operation.

In order to be able to conclude that one mode of operation leads to faster recovery, than another, one would have to

1. define a population to which the claim applies
2. choose a sample from this population
3. apply both treatments to each unit in the sample
4. observe the results of each treatment on every unit in a way not influenced by the fact that another treatment was or will be applied to the unit
5. compare the responses to each treatment

Different real applications of experiments to requirements above are of different levels of difficulty. In the surgery example, item 1 may be taken as all patients, who require an operation for a particular condition, but it should be clear that whether or not an actual person belongs to this population is a matter of decision, usually not without subjective elements. The issue of a proper sample in item 2 is not pressing if the population of interest consists of units which can be seen as homogeneous with respect to their responses to the treatments considered,[6] but in many populations, in particular those applied to animals or humans, units of the populations may be heterogeneous, and proper sampling becomes an issue. See the discussion of external validity at the end of this subsection.

The requirement in item 3, in the surgery example, means to perform both kinds of operation on the same individuals, which is, obviously, not possible.[7] More generally, this is the situation in many experiments, where treatments destroy the experimental units (e.g., seeing whether artillery shells fire or what pressure a material can withstand). Destroying is a special case of changing the experimental units, so the result of the response to the treatment applied as the second one will be influenced by the first treatment, too. One has to assume the latter situation occurs, whenever the experimental units are humans. This shows the difficulty with item 4.

Assuming the counterfactual requirement in item 3, and also the one in item 4 could be achieved, one would estimate the causal effect of the new treatment, as required in item 5, relative to the old treatment as

$$\begin{aligned} &\mathit{Average}(\text{recovery time after the new treatment}) \\ -&\mathit{Average}(\text{recovery time after the old treatment}), \end{aligned} \tag{8.1}$$

and the better treatment could be selected based on the sign of the above quantity. Of course, (8.1) is a sample estimator of the relevant population quantity, the expectation of (8.1), and by taking into account the sampling procedure, it may be subject to tests of various hypotheses (e.g., whether it is different from zero). Also, quantities different from the average in (8.1) may be of relevance, when choosing a treatment, like the minimum or maximum recovery times.

In the sequel, we concentrate on an approach to deal with the counterfactual condition in item 3.

In fact, there is a weaker condition which is sufficient for estimating the expected value of the causal effect, one which may be achieved. This condition is that the expectation of the response to either treatment is the same in both treatment groups. More precisely, that

$$\begin{aligned} &E(\text{recovery time after new treatment} \mid \text{old treatment was received}) = \\ &\quad E(\text{recovery time after new treatment} \mid \text{new treatment was received}) \end{aligned}$$

and

$$\begin{aligned} &E(\text{recovery time after old treatment} \mid \text{old treatment was received}) = \\ &\quad E(\text{recovery time after old treatment} \mid \text{new treatment was received}). \end{aligned}$$

If this was true, then one would have

[6] See also the concept of exchangeability below.

[7] Such impossible conditions are called counterfactual in the literature on causality.

$$E(\text{recovery time after new treatment}) = \\ E(\text{recovery time after new treatment} \mid \text{new treatment was received})$$

and

$$E(\text{recovery time after old treatment }) = \\ E(\text{recovery time after old treatment} \mid \text{old treatment was received})$$

and then the causal effect could be obtained as

$$E(\text{recovery time after new treatment} \mid \text{new treatment was received}) - \\ E(\text{recovery time after old treatment} \mid \text{old treatment was received})$$

and both of these quantities may be estimated from the data, without any counterfactual assumption.

This condition is implied if the distributions of recovery times (after the new and the old surgery procedures) in the two groups (those who actually received the new and those who received the old treatment) would not change, if one individual from one of the groups would be swapped with another individual from the other group. This property of the distributions in the two groups is called exchangeability. Exchangeability is achieved, if the experimental units may be seen as homogeneous with respect to their responses to the treatments, and it does not hold, e.g., if one group contains the more severe and the other the less severe cases. But exchangeability is achieved – as will be shown – if the individuals are allocated randomly into the two treatment groups. Random allocation means that the probability of being assigned to one or the other treatment group is constant for all individuals, or, equivalently, does not depend on any of their characteristics, like how severe is their illness or what is their age. Somewhat informally, this characteristics is often referred to as the two groups being identical (except for the treatments received), at least in expectation.

Random allocation[8] cannot be replaced by trying to allocate subjects in a way that the two groups become identical. The main reason for this is that the surgeons may avoid assigning the less severe cases to one treatment group and the more severe cases to the other treatment group, but other aspects that also require balancing of the two groups may not occur to them (e.g., they may not think of trust as a factor influencing recovery times) or may not be known by them (e.g., two patients with the same symptoms may have very different diseases and potentially different recovery times after surgery).

In addition to random allocation, there are further requirements to make the two groups comparable and the differences in response, if found, attributable to the different treatments received. These requirements may be seen as necessary for the two groups to remain identical even during the experiment, in terms of their potential responses to either one of the treatments. Some patients may have high expectations

[8] Random allocation is also called randomization.

of new surgical procedures; others may trust traditional methods better. It is important that the recovery times are only influenced by the different surgical procedures and not by the different expectations the subjects may have with respect to the different methods of operation. Therefore, the subjects are kept blind, that is, are not told whether they have undergone the new or the old method of operation.[9]

But it is not only the subjects, who may have expectations with respect to the outcomes of the different methods of operation, but also the surgeons, who invented the new procedure. Oftentimes, not only expectations but also interests are involved. To make sure the patients are declared recovered, when they really are, and not when it would be beneficial for the inventors, also the evaluators, who decide about the status of every patient, should be kept blind, as to which method of surgery was used. If this is also part of the design of the experiment, it is called a double-blind experiment.

A special kind of experimental design is needed, when the goal is not the comparison of different treatments, like the old and new surgical procedures in the example above, rather to decide, whether the application of a certain treatment is more beneficial than applying no treatment at all. Also in such tasks, the response to treatment needs to be compared to some control, like the response to the new surgical procedure was compared to that of the old one in the previous example. In these cases, a so-called placebo is applied. A placebo is a procedure that appears to be a treatment for the subjects in the experiment but has no effect, at least no effect along the lines the real treatment may have. In fact, it may happen that those treated with a placebo show a larger effect (e.g., faster recovery) than those who received the real treatment. In these cases, the usual conclusion is that idea of having been treated has an effect, which is called the placebo effect, but the treatment that was to be tested has a less beneficial effect than the fake treatment by the placebo, so the real effect is negative. However, applying a placebo makes it possible to keep the subjects participating in the experiment blind.

In summary, a good experiment is randomized, controlled[10], and double blind. If the responses in the two treatment groups in a well-designed and carefully executed experiment are different, the differences are usually attributed to the different treatments received (and not to the potential differences among the subjects in the two groups, including their potentially different responses to the treatments actually received), although only an estimate of the causal effect, not the causal effect itself, is observed. This forms the basis of the required policy decision, e.g., deciding which surgical procedure to apply.

To formalize why designed experiments are appropriate to infer about effects, let T denote the treatment variable and R the response variable. In the examples so far,

[9] This practice, although seems necessary, raises ethical issues in many settings, restricting the applicability of experiments to humans.

[10] Controlled refers to the existence of comparison. If the recovery times after the new treatment were not compared to the recovery times after the old treatment or to recovery times after a placebo treatment, nothing could be said about the relative efficacy of the new treatment.

both T and R were binary, but this is, in general, not a requirement.[11] One more variable is needed to describe designed experiments, and this is allocation, denoted as A. Allocation tells which treatment group a subject is assigned to. Because the treatment groups may be identified with the treatments to be compared, A has the same categories as T does. Because of randomization, A is a random variable, and, as the main feature of random allocation, the conditional distribution of A is common for all subjects in the experiment. The treatment received by an individual participating in the experiment is entirely determined by allocation,[12] so $T = A$; however, in addition to the actually received treatment, hypothetical treatments will also be considered. The response may, of course, depend on the treatment received, so the behavior of R is best described by the conditional distribution of R given $T = t$, where t is any treatment considered. The conditional distribution $R|T = t$ depends on the characteristics of the subjects in the experiment.

Usually, the choice of the better treatment depends on a comparison of certain parameters of the distribution of R, when different treatments are received. For example, if t_1 and t_2 are the two treatments to be compared, and r_1 and r_2 are the two possible responses, like the patient's condition improved or did not improve, instead of the average considered above, of interest may be any of the following quantities:

$$P(r_1|T=t_1) - P(r_1|T=t_2)$$

$$\log P(r_1|T=t_1) - \log P(r_1|T=t_2)$$

$$\log(P(r_1|T=t_1)/P(r_2|T=t_1)) - \log P(r_1|T=t_2)/P(r_2|T=t_2)),$$

where the first one is the increase in the probability if improving (after having received t_1, as opposed to t_2), the second one is the log relative proportion of cases improved, and the third one is the log odds ratio. When R is numerical, like the number of days till recovery or the blood pressure, then of interest may be

$$E(R|T=t_1) - E(R|T=t_2),$$

which is the expected difference in R (after having received t_1, as opposed to t_2), as above. This quantity does not only tell which is the more beneficial treatment but also gives the effect size. The differences in the responses formulated all use the expression "after having received treatment t_1 as opposed to t_2". The reason for using designed experiments to collect data is that this expression, at least in expectation, may be replaced by saying because of having received t_1 as opposed to t_2.

All the above quantities are differences of functionals of the conditional distributions of R and will be denoted as

$$\mathscr{F}(R|T=t_1) - \mathscr{F}(R|T=t_2).$$

[11] Both treatment and response may be continuous, too. For example, treatment may be the dosage of a drug administered, and response may be the reduction in systolic blood pressure achieved.

[12] However, for real experiments, especially those involving humans, this may not be true; see the discussion of non-compliance later in this subsection.

The crucial feature of designed experiments is that A is independent of any characteristics of the subjects, including the conditional distribution of $R|T=t$, which describes how they would respond if treated in a particular way. Therefore, using that X independent of Y means that the conditional distribution $X|Y=y$ does not depend on y and, thus, is the same as the unconditional distribution of X,

$$\mathscr{F}(R|T=t_1|A=t_1)=\mathscr{F}(R|T=t_1|A=t_2)=\mathscr{F}(R|T=t_1) \tag{8.2}$$

and

$$\mathscr{F}(R|T=t_2|A=t_1)=\mathscr{F}(R|T=t_2|A=t_2)=\mathscr{F}(R|T=t_2). \tag{8.3}$$

This means that the response to $T=t_1$, which is described by the conditional distribution $R|T=t_1$, and the quantity of interest based on this conditional distribution, $\mathscr{F}(R|T=t_1)$, is the same, whether it is determined only for those who were allocated to receive t_1 or for only those who were allocated to receive t_2 or for everybody, irrespective of treatment allocation, and similarly for the response to t_2. Obviously, under the assumption that everybody receives the treatment into which they were allocated, the second term in (8.2) and the first term in (8.3) are counterfactual and cannot be observed and estimated from data. Conditioning on the actual value of $A=t_1$ selects a group of subjects, and conditioning on $A=t_2$ selects the remaining subjects. The quantities above measure how these two groups of subjects would respond to each of the treatments, and the equations say that they would behave the same, in the sense of producing the same response distributions or, at least, the same value of the functional of interest. Consequently,

$$\begin{aligned}&\mathscr{F}(R|T=t_1|A=t_1)-\mathscr{F}(R|T=t_2|A=t_1)\\ &=\mathscr{F}(R|T=t_1|A=t_2)-\mathscr{F}(R|T=t_2|A=t_2)\\ &=\mathscr{F}(R|T=t_1|A=t_1)-\mathscr{F}(R|T=t_2|A=t_2),\end{aligned}$$

meaning that the quantity of interest, for the comparison of the effects of treatments t_1 versus t_2, is the same for those who were allocated to receive treatment t_1 and for those who were allocated to receive treatment t_2, and both are equal to the quantity that is obtained by a comparison of those who actually received t_1, because they were allocated to be treated by t_1 and of those who actually received t_2, because they were allocated to that treatment group. Further, because of (8.2) and (8.3),

$$\begin{aligned}&\mathscr{F}(R|T=t_1|A=t_1)-\mathscr{F}(R|T=t_2|A=t_2)\\ &=\mathscr{F}(R|A=t_1)-\mathscr{F}(R|A=t_2).\end{aligned} \tag{8.4}$$

The importance of (8.4) is that its right-hand side may be estimated based on the actual data. These results are summarized in the next theorem.

Theorem 8.1. *Suppose, in an experiment which compares the responses R to treatments t_1 and t_2, the subjects available are allocated randomly to the two treatment groups, independently of their characteristics, so that the conditional distribution $R|T$ does not depend on allocation. Assume further that each subject receives the* *treatment they were allocated to. Then, for the effect of the treatment, the following holds:*

$$\mathscr{E} = \mathscr{F}(R|T = t_1) - \mathscr{F}(R|T = t_2) = \mathscr{F}(R|A = t_1) - \mathscr{F}(R|A = t_2). \tag{8.5}$$

If $\hat{\mathscr{F}}(.)$ is an unbiased estimator of $\mathscr{F}(.)$, which may be applied to the observed distribution of R, then the difference of the $\hat{\mathscr{F}}(.)$ values obtained for the observed distributions of $R|A = t_1$ and of $R|A = t_2$ is an unbiased estimator of $\mathscr{E}$. □

The expectations in the above results are interpreted with respect to the random allocation. The unbiased estimator in Theorem 8.1 may be (a function of) a relative frequency or (a function of) a probability or the sample mean for the expected value.

Real experiments, in particular those involving human subjects, often fall short of the idealized construction above.

How the lack of blindness may introduce biases into the estimation procedure depends heavily on the circumstances of the experiment, including how objective the measurements are, how much self-reporting is involved, what the subjects know about the treatments, etc. These aspects will not be discussed here in detail. There is, however, one related aspect of experiments, which needs to be discussed, and this is whether the subjects available for the experiment are human beings or else material samples[13] which do not reflect upon the fact of being part of an experiment. Human beings often fail to follow the experimental protocol, e.g., fail to take the pills on time (or at all), called non-compliance, and may decide to stop participating in the experiment, in particular if that involves a longer period of time, or may become ineligible, called dropout. Non-compliance means that allocation does not imply treatment, and the proof of Theorem 8.1 fails. The effect of dropout depends on the mechanism that leads to dropout. For example, those who feel they have become entirely healthy may stop participating in a medical experiment, but those who die during the experiment (because of reasons related to but also unrelated to the treatment) are, technically, also in the dropout category.

Randomized double-blind experiments are usually considered appropriate to make valid inferences with respect to the existence of an effect attributable to a treatment, as opposed to another treatment, and oftentimes such effects are also considered causal. Strictly speaking, this validity is internal, because the argument above referred to the subjects available for the experiment and unbiased estimates of the effects could be obtained for them. Unbiased estimation in Theorem 8.1 meant that if all allocation was considered, the average estimated effect would be equal to the counterfactual effect which would be obtained if both treatments were applied to every unit participating in the experiment. Whether this is an unbiased estimator of what one would obtain if all members of the population participated in the experiment, in which case it would also have external validity, depends on whether the subjects available for the experiment may be considered a random sample from the population of individuals defined in item 1. This does not seem problematic, when the experiment is about material samples, but may be problematic if the experiment

[13] The word sample is used in the nonstatistical sense here, with the meaning of a specimen.

involves human beings. In many experiments, participants are recruited in ways which do not seem to be proper random sampling procedures. In certain fields, most experiments are done with university students; in others, it is customary to offer monetary compensation to participants, raising the issue whether young people who are not university students, or those who would not participate in the experiment even if paid for, would have given the same responses.

The final cautionary comment about experiments is that the assumed independence of allocation and of response to treatment (also a condition in Theorem 8.1) may not hold in the case of any real experiment. Random assignment into the treatment groups implies that when randomization is applied to a large number of experiments, the conditional distributions of response given treatment would be about the same for those assigned to receive t_1 or t_2. However, random allocation for a single experiment guarantees little certainty that this would be the case for the actual allocation. In this respect, a larger sample size is relevant (just like to improve the precision of the extension of the results to the whole population), but experiments are often difficult and expensive, and experimenters often have to settle for small sample sizes. This seems less of a problem when the population of interest may be seen as homogeneous. In such cases the issue of sample size usually is only related to the precision of estimating the size of the effect.

In summary, although experiments are a well-established method of data collection to infer effects, the situation described in Theorem 8.1 is somewhat idealistic, in particular when the experiment involves human beings.

8.3.2 Observational Studies

In designed experiments, the researcher decides through the application of a random allocation procedure, who receives which treatment. However, there are many data collection procedures where the researcher has no influence on the treatment received by the observed individuals. Such data collection procedures are called observational studies. Evaluation of effects of treatments may also be needed in cases when designed experiments are not feasible; only observational data are available. For example, when effects of using different drugs are to be compared with the understanding that all drugs in the study are harmful, organizing an experiment where different individuals would take different drugs for a certain period of time is usually considered unacceptable. However, data collected from drug users may be available. Collecting and analyzing such data is an observational study. In other cases, experiments are possible, but a large scale experiment would be prohibitively expensive. For example, insurance companies may wish to offer lower premiums for drivers of cars which require cheaper repairs. When cars are to be compared as to the average damage after an accident[14] a few of each make, a model may be tested in an experiment. But there may be several types of crashes and other conditions to take

[14] Before the premiums are decided, the insurance company will also take into account the differences in the likelihood of an accident.

into account, and instead, it is of great value to analyze data collected from all real accidents that involved the cars of interest. Such data are observational.

In most observational studies, inferring effects is problematic, because those choosing different treatments may be different and their responses to the actually received treatments may also be different. Consequently, if different responses are observed in a comparison of groups who have received (i.e., chose) different treatments, this is not necessarily because of the different treatments but may also be a consequence of the differences among those choosing different treatments. In this case, (8.2) and (8.3) and Theorem 8.1 do not hold. For example, users of different drugs may be very different, and also their reactions to different drugs may be not comparable. To give an example of the mechanisms which may be behind such effects, take the case of a car manufacturer, who advertises that its cars are very reliable. After a number of years, a large study is conducted to determine whether these cars are more reliable than other cars in the same price range. Data from roadside assistance providers are collected and found that while the cars of this manufacturer released within the last 2 years had 16.8 mechanical or electrical failures per 1000 registered, the average of the same value for comparable cars is 32.3. Obviously, 16.8 is much less than the class average 32.3, about half of it, so these cars are more reliable. Here, the observational units are cars, and the treatments to be compared are being produced by this manufacturer or by another one. So it seems cars produced by this manufacturer operate more reliably during the first 2 years of ownership than cars produced by other manufacturers. But is it really the case that they are more reliable because they were produced by this manufacturer? What leads to the reliable operation of these cars? Isn't it well possible that cars which are considered reliable are bought (in larger fraction) by people who think reliability is the most important feature of a car, as opposed to, say, giving an enjoyable driving experience? Isn't it, further, quite likely that people who think reliability is the most important feature of a car will take better care of their cars in terms of proper maintenance and then their cars will require less frequent roadside assistance? So, is it possible that cars produced by this manufacturer are not more reliable than other cars, just are simply being better taken care of? This question cannot be decided based on observational data; however, concluding that these cars are more reliable than others, based on the above data, is quite problematic.

There are efforts in certain fields to handle the problem described above by applying matching to find comparable groups from among those who chose different treatments. In the case of the example above, matching would mean that not all cars manufactured by this company are compared to all cars manufactured by others, but only cars with similar owners are compared. One way to implement this idea is to look at the age, income, educational level, gender, or any other characteristic of the owners of this particular brand and select a similar group of owners from among those driving different brands. Here, matching is done in the hope that the matched groups of owners will take equally good care of their cars and if difference in the fractions of cars requiring roadside assistance is found, it can be attributed to the different qualities of the brands. Unfortunately, this assumption may be entirely wrong, and it is usually hard or even impossible to test. The assumption behind matching would be implied, if the characteristics used for matching determined the level of care. This can be very questionable, in particular that the variables available for matching are usually very few, and the effects of the variables which may be available for matching on the level of maintenance may be entirely unknown.

In the case of epidemiological studies, where matched controls are often used, those exposed to an assumed risk factor of a disease are compared to those not exposed to it, and the prevalences in the two groups are compared. The exposed and unexposed groups are not determined by the researcher using randomization, rather exposure or no exposure are usually results of complex procedures, which contain choices of the individuals involved but also environmental or societal factors; thus the data are observational. Matching the two groups based on sex and age, as is often done, does not seem to imply that the responses are comparable, and a difference, if found, may be attributed to the lack or presence of the risk factor. A related design, called case-control study, compares patients suffering from a medical condition to another group free from this disease. The rates of exposure to an assumed risk factor are compared in the two groups, often using matched individuals. Also in these cases, matching is usually limited to a few variables, which does not imply, in general, that differences in the exposure rates may be seen as leading to the different health statuses of the two groups.

Obviously, in order to make statistical inference, like estimation of a parameter value or test of a hypothesis in an observational study, it requires proper sampling procedures. Some of the theoretical properties of random sampling procedures were discussed in Chap. 2 of this book.

Some designed experiments, in their real implementation, turn out to have observational features, leading to more complex designs. For example, almost all real experiments involving humans have a certain level of non-compliance and dropout. Thus, the group of individuals, who received a particular treatment, is a result of a combination of the randomization performed by the researcher and of the choices of the individuals participating in the experiment or of outside factors. The former aspect is experimental, the latter is observational, and the consequences were discussed in Sect. 8.3.1.

8.4 Some Theories of Causal Analysis Based on Observational Data

Earlier in this chapter, the traditional reluctance of statisticians to make causal conclusions was mentioned briefly. There seems to be an agreement that based on designed experiments, causal conclusions are justified, along the lines described in the previous section, using the observed difference in the responses of those treated and of those not treated as an estimate of the expected difference in the responses if everybody was treated and if nobody was treated.[15] This comparison uses counterfactual (or potential) outcomes. The idea of this comparison essentially stems from [61] and has been further developed by a number of researchers; see [69]. The skepticism regarding causal conclusions applies to observational data or more generally to comparisons based on data sets which were not obtained from clear random allocation. The new developments which seem to have the potential to change this position include work reported in [68] and in several subsequent publications and also in some of the ideas described in [62]. The approaches put forward by Rubin and Pearl are fundamentally different, and their relationship, perhaps the superiority of one over the other one, is subject to ongoing scientific debate.

Although a deeper discussion of these theories of causality is outside of the scope of this book, as a very superficial summary, one may say that Rubin shows that under appropriate conditions, the ideas applied to data from designed experiments may be applied to data from observational studies. Comparing the responses of those who received t_1 to the responses of those who received t_2 is not going to give an estimate of the (causal) effect of one treatment versus the other, but such comparisons based on certain subsets of the two groups may give unbiased estimates of the effect. The key idea comes from a certain formulation why observational studies are not appropriate to reach causal conclusions: because those who chose a particular treatment may have had good reasons to choose that treatment (perhaps including a particular response to the treatment), and therefore their responses to different treatments cannot be compared. As an example, think of individuals who do and do not choose to smoke. Smokers do find some benefit in smoking which nonsmokers don't, so their overall physiological, mental, or emotional responses are different.

It will be shown next that the effect size may be estimated by comparing those who have and those who have not received the treatment from among those who would have received the treatment with equal probability. This probability is called the propensity score. The existence of the propensity score and its usefulness in matching treated to untreated rely on a number of assumptions.

Let the vector X denote all characteristics of the individuals in the study.[16] The meaning of all characteristics is that if I is the set of individuals in the study and if g is a function defined on I, then there exists a function h, such that for all $i \in I$,

$$g(i) = h(x(i)).$$

Therefore, individuals with the same $X = x$ need not and cannot be distinguished.

The main difference between designed experiments and observational studies may be formulated by saying that in the former, X is independent from the allocation A, while this may not hold in observational studies. Let this independence be denoted as

$$X \perp\!\!\!\perp A$$

[15] Or if everybody was treated with one treatment and if everybody was treated with the other treatment.

[16] We disregard here the sampling aspect.

and under full compliance, the treatment is also independent of the characteristics:

$$X \perp\!\!\!\perp T.$$

This independence may not hold in observational studies, so let us try to achieve a weaker[17] property, namely, that instead of the above independence, a conditional independence holds, when conditioned on a fixed value of some function S of X, called a balancing score:

$$X \perp\!\!\!\perp T | S(x),$$

for all values x of X. The precise meaning of the above conditional independence is that X and T are independent on every subset of I, where $S(X) = S(x)$ for some value x of X. If such a balancing score existed, then among those who are at the same value of the balancing score, those treated by t_1 and those treated by t_2 would be identical in every aspect, including their factual or counterfactual responses to either of the treatments, just like in a designed experiment.

Technically, there exists at least one function S, for which the above conditional independence holds, namely, when S is the identity function, so that $S(x) = x$. If two individuals participating in the experiment are identical in every aspect (both are characterized by $X = x$), then their responses to either one of the treatments would be the same, irrespective of which treatment they actually received, and if they did receive different treatments, their responses could be compared to assess the effects of the different treatments. This, again, underlines that X should contain all characteristics, so that the conditional distribution of response given treatment for an individual depends only on x, for every individual.

The theory so far does not appear to be very useful in practice, as one cannot really decide whether two individuals have identical values on enough variables so that their identical response distributions would be implied, that is, finding or constructing X with the properties assumed does not seem feasible in practice. However, in [68] a simple but useful observation is made about the existence of a balancing score.

To formulate the result, the foregoing, somewhat heuristic considerations need to be made more precise, and the underlying assumptions need to be made explicit. To emphasize this aspect, the notation is also changed a bit. Assume each individual i in the population of interest is characterized by a probability of receiving (choosing) tr_1. Denote this probability by $P(T(i) = tr_1)$ and then $P(T(i) = tr_2) = 1 - P(T(i) = tr_1)$. Further, assume that individual i having received one of the treatments gives responses with certain probabilities, and let $P(R(i, tr_j) = rp_k)$ denote the probability that individual i having received treatment tr_j gives response rp_k. The probabilities defined have nothing to do with random sampling; they are assumed to be characteristics of the underlying population and of every individual in the population.

It has to be emphasized that the existence of these probabilities is a very strong assumption, because there is often no possibility neither to check its appropriateness

[17] Weaker is not used here in a strict mathematical meaning. If independence was true, it would not necessarily imply the conditional independence.

in the actual research problem nor to estimate the probabilities which are assumed to exist. The probability of choosing a particular treatment for an individual may be a subjective probability expressing the strength of inclination of that individual to choose one of the treatments, or it may be a frequentist probability in case each individual has several opportunities to choose one of the treatments. For the concept of subjective and frequentist probability, see [46].

Further, let X be a variable defined on the population, such that if for two individuals i and i' from the population $x(i) = x(i')$, then for every j and k,

$$P(T(i) = tr_1) = P(T(i') = tr_1) \text{ and } P(R(i, tr_j) = rp_k) = P(R(i', tr_j) = rp_k).$$

This is the sense in which the variable X is assumed to contain all relevant information. Note that X is not a random variable in the population; it will become a random variable if random sampling is applied to select those observed. The above equations mean that both $P(T(i) = tr_j)$ and $P(R(i, tr_j) = rp_k)$ may be seen as depending on the value of $x(i)$ only, instead of i.

Define

$$S(x) = P(T(x) = tr_1). \tag{8.6}$$

The function $S(x)$ is called the propensity score. The propensity score may, by the assumption made, be seen as a function defined for every individual i in the population: $S(i) = S(X(i))$.

Proposition 8.1. *With the above assumptions made, the propensity score (8.6) is a balancing score in the sense that*

$$P(T(i) = tr_j | X(i) = x(i), S(i) = s(i)) = P(T(i) = tr_j | S(i) = s(i)),$$

Proof. First note that

$$P(T(i) = tr_j | X(i) = x(i), S(i) = s(i)) = P(T(i) = tr_j | S(i) = s(i)),$$

because by the definition of the propensity score, the conditional probability is equal to $s(i) = s(x(i))$, irrespective of what is the value of x, so there is no need to condition, in addition to $s(i)$, also on $x(i)$.

As seen in (6.30), if X and thus $S(X)$ were also random variables (which they will be in a random sample), then this would be $X \perp\!\!\!\perp T | S(x)$, which was the definition of a balancing score. □

Let Q be a simple random sample from the population of interest, and concentrate on those i in Q, for whom $S(X(i))$ is equal to a specified value, say s. Denote this subsample as Q_s. If there are several such strata in the sample, for different values of s, the following procedure applies to all, and the estimates obtained below may be combined taking into account the relative sizes of the strata.

In Q, X is a random variable, so Proposition 8.1 implies that

$$X \perp\!\!\!\perp T | S,$$

and thus for those in Q_s,

$$X \perp\!\!\!\perp T. \tag{8.7}$$

This gives rise to the following result.

Theorem 8.2. *Let the two treatment groups within Q_s be Q_{s1} and Q_{s2}, with respective sizes $|Q_1|$ and $|Q_2|$. Further, let Y_l be the number of observations giving rp_l in Q_{sl}, $l = 1,2$. Then,*

$$E(Y_1/|Q_1| - Y_2/|Q_2|) = P(R = rp_1|T = tr_1) - P(R = rp_1|T = tr_2),$$

that is, $Y_1/|Q_1| - Y_2/|Q_2|$ is an unbiased estimator of the effect of receiving tr_1 versus tr_2 within Q_s.

Proof. Because of (8.7), within Q_s, the conditional distribution of X given T does not depend on T, so it is identical in Q_{s1} and Q_{s2}. Thus, any characteristic which depends on X has the same distribution in these two groups. In particular, if $i \in Q_{s1}$ and $i' \in Q_{s2}$, then

$$P(R(i) = rp_1|T(i) = tr_1)) = P(R(i') = rp_1|T(i') = tr_1)$$

and

$$P(R(i) = rp_1|T(i) = tr_2)) = P(R(i') = rp_1|T(i') = tr_2),$$

where the conditional probabilities are to be interpreted irrespective of which treatment i or i' actually received. The above equations mean that the response probabilities to any treatment are the same in Q_{s1} and Q_{s2} i. That is,

$$P_{i\in Q_{s1}}(R(i) = rp_1|T(i) = tr_1)) = P_{i\in Q_{s2}}(R(i) = rp_1|T(i) = tr_1))$$
$$= P_{i\in Q_s}(R(i) = rp_1|T(i) = tr_1))$$

and

$$P_{i\in Q_{s1}}(R(i) = rp_1|T(i) = tr_2)) = P_{i\in Q_{s2}}(R(i) = rp_1|T(i) = tr_2))$$
$$= P_{i\in Q_s}(R(i) = rp_1|T(i) = tr_2)),$$

and thus

$$P_{i\in Q_s}(R(i) = rp_1|T(i) = tr_1) - P_{i\in Q_s}(R(i) = rp_1|T(i) = tr_2)$$

$$= P_{i\in Q_{s1}}(R(i) = rp_1|T(i) = tr_1)) - P_{i\in Q_{s2}}(R(i) = rp_1|T(i) = tr_2)).$$

Then, $Y_1/|Q_1|$ and $Y_2/|Q_2|$ are unbiased estimators of the last two quantities. □

Although $P(R = rp_1|T = tr_1) - P(R = rp_1|T = tr_2)$ is often interpreted as the causal effect of receiving tr_1 as opposed to tr_2, it may perhaps be more precisely called the effect attributable to receiving tr_1 as opposed to tr_2.

In practical applications of the machinery developed above, in addition to the variables showing treatment and response, the data contain a number of other variables, often called covariates, which may or may not have the property X was assumed to have so far. Nevertheless, let us use X to refer to them. In the first step of the analysis, those components of X are sought for, which appear to be related to receiving or not receiving the treatment. The actual method of analysis is often the one to be described in Sect. 11.1, or if X also has continuous components, some variant of logistic regression is used. As a result, estimates are obtained as to how being in a certain category of a component of X affects the odds of receiving versus not receiving the treatment. Second, using these coefficients, for each individual in the data, an estimate for its odds of receiving versus not receiving the treatment is obtained. In the third step, those observations, for which the odds are similar, are collected as Q_s, and Theorem 8.2 is used to estimate the effect within each such Q_s, and finally these estimates are combined using the relative sizes of the Q_s groups.

This method of estimating causal effects from observational data is often called Rubin's causal model, and its essence is the definition of such strata, so that exchangeability holds within each, and then the possible comparison of effects gives the same result as the counterfactual comparison would give, at least in expectation.

The other remarkable approach to causal analysis is based on [62]. Another excellent reference is [87], which discusses many of the philosophical and applied aspects of the theory. Pearl's theory of causality relies on the concept that causes may only be factors that may be manipulated, and applying a treatment and observing the response to it differ from conditioning on it, which is selecting the cases when it occurred.[18] Accordingly, a different calculus is needed to infer effects. The methods of establishing and analyzing causal effects suggested by Pearl are closely related to directed graphical models, which will be briefly discussed in Chap. 13. The fundamental idea is illustrated next.

Assume that the phenomenon to be analyzed is fully described by three variables, say X, Y, and Z. This assumption means that all other possible aspects are jointly independent from these three variables.[19] In this setup, one is interested in finding out whether effects exist among the three variables and if they do, in determining their directions. If the existence of such effects is established, they are often called causal effects. Note that the meaning of effect (or of causal effect) is not yet decided, rather this concept is operationalized now in an intuitively attractive way, one that will make it possible to decide about existence and direction. The building up of an intuition which corresponds to this theory is also helped by applying directed graphs to illustrate the causal effects to be established; see later. In the universe defined, all causal effects, if any, occur among the three variables. Consider first variables X and Y. If they are independent, then there is neither an effect between them nor is the third variable influencing both of them. The first of these conclusions seems intuitively clear; the second one needs some explanation. In general, one may expect that if Z has an effect on both X and Y, then with Z increasing (more precisely, when Z is made to grow), either both would increase or both would decrease or, perhaps, one would decrease, the other one decrease. In either of these cases, X and Y would not be independent from each other.

[18] Of course, this is very close to the distinction between a designed experiment and an observational study.

[19] This is a very strong assumption, and its applicability will be discussed later.

Of course, the general expectation described above is not correct in all cases. However, if effect was a linear effect, so that

$$X = aZ + E_X, \;\; Y = bZ + E_Y, \tag{8.8}$$

where a and b are constants and the expectation of both E_X and E_Y is zero and are both independent of Z and from each other, then the covariance between X and Y would be

$$\begin{aligned} E(X - E(X))(Y - E(Y)) = E(XY) - E(X)E(Y) = \\ E(abZ^2 + aZE_Y + bZE_X + E_X E_Y) - E(aZ)E(bZ) = \\ abE(Z^2) - ab(E(Z))^2 = ab\ Var(Z) \end{aligned} \tag{8.9}$$

thus if Z is really a variable, that is, it has positive variance, then the covariance is not zero, except if at least one of a and b is zero. If, in addition, multivariate normality is assumed, then zero covariance is equivalent to independence. Therefore, this setup justifies the conclusion that if X and Y are independent, then a and b cannot be both different from zero. The interpretation of E_X and of E_Y is that they are not variables of interest on their own, rather represent some kind of measurement or sampling error.

On the other hand, if effect is not identified with a linear effect, it can very well happen, that X and Y are independent, yet they are both not independent from Z. Such a situation is illustrated for three binary variables in Table 8.3. The three two-way marginals of the distribution are shown in Table 8.4. It is seen easily that while X and Y are independent, Z is not independent from (or, if one wishes, has effects on) both X and Y. This is related to the presence of a second-order interaction, that is, that the value of the second-order odds ratio is not 1. See also the comments about the comparison of multivariate normal data and categorical data in Sect. 1.3.[20]

Table 8.3 X and Y are independent, but Z is related to both

	$Y : yes$	$Y : no$	$Y : yes$	$Y : no$
$X : yes$	0.05	0.2	0.07	0.08
$X : no$	0.15	0.35	0.03	0.07
	$Z : yes$		$Z : no$	

To continue the exposition of the ideas underlying Pearl's theory of causality, the situation when X and Y are independent excludes the possibilities that Z would have an effect on both and that one of them would have an effect on the other. This latter setup, one having an effect on the other one, could occur not only in the way of X influencing Y directly or the other way round but also indirectly through Z.

[20] For a more technical discussion of how the ideas presented here may be extended to categorical data, see [45].

Table 8.4 The two-way marginals of the distribution in Table 8.3

	Y : *yes*	*Y* : *no*		*Z* : *yes*	*Z* : *no*		*Z* : *yes*	*Z* : *no*
X : *yes*	0.12	0.18	*X* : *yes*	0.25	0.15	*Y* : *yes*	0.20	0.10
X : *no*	0.28	0.42	*X* : *no*	0.40	0.10	*Y* : *no*	0.55	0.15

Adopting the convention that a direct effect is denoted by an arrow, certain causal situations may be illustrated with directed graphs. Of course, the assumption that causality is something that may be described by arrows between the nodes of the graphs representing the variables of interest is another simplification applied in this theory.

Figure 8.1 shows all possible directed graphs for three nodes (variables). In general, graphs [s] and [z] are not identified with any causal structure. One reason for this is that both represent cyclic structures, and in both any variable has an effect on any other, so the causal structures could not be distinguished.

When X and Y are independent, many of the graphs in Fig. 8.1 are excluded. Graphs [a], [b], [g], [h], [i], [j], [o], [p], [q], [r], [t], [u], [v], [x], [y], and [z] mean direct effect between X and Y, which is not possible. Graphs [l] and [m] show indirect effects through Z. Further, graphs [c], [d], [e], [f], and the unmarked last graph, the so-called empty graph, show, in addition to X and Y, other pairs of variables unrelated or independent, which is not a claim we wish to make in the present situation. Graph [n] shows effects of Z on X and on Y, which was excluded when the latter two variables are independent. One is left with graph [k] to describe the causal effects when X and Y are independent. This graph means that X and Y both have effects on Z and no other effects are present.

A number of comments are needed to put the foregoing argument in a proper context. The argument is not a deduction which, based on simple principles, derives the direction of causal effects. Rather, it is a somewhat heuristic argument which builds on intuition and develops intuition further. It is assumed that the three variables considered fully describe the causal structure. It is assumed further that arrows in a directed graph can represent causal effects (true when linear effects are considered for multivariate normal data, not necessarily true for categorical data). One of the main advantages of this view is that the study complex causal systems can be facilitated by the study of complex directed graphs. Directed graphs which contain a directed cycle, just like the graph [z], are excluded, and only directed acyclic graphs, DAGs, are given such interpretation. Statistical models based on DAGs and their generalizations will be described briefly in Chap. 13.[21] The presence of an arrow means there may be an effect but does not necessarily mean that there is an effect. But if there is an effect, it goes in the direction of the arrow.

Another remarkable situation is when X and Y are conditionally independent given Z. There are two types of graphs which are considered as possible descriptions

[21] It is also going to be pointed out in Chap. 13 that graphs cannot describe the structure of relationship among categorical variables, unless an explicit assumption of no higher order interactions is made; see the discussion of path models.

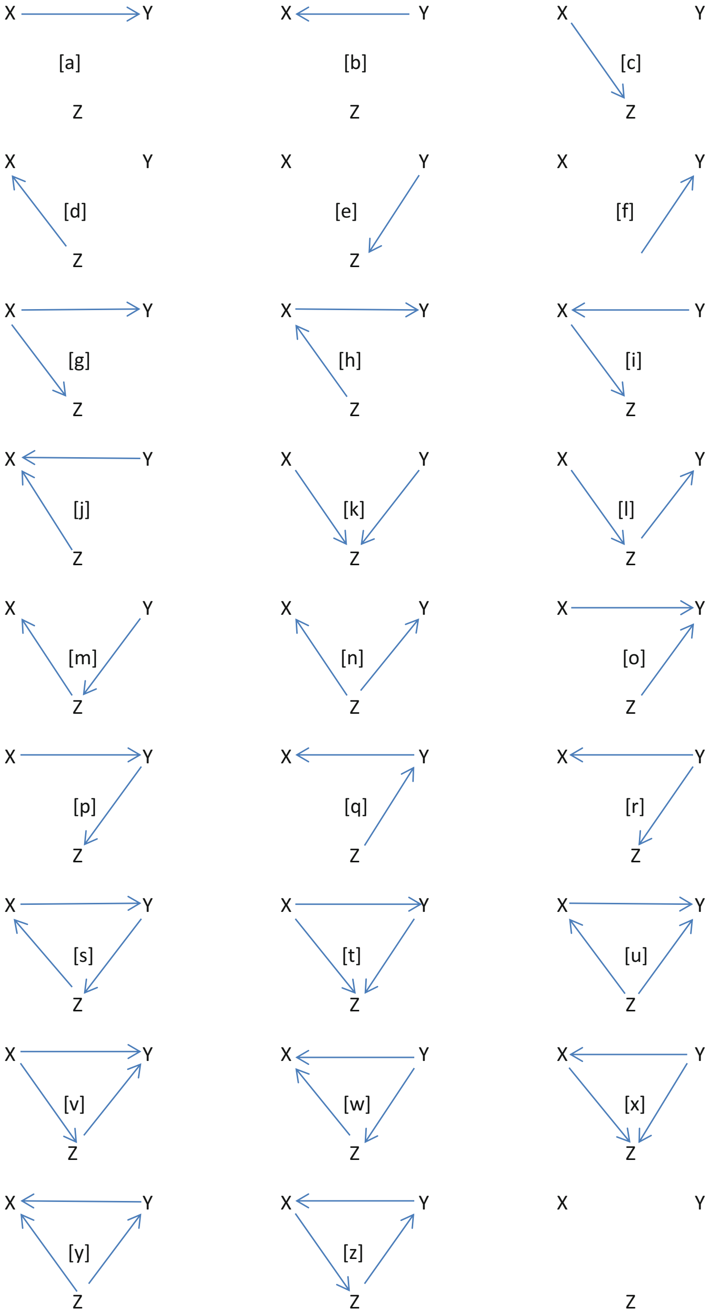

Fig. 8.1 Graph representation of causal relationships among three variables

of the causal structure in such cases. Either Z has an effect on both X and Y, like in graph [n], or one of X and Y has an indirect effect on the other one, through Z, like in graphs [l] and [m]. Conditional independence is linked to the causal structures depicted in these graphs by saying that if some intervention kept Z constant, X and Y would be independent, just like as if conditioned on a particular category of Z. In the setup described in (8.8), when Z is held constant, the covariance of X and Y in (8.9) is zero, because if Z is held constant, its variance is zero. This is the causal structure described in graph [n] and similarly in the causal structures described by graphs [l] and [m]. If effect is interpreted as in (8.8), then for [l], the second equation holds, and if Z is held constant, Y remains constant. As a constant is independent from any variable, when [l] holds and Z is held constant, X and Y are independent. In the data, this relationship manifests itself, as X and Y being conditionally independent, given Z. In the relationship depicted by graph [m], the roles of X and Y are swapped, but otherwise the same argument applies.

The summary is that when only three variables are relevant, and X and Y are independent, then the causal structure is depicted by graph [k], and if X and Y are conditionally independent given Z, one of the graphs [l], [m], and [n] describes the causal structure in this theory.

Such causal arguments are applied quite frequently in many fields of science, including economics, biology, and sociology. In econometrics, essentially the same idea is used under the name of instrumental variable. Suppose, the question to answer is whether X has an effect on Z or Z has an effect on X. The instrumental variable is Y, which is supposed to influence Z but not X directly and is also supposed to be not influenced by X directly. Thus, there is no arrow between X and Y, and there is an arrow pointing from Y to Z. Now, if X and Y are independent, [k] is the causal structure, so the effect goes from X to Z. On the other hand, if X and Y are conditionally independent, given Z, then the graph is [m], and the effect goes from Z to X. Among other relevant materials, [60] contains an example of the application of the instrumental variable method.

8.5 Things to Do

1. Assume that in the data, almost all observations are in cell $(1,1)$. Do these data imply association or independence?
2. Generate a 2×2 table, where the distribution of the two variables is independent, but the distribution is almost entirely concentrated in the off-diagonal cells.
3. If younger patients tend to be treated with the new surgical procedure and older patients tend to be treated with the old procedure, the difference in recovery times may be a consequence of the different treatments but also of the different conditions of the patients. The effects of treatment and of age cannot be separated and are said to be confounded. Why is it that randomization makes confounding unlikely?

4. Why is it that randomization does not eliminate the effect of the expectation of the subjects as to the effect of a particular treatment, and, thus, blind designs are needed?
5. Suppose subjects in an experiment receive a new drug in the form of a yellow pill that they take twice daily. What would be the placebo in this case?
6. Why is it that if the subjects participating in an experiment have an expectation regarding the effect of being treated, irrespective of the actual treatment received, randomization removes this effect from the comparison of the responses given in the two treatment groups? Does this remain true if one of the treatments is a placebo?
7. Explain why comparing sales volume after a commercial campaign to sales volume before it is not an experiment to estimate the effect of the campaign.
8. Formulate Theorem 8.1 for more than two treatments.
9. Suppose some of the participants in a longer experiment to compare two different treatments for a medical condition feel so well that they stop participating. How does this affect the results?
10. Suppose in the experiment described above, some participants die because of the medical condition the experiment was intended to investigate. What is the effect on the results?
11. Suppose some participants of the experiment described above die because of reasons independent from the medical condition investigated. What is the effect on the results?
12. A physician working in a clinic finds that there are several patients attending the clinic with a particular medical condition and observes that many of them come from a certain part of the town. Before concluding that there may be an environmental factor, like pollution, present in that part of the town that leads to this medical condition, she chooses, as control group, patients who attend the clinic for other reasons and finds that a much smaller fraction of them is from that particular part of the town. Is it now justified to conclude that there may be an environmental factor in the background?
13. Is the study above experimental or observational? If observational, are the controls matched? If experimental, who performed the randomization?
14. Is it justified, based on the study above, to conclude that there are more cases in that part of the town?
15. Find three research articles dealing with the adverse effect of smoking on humans. What data are those articles based upon? Did the authors conduct experiments with human beings? If not, what is the basis of the conclusions suggested?
16. Formulate and prove Theorem 8.2 for the case when the sample contains individuals with different values of the propensity score.

Chapter 9
Simpson's Paradox

Abstract Variants of a widely discussed problem related (but not restricted) to causal inference are called Simpson's paradox. In one version, the paradox is that while a new drug may be better than the old drug for both male and female patients, when the data are combined, for all individuals, the old drug appears better. In these cases, the odds ratio is used to determine which treatment is better. First, the paradox is illustrated, and a brief overview of some of the published arguments is given, which aim at explaining what is wrong. Most of these theories say that the paradox occurs as a result of properties of the data or of the data collection procedure. This chapter takes a different position. It is argued that the odds ratio may not be appropriate to measure effect size, because it fails to take into account how popular the compared treatments were, which is relevant information collected in observational studies. A competing, consistent measure of effect (and a concept of effect) is developed, which never commits the paradox. Finally, the last section does not suggest neither the odds ratio nor the measure developed in the previous section to be used universally; rather, it is argued that for a good choice of the better treatment, additional aspects, not only the numbers of positive and negative responses, need to be taken into account.

The chapter starts with presenting facts related to the lack of apparent consistency between decisions based on marginal and conditional tables, which is seen by many as paradoxical.

9.1 Simpson's Paradox

Simpson's paradox refers to a surprising (paradoxical) situation, when inference is to be made based on data, but the difficulties are not related to statistical inference, that is, generalization from data observed in the sample to the underlying population, rather, would also be present if one had access to population data.

The following example, motivated by [21], illustrates a real situation, where a policy decision is needed, based on the analysis of causal or evidential effects. You enter your new job, as shipping manager of a company which sells goods through mail order. The CEO tells you that there have been in the past too many complaints from customers concerning shipping problems, and you should analyze the existing records, and based on your analysis, in the future, only the services of the shipping company that has performed better so far will be used. Your company ships twice a day, in the morning and in the afternoon, and the previous shipping manager has left detailed records of how well the two shipping companies, A and B, performed. The data (hypothetical) are shown in Table 9.1. Using the data in Table 9.1, the records of the two shipping companies can be summarized as shown in Table 9.2. Table 9.2 is a marginal table of Table 9.1, and the data in Table 9.1 are the conditional distributions from Table 9.2.

Table 9.1 Shipment data

Morning shipments			Afternoon shipments		
	OK	Not OK		OK	Not OK
Company A	500	15	Company A	5000	25
Company B	6000	150	Company B	2000	4

Table 9.2 Summary shipment data

All shipments		
	OK	Not OK
Company A	5500	40
Company B	8000	154

Using the material in Sect. 8.1, you compare the conditional odds of a delivery without a problem to a delivery resulting in a customer complaint, for the two shipping companies. For A, the odds is $5500/40 = 137.5$, and for B, it is $8000/154 = 51.95$, so company A has performed better, by a large margin: based on the available data, the odds ratio is $137.5/51.95 = 2.65$, quite far from 1. You also want to see whether the advantage of using the services of company A is the same in the morning as it is in the afternoon. The odds ratio in the morning is $(500/15)/(6000/1500) = 0.83$, and in the afternoon, it is $(5000/25)/(2000/4) = 0.4$.

So it turns out that for all shipments, company A is doing better. But for both morning and afternoon shipments, B is clearly performing much better. If you do not think that every shipment is either in the morning or in the afternoon, then you should use the services of company A. But if you think that every shipment is in the morning or in the afternoon, then irrespective if morning or afternoon shipments are concerned, you should choose company B. This is quite confusing.[1]

The possibility illustrated in the above example is that both conditional odds ratios are on the same side of 1 and the marginal odds ratio is on the other side of 1, which is called Simpson's paradox. In fact, Simpson's paradox has even more disturbing variants to it. By relabeling the variables in Tables 9.1 and 9.2, one may imagine a situation when two drugs, the old one and a new competitor, are compared, and it turns out that for all patients, the old drug is better, but for both male and female patients, the new drug is better.

You may try to avoid confronting the difficulty of Simpson's paradox by a number of arguments.

You may say real data sets with such properties do not exist. Unfortunately, there is a decent number of published cases, where Simpson's paradox did occur in reality. The first such widely discussed case was reported by Bickel et al. [12]. It was noted that the graduate school of the University of California, Berkeley admitted a higher fraction of the male applicants than that of the female applicants. As there was no reason to believe that the male applicants would be better prepared to graduate studies than the female applicants, the question of possible gender discrimination was raised. But the university refuted this argument by arguing that admission decisions were made at the departmental level, and in each department, a larger fraction of female applicants were admitted than of male applicants,[2] so the real advantage was on the side of the female applicants. The explanation given was that some departments were easy and others difficult to get in. Male applicants applied in disproportionately large fraction to departments where it was easy to be admitted, while female applicants rather applied to departments where it was difficult to get in. Thus, although, in (almost) every department, female applicants were admitted in larger fraction, overall, a larger fraction of male applicants got admitted. In this case, the conclusion based on the conditional distributions was suggested to be relevant.

Whether or not one feels this explanation is satisfactory in all inferential aspects, technically this is certainly correct. A similar feature can be seen with the data in Table 9.1: Deliveries in the afternoon work out much better for both companies, and a larger fraction of shipments by company A occurred in the afternoon than of shipments by company B, see the discussion later.

Another real example of Simpson's paradox was published in [91]. Those who wish to apply to medical school in the USA have to take a biology-related test. However, not all of those who take the test apply to medical school. When analyzing the entire population of test takers, it turned out that at each level of the test score, a larger fraction of the black test takers applied to medical school than of white test takers. Disregarding the test score, a larger fraction of the white test takers, than of

[1] In the biostatistics literature, a somewhat different phenomenon is called Simpson's paradox, and the properties and the main results are different.

[2] This was only approximately true, in the sense that this was the case in the largest departments and where the difference in admission rates was not negligible. Other examples, to be discussed, showed exact Simpson's paradoxes.

the black test takers, applied to medical school. Are then the black or the white test takers applying to medical school in larger fractions?

One cannot say, therefore, that data exhibiting Simpson's paradox do not occur in reality. Another possible argument is that the question you, as shipping manager, have to answer is not which shipping company worked better in the past, rather, which shipping company may be expected to work better in the future. The question is about intervention, that is, a policy decision, namely, which company to choose to receive a better-quality service. As was described earlier, such questions may only be answered using data that arise from a designed experiment. As there is no information how the data were collected, in particular, how the shipments were allocated to the two companies, no predictions may be made. This argument is essentially correct in the current example, but does not handle the issue of Simpson's paradox in all possible cases. Your CEO may argue that yes, this is a policy decision that you are about to make, but what could be a better predictor of expected future performance than the observed past performance? The CEO may argue further that whatever the allocation mechanism was in the past, unless you design a new procedure to allocate shipments to the two companies, the existing data remain the best predictor of what could be expected. The question, whether or not this argument is correct, is not one that has a clear answer. There is certainly no mathematical argument that would decide, because the difference in the views held by you and by the CEO is not a matter of correct or incorrect inference within a model (here model means a collection of relevant characteristics) of reality, rather choosing between different models. Based on this argument, the CEO insists you pick the better company, and you are back at Simpson's paradox. In some important real problems, the pressure to make a decision based on the existing information, even if the appropriateness of the data is not clear, may be very strong. For example, if there is a serious epidemy in a country, a public health official may feel a strong pressure, and certainly a strong need, for quickly choosing the best vaccine against the illness, even if the data available about the efficacy of the available vaccines are far from perfect.

As a final obstacle to resorting to the argument that the data are not appropriate to make a decision, it may very well turn out that the previous shipping manager applied random allocation, as was described in Sect. 8.3.1. Indeed, there is no guarantee that a designed experiment cannot yield data exhibiting Simpson's paradox.

Another argument you may agree with, which is often used to "explain" what is behind Simpson's paradox, is considering it as an extreme case of confounding. For illustrations of this view, see, e.g., [59] or [65]. The goal of an experimental setup is to be able to attribute differences in responses to differences in treatments. So the treatment groups are different according to the treatment they received, but they may also be different according to other characteristics. This is usually not a major concern in a designed experiment, because randomization makes such differences unlikely. But in observational studies, such differences may exist. For example, in a case-control study, patients affected by a certain disease are compared to those not affected by this disease, from the point of view of exposure to a risk factor. But those in the exposed group may be different from those in the not exposed group in many ways, in addition to exposure. For example, if the risk factor is having been exposed to alcohol for an extended period of time, say 20 years, then those in the exposed group will tend to be older than those in the not exposed group. Or, if exposure to an occupational risk is investigated, the two treatment groups may have different gender distributions, e.g., the exposed group may contain much more men than the not exposed group. In such cases, the additional difference (age or gender in these examples) is said to be a confounder. That is, a confounder is a factor, whose effect cannot be distinguished from the effect of the treatment in the actual data collection procedure. Matching may help in reducing the severity of the implications of this problem, but in many cases, the limited number of matching variables does not make it reasonable to think that the problem may not occur.

So, what confounding variables may be present in the shipping data? Without further speculation about how the data came about, there is one option, and this is whether the shipment was done in the morning or in the afternoon. Table 9.3 shows the results of morning and afternoon shipments: the afternoon shipments resulted in much fewer complaints than the morning shipments. Also, Table 9.4 shows that this is the case, irrespective of which company was used for the shipment. In this form of data, there is no paradox.[3] You may suggest to the CEO that the data do not show the effect of the shipping company; rather they show the effect of the time of the day; just confounding present in the data covered this fact. Indeed, the company that appeared to perform better did a much higher fraction of afternoon deliveries. So, you may suggest choice is not between the companies, rather between the shipping times. Giving up shipping in the morning seems to improve customer satisfaction. Unfortunately, the CEO may very well tell you that a major factor of customer satisfaction is that you promise that the goods ordered will leave the warehouse within 12 hours, so there is no way to give up morning shipping, and you have to choose one of the companies or else the CEO may choose a different shipping manager.

Table 9.3 Summary shipment data rearranged

All shipments		
	OK	Not OK
Morning	6500	165
Afternoon	7000	29

One also has to be careful not to think that a confounder can only be a variable recorded in the data. It may very well happen, although there is no such information in the available data, that the addresses to which the two companies delivered were not evenly distributed in the area of operation. It may be the case that company *A* not only performed fewer of the morning shipments (which are more problematic than afternoon shipments) but it also delivered more to those areas where people

[3] Of course, the paradox will not go away, in general, just by changing the order of conditioning with all data sets. This is a particular feature of these data.

Table 9.4 Shipment data rearranged

Company *A*			Company *B*		
	OK	Not OK		OK	Not OK
Morning shipment	500	15	Morning shipment	6000	150
Afternoon shipment	5000	25	Afternoon shipment	2000	4

with a weaker tendency to complain live. If company *B* delivered more frequently to areas with more complaint-prone population, its record would appear to be worse than that of company *A*. While “how much ready to complain” may be seen as a confounding factor, most likely there is no variable recorded in the data with that meaning.

The other consequence of the foregoing argument is that, in the case described, the observed data are going to have little predictive power regarding the success rate of delivery, if a new policy to distribute shipments between companies *A* and *B* is applied. This underlines, again, the importance of allocation when causal effects are to be found out from data and, based on the analysis, policy decisions are to be made. If complaint-prone individuals are more likely to be served by company *B* and individuals who are less likely to complain are more likely to be served by company *A*, then allocation into the treatments (company *A* or *B* delivers) is not independent from all possible features of individuals. Random allocation is intended to achieve this independence, but there is no guarantee that after any actual random allocation, the conditional distributions of the treatments would be the same for each and every group of individuals. Therefore, random allocation provides protection against confounding (and possible Simpson paradox) only in expectation, but not in every real application.

Simpson’s paradox can also be presented independently of any interpretation or need to make policy decisions. This approach will be helpful in separating the structural and interpretational aspects of the paradox. If one is given four positive numbers, a, b, c, and d, and each is split into two positive parts as

$$a = a_1 + a_2,\ b = b_1 + b_2,\ c = c_1 + c_2,\ d = d_1 + d_2$$

and it happens to be the case that

$$a_i + d_i < b_i + d_i,\ \text{i=1,2}, \tag{9.1}$$

then

$$a + d < b + d. \tag{9.2}$$

The above implication is obvious. The somewhat surprising fact is that if addition is replaced by multiplication in the last two inequalities, then the implication is not true anymore. For example,

$$5500 = 50 + 5000,\ 40 = 15 + 25,\ 8000 = 6000 + 2000,\ 154 = 150 + 4,$$

then

$$500 * 150 < 15 * 6000 \text{ and } 5000 * 4 < 25 * 2000,$$

but

$$5500 * 154 > 40 * 8000.$$

Thus, while addition and multiplication are both monotone operations, in this context, they behave very differently.

This numerical example (the shipment data) highlights an additional face of the paradox. In both readings of the paradox, one has an expectation, which seems well grounded, but turns out to be wrong. One expectation is that if a drug is better than another one for both male and female patients, then it has to be better than the other one for all patients. The other expectation is that if the inequalities (9.1) hold, then the inequality in (9.2) has to hold, too. That the second expectation is factually wrong gives rise to the idea that, perhaps, the odds ratio is not a good way to measure effect, because it cannot do according to the first expectation. This idea is explored in the next section.

9.2 Consistent Treatment Selection

The material in this section is based on [76] and answers the following question: given that the odds ratio, when used as a measure of effect, does not always lead to decisions which are consistent in the sense that if one treatment is better under all conditions than the other one, then it is also better unconditionally,[4] what other measures of effect could be used instead, to always have a consistent decision? Remember, however, that another measure of effect, in fact, measures effect in a different sense, so effect is defined differently, if one chooses a new measure.

Before answering the question above, one more limitation of the odds ratio as a measure of effect is identified. As discussed in Sect. 6.1.2, the odds ratio is variation independent from the marginal distribution.[5] So in a treatment by response table, the odds ratio is not sensitive to the allocation of the observations in the different treatment categories. This is fine, if the data are from designed experiments. In those situations, the allocation in the treatment categories depends on the decisions of the experimenter and on chance, because of randomization, and therefore, allocation is not informative with respect to the effects of the different treatments. The situation is different, when the data come from an observational study. In this case, the allocation may reflect the preferences of the individuals from whom data are available as to the different treatments. To be more specific, let the data compare two treatments based on two responses in the form of a 2 × 2 table.

[4] Consistent, therefore, means that Simpson's paradox never occurs.

[5] This was seen as a desirable property, when the odds ratio was used to measure the strength of association.

$T =$

Response	Positive	Negative
Tr1	a	b
Tr2	c	d

For simplicity of presentation, the entries a, b, c, and d are assumed to be positive, but some of the results of this section do not require this assumption. The odds ratio compares the conditional odds for a positive versus a negative response:

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}.$$

When its value is greater than 1, $Tr1$ is deemed better, when it is less than 1, $Tr2$. The odds ratio fails to reflect on how popular the treatments are. Its value is the same for the following two data sets.

$T_1 =$

Response	Positive	Negative
Tr1	30	10
Tr2	40	30

$T_2 =$

Response	Positive	Negative
Tr1	30	10
Tr2	4000	3000

The odds ratio is more than 1 for both sets of data, suggesting that $Tr1$ is better than $Tr2$. Although the sample sizes are different and the confidence intervals for the true value of the odds ratio (see Sect. 6.1.1) are of different lengths, the main uncertainty whether it is really the first treatment which is better in both cases has a different reason. In the first set of data, $Tr1$ was given to 40 individuals, $Tr2$ to 70. In the second set of data, the number of those receiving $Tr1$ is the same as in the first set, but the number of those receiving $Tr2$ is 7000, that is 100 times as much as in the first set of data. Does this have any relevance? If the data sets are from experiments, then this difference is unrelated to how useful the treatments are. In a designed experiment, allocation is not informative. If, however, the data sets came from an observational study, then while in the first set of data there is not much difference between the numbers of those who selected either treatment, in the second set of data, $Tr2$ looks much more popular. This is information, which may be relevant for how useful the treatments are. In the case of an observational study, the allocation into treatment categories is potentially informative and, consequently, has to be taken into account when the better treatment is selected. And the odds ratio cannot do this.

In the second set of data, one sees three times as many positive than negative outcomes for $Tr1$, and the same ratio is 1.33 for $Tr2$. But if one looks at not the ratio, rather the difference of positive versus negative outcomes, it is 20 for $Tr1$ and 1000 for $Tr2$. Thus, $Tr2$ led to much more positive outcomes than $Tr1$ did, even if the number of negative outcomes is subtracted. Whether or not the choice of the better treatment should depend on a comparison of ratios or differences will be discussed in the next section. It is clear, however, that the comparison of the differences in the numbers of positive versus negative responses is sensitive to allocation. In the first set of data, the difference is 20 for $Tr1$ and is 10 for $Tr2$; in the second set of data, it is 20 for $Tr1$ and is 1000 for $Tr2$. So while for the first set of data, also the difference deems $Tr1$ better, for the second set of data, $Tr2$ seems better based on the difference in the numbers of positive and negative responses.

To formalize the foregoing argument, let T be the 2×2 treatment by response table and let γ be a decision function, so that when $\gamma(T)$ is 1, $Tr1$ is better; when it is -1, $Tr2$ is better; and when it is 0, the two treatments appear equally good. The sign of the logarithm of the odds ratio (which is called cross product ratio in this context)

$$CPR = sgn(log\frac{a/b}{c/d}) = sgn(log\frac{ad}{bc})$$

is one such decision function, and also the sign of the logarithm of the comparison of differences

$$CSR = sgn(log\frac{a-b}{c-d}) = sgn(log\frac{a+d}{b+c}),$$

called the cross sum ratio, is such decision functions.

In order to be meaningful, decision functions need to have certain properties, which will be called axioms. The first two axioms show data structures, based on which no treatment is better than the other one.

Axiom 1:

$$\gamma\left(\begin{array}{|c|c|}\hline a & a \\ \hline b & b \\ \hline\end{array}\right) = 0.$$

Axiom 2:

$$\gamma\left(\begin{array}{|c|c|}\hline a & b \\ \hline a & b \\ \hline\end{array}\right) = 0.$$

The next two axioms mean that decision functions are antisymmetric in both the treatments and the responses.

Axiom 3:

$$\gamma\left(\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline\end{array}\right) = -\gamma\left(\begin{array}{|c|c|}\hline c & d \\ \hline a & b \\ \hline\end{array}\right),$$

Axiom 4:

$$\gamma\left(\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline\end{array}\right) = -\gamma\left(\begin{array}{|c|c|}\hline b & a \\ \hline d & c \\ \hline\end{array}\right)$$

The last two axioms show data structures when $Tr1$ is better than $Tr2$.

Axiom 5:

$$a > b,\ c \leq d \Rightarrow \gamma\left(\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline\end{array}\right) = 1,$$

Axiom 6:

$$a > c,\ d = b \Rightarrow \gamma\left(\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline\end{array}\right) = 1.$$

Only decision functions fulfilling the above axioms will be considered. Two decision functions γ_1 and γ_2 will be considered identical, if

$$\gamma_1(T) = \gamma_2(T)$$

for all 2×2 treatment by response tables T. A decision function γ is consistent, if the following holds.

$$\text{if } \gamma(T_1) = \gamma(T_2), \text{ then } \gamma(T_1 + T_2) = \gamma(T_1). \tag{9.3}$$

Consistency of γ means that Simpson's paradox never occurs when γ is used to choose the better treatment.

It is clear that all axioms hold for both the *CPR* and the *CSR*. On the other hand, the *CPR* is not consistent, while the *CSR* is consistent, that is, decisions based on it always avoid Simpson's paradox, whatever is the structure of the data available. In fact, the *CSR* is the only decision function with this property.

Theorem 9.1. *Let γ be any decision function subject to Axioms 1–6. Then γ is consistent if and only if it is equal to the CSR.*

Proof. The proof is in three parts, showing that when the *CSR* is 1, -1, and 0, then γ takes on the same value.

First, let $a + d > b + c$.

If $a > b$ and $c < d$, then Axiom 5 implies that $\gamma(T) = 1$.

If $a \leq b$ and $c < d$, then let x be any number such that $d - c > x > b - a$. Then T splits into the following tables.

$$T_1 = \begin{array}{|c|c|}\hline a - \frac{1}{2}min(a,c) & a - \frac{1}{2}min(a,c) \\ \hline c - \frac{1}{2}min(a,c) & d - x - \frac{1}{2}min(a,c) \\ \hline \end{array} \quad T_2 = \begin{array}{|c|c|}\hline \frac{1}{2}min(a,c) & b - a + \frac{1}{2}min(a,c) \\ \hline \frac{1}{2}min(a,c) & x + \frac{1}{2}min(a,c) \\ \hline \end{array}$$

Then $\gamma(T_1) = 1$ because of Axioms 3 and 6, and $\gamma(T_2) = 1$ because of Axioms 4 and 6. Then by consistency, $\gamma(T) = 1$.

If $c \geq d$, then let x be any number such that $a - b > x > c - d$. Then T splits into the following tables.

$$T_1 = \begin{array}{|c|c|}\hline a - x - \frac{1}{2}min(b,d) & b - \frac{1}{2}min(b,d) \\ \hline d - \frac{1}{2}min(b,d) & d - \frac{1}{2}min(b,d) \\ \hline \end{array} \quad T_2 = \begin{array}{|c|c|}\hline x + \frac{1}{2}min(b,d) & \frac{1}{2}min(b,d) \\ \hline c - d + \frac{1}{2}min(b,d) & \frac{1}{2}min(b,d) \\ \hline \end{array}$$

Here $\gamma(T_1) = 1$ because of Axiom 5, and $\gamma(T_2) = 1$ because of Axiom 6. Consistency implies that $\gamma(T) = 1$.

Second, let $a + d < b + c$. Swapping the two columns of T and applying Axiom 4 show that $\gamma(T) = -1 = CSR(T)$.

Third, let $a + d = b + c$.

If $a < c$, choose a positive number x such that $x < min(a, b, c - a)$. Then split T into

$$T_1 = \begin{array}{|c|c|}\hline x & x \\ \hline c - a + x & c - a + x \\ \hline \end{array} \quad T_2 = \begin{array}{|c|c|}\hline a - x & b - x \\ \hline a - x & b - x \\ \hline \end{array}$$

With reference to Axioms 1 and 2, consistency implies that $\gamma(T) = 0$.

If $a > c$, apply Axiom 3 and the previous result to obtain that $\gamma(T) = 0$.

If $a = c$, apply Axiom 1.

□

Note that the concept of equality of decision functions means that equal decision functions always lead to the same decision, not that they would be based on the same formula.

On the other hand, a fundamental property of the *CPR* implies that for some sets of data, Simpson's paradox will occur.

Theorem 9.2. *If Axioms 1–6 hold and*

$$\gamma\left(\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline\end{array}\right) = \gamma\left(\begin{array}{|c|c|}\hline ta & tb \\ \hline uc & ud \\ \hline\end{array}\right) \tag{9.4}$$

for every table and all positive t and u, then γ is not consistent.

Proof. By assumption

$$\gamma(T) = \gamma\left(\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline\end{array}\right) = \gamma\left(\begin{array}{|c|c|}\hline a & b \\ \hline \frac{b}{d}c & \frac{b}{d}d \\ \hline\end{array}\right) = \gamma\left(\begin{array}{|c|c|}\hline a & b \\ \hline \frac{b}{d}c & b \\ \hline\end{array}\right)$$

and by Axiom 6, $\gamma(T) = 1$ if $CPR(T) = 1$, and by Axiom 3, $\gamma(T) = -1$ if $CPR(T) = -1$. Finally, by Axiom 2, $\gamma(T) = 0$ if $CPR(T) = 0$ and the *CPR* is not consistent. □

Property 9.4 can be interpreted as variation independence from allocation: when allocation into the treatment categories changes but the relative frequencies of the responses do not, the same treatment is selected. The next result is about a kind of variation independence from proportional changes in the responses.

Theorem 9.3. *If Axioms 1–6 hold and*

$$\gamma\left(\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline\end{array}\right) = \gamma\left(\begin{array}{|c|c|}\hline ta & ub \\ \hline tc & ud \\ \hline\end{array}\right) \tag{9.5}$$

for every table and all positive t and u, then γ is not consistent.

Proof. By assumption

$$\gamma(T) = \gamma\left(\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline\end{array}\right) = \gamma\left(\begin{array}{|c|c|}\hline a & \frac{c}{d}b \\ \hline c & \frac{c}{d}d \\ \hline\end{array}\right) = \gamma\left(\begin{array}{|c|c|}\hline a & \frac{c}{d}b \\ \hline c & c \\ \hline\end{array}\right)$$

and the proof can be completed similarly to that of Theorem 9.2 □

The previous results show that variation independence from marginal adjustments, in the sense of (9.4) or (9.5), leads to the possibility of Simpson's paradox. This restricts the usability of the odds ratio as a measure of effect, while the same properties are very desirable when the odds ratio is used as a measure of association.

A multivariate generalization of the above results was given in [78]. The setup for the generalization is that there are k binary variables, out of which 1 may be response, while the other $k-1$ are treatments. If for the response variable category 1 is the positive outcome and category 2 is the negative outcome, while for the treatment variables category 1 means treatment and category 2 means no treatment or placebo treatment, then a continuous function γ_k defined on all 2^k distributions such that

$$\gamma_k \text{ is a monotone increasing function of } p(1,1,\ldots,1) \tag{9.6}$$

may be seen as a measure of how much the treatments are associated with the positive outcome. When the value of γ_k is positive, the treatments are associated with the positive response; when it is negative, they are associated with the negative response.

The property that γ_k never commits Simpson's paradox is formulated in the multidimensional case as directional collapsibility. If for any of the variables V_j, the conditional distribution of the remaining variables is denoted as $T_{k-1}(V_j = j)$, then directional collapsibility is that

$$\text{if } sgn(\gamma_{k-1}(T_{k-1}(V_j=1))) = sgn(\gamma_{k-1}(T_{k-1}(V_j=2))), \text{ then}$$
$$sgn(\gamma_{k-1}(T_{k-1}(V_j=1))) = sgn(\gamma_{k-1}(T_{k-1}(V_j=2))) = sgn(\gamma_{k-1}(T_{k-1}(V_j=+))), \tag{9.7}$$

where $T_{k-1}(V_j = +)$ is the marginal distribution of the other variables. Directional collapsibility means that if conditioned on any value of V_j, the other treatments are associated with the positive (negative) response, then the same is true disregarding treatment V_j.

Then [78] gave the following result.

Theorem 9.4. *Let γ_k have the properties (9.6) and (9.7). Then, for every T_k,*

$$sgn(\gamma_k(T_k)) = sgn(logDI_k(T_k)),$$

that is, γ_k, as a decision function is equal to $logDI_k$, where DI_k is defined as follows. Let c be a cell in the 2^k table, let $p(c)$ be the probability in this cell, and let Σ_c denote the sum of the indices of this cell. Then

$$DI_k(T_k) = \Sigma p(c)^{-1^{\Sigma_c}}. \tag{9.8}$$

□

The theorem means that subject to (9.6), the only decision function which always avoids Simpson's paradox is DI_k. It is easily seen that $sgn(logDI_2) = CSR$.

9.3 General Aspects of Measuring Effects and Associations

The results of the previous section indicate that subject to mild conditions, it is essentially only the CSR and its generalizations which, independently of the data, avoid Simpson's paradox, that is, which always lead to consistent decisions. This, however, does not imply that the *CSR* should be necessarily preferred over the *CPR*, or any other measure, when it comes to choosing the better treatment. There are a number of factors which need to be taken into account, when a measure of association or a measure of effect is selected. Many such factors were considered in [77].

An important aspect, already discussed in this book, is whether the data available for the decision-making come from a designed experiment or from an observational study. In the former case, allocation is not informative, and variation independence from the marginal distributions, like in the case of the *CPR*, is desirable. When the data come from an observational study, allocation in the treatment categories is informative, variation independence from the marginals is not advisable, and the *CSR* may be used, which also avoids Simpson's paradox.

The foregoing argument suggests that when the data are from a designed experiment, one cannot avoid running into Simpson's paradox occasionally, because the concept of effect, which is relevant, the one measured by the *CPR*, does not provide overall protection against the paradox. Recall the discussion in Sect. 9.1, explaining that designed experiments, specifically random allocation, do not, in all instantiations, provide data which do not exhibit the paradox, when the *CPR* is used. In expectation, they do, which is an important argument for random allocation, but this should not be taken as an overall guarantee.

A further aspect to take into account, when the better decision is selected, is whether the selected treatment will be offered or imposed. A treatment is offered, if it is made available, and individuals may choose whether or not to use that treatment. Examples of such treatments include training courses offered free of charge to unemployed people to increase their chances of finding a job or making smoking cessation aids available at a subsidized price. A treatment is imposed, if individuals in a certain group will be subject to that treatment. Examples include elementary school education or compulsory vaccination of children. When a treatment to be offered is to be selected, the question of how happily individuals choose that treatment, in comparison to other competing treatments, is of relevance in choosing the best treatment. Not only how efficient is the treatment, if selected, is what determines how strong an effect may be expected on the population level but also how likely is that the treatment will be selected. In such cases, observational data (showing both choice and effect) are relevant, and the *CSR* is to be used. In cases when the selected treatment is to be imposed, experimental data are relevant (showing attributable effects), and the *CPR* is to be used.

The concepts and measures of effect considered so far (*CPR*, *CSR*) assume that some comparison of the numbers of positive and negative responses (ratio or difference) is a good indication of the magnitude of the effect. There may be situations, however, when this is not the case. One usually thinks of the outcomes as improvement or no change or got much better or got a little better. But if the negative response is catastrophic, e.g., the condition that was to be treated got much worse or the patient died, one may wish to avoid the catastrophic outcome, at all, and a treatment with no negative response may be seen preferable to another treatment with some negative responses, irrespective of the numbers of positive responses.

The summary of all these considerations is that the 4 numbers in a 2×2 treatment by response table do not contain all information which is necessary to choose the better treatment. At least, the kind of decision to be made (offer or impose the better treatment), the method of data collection (observational study or designed experiment), and the exact meanings of the positive and negative responses need to be taken into account. A decision function applied to numbers in the treatment by response table cannot take these important differences into account. In other words, when doing so, one tries to apply the same method to different problems, and one cannot be surprised if the results do not always appear consistent logically.

9.4 Things to Do

1. Find the Berkeley admission data in [12] and see how close they are to exhibiting Simpson's paradox.
2. Obtain the medical school application data from [91] and see if you can find a structure similar to the one described for the Berkeley admission data.
3. Study the analysis given in [76] of the medical school admission data.
4. Given the data collection procedures in the Berkeley admission and the medical school application data, determine whether they are experiments or observational studies, if there was sampling applied to select the individuals, and discuss the issue of external validity, if it is applicable.
5. Is it possible that in one country, the death rate (percentage of those who die each year) is higher in another country, but in the latter country, the death rate is higher within any age group?
6. Find a paper which defines Simpson's paradox as the situation when there is no association in the conditional tables, but there is association in the marginal table.
7. Prove that Axioms 1–6 hold for the *CPR* and the *CSR*.
8. Prove that the *CSR* is a consistent decision function.
9. Develop decision functions which are equal to the *CSR* and which are equal to the *CPR*.
10. Develop decision functions which are not equal to neither the *CSR* nor the *CPR*.
11. Decide whether the positivity of a, b, c, and d is required for Theorem 9.1 to hold.
12. Complete the proof of Theorem 9.3.
13. Compare Axioms 1–6 with condition (9.6).
14. Find policies or treatments in a university environment which are offered and which are imposed.

Chapter 10
Log-Linear Models: Definition

Abstract This chapter introduces log-linear models which are the most widely used simple structures in the analysis of categorical data. Their simplicity comes from a multiplicative structure, where the multipliers depend on subsets of the variables, but not on all variables together. These subsets are the allowed interactions, and larger subsets of variables exhibit no interaction, as measured by the conditional odds ratio. When the logarithm of this multiplicative structure is taken, one obtains a linear structure on the logarithmic scale. More formally, a log-linear model is defined by a mixed parameterization of the distributions on the contingency table, introduced first. It consists of conditional odds ratios on an ascending class of subsets of the variables and of marginal distributions on the complement descending class. The log-linear model is obtained by setting all the conditional odds ratios on the ascending class equal to 1. The resulting models include generalizations of independence for three-way tables, which are discussed in detail, like joint, multiple, and conditional independence and also the model of no second-order interaction. Log-linear models may also be equivalently defined through defining a log-linear representation of the cell probabilities and then setting some of the log-linear parameters to zero.

Log-linear models and their various modifications play a central role in the analysis of categorical data. A model in statistics is a set of assumptions with respect to the true population distribution but is also interpreted as the subset of all possible distributions which are characterized by these properties. The fundamental property of log-linear models is a multiplicative structure of the cell probabilities (which, when logarithms are taken, becomes additive). The models assume the existence of various effects, each being present in some of the cells of the contingency table and each having a multiplicative contribution to the cell probabilities, where they are present. The simplest such model is independence in a two-way contingency table. In every cell, there is an effect of the row it belongs to and of the column it belongs to. In this model, the numerical values of the effects are identified with the row and column marginal probabilities:

$$p_{ij} = p_{i+}p_{+j}, \text{ for all } i, j.$$

For a more precise statement, however, see Sect. 10.2, where log-linear models will be introduced as generalizations of independence. Log-linear models assume a special structure of the parameters of the distributions, and parameterizations of discrete distributions are discussed first. Log-linear models are exponential families, and many of the results in Sect. 7.2 apply to them.

10.1 Parameterizations of Multidimensional Discrete Distributions

This section relies directly on the material presented in Sects. 6.1 and 6.2, and the immediate goal is to extend the parameterization given in Theorem 6.3 for distributions on 2×2 tables to distributions on arbitrary contingency tables. There are many such generalizations, using parameters of different meanings.[1]

To obtain such a generalization, consider all subsets of the variables in a partial order with respect to inclusion. For three variables, A, B, and C, such a partial order may be represented as follows:

$$\begin{array}{c} ABC \\ AB\ AC\ BC \\ A\ B\ C \\ \emptyset \end{array}$$

In this array, for every subset, the bigger ones are above it, and the smaller ones are below it. For four variables, A, B, C, and D, the partial order of the subsets of the variables is as follows:

$$\begin{array}{c} ABCD \\ ABC\ ABD\ ACD\ BCD \\ AB\ AC\ AD\ BC\ BD\ \ CD \\ A\ B\ C\ D \\ \emptyset \end{array}$$

If 2^V denotes the power set of the variables, then a class $\mathscr{C} \subseteq 2^V$ is called ascending; if for all subsets of the variables S, $S \in \mathscr{C}$ implies that $T \in \mathscr{C}$, for all $T \supseteq S$. For example, for four variables $\{AB, ABD, ABC\}$ is not ascending, but $\{ABD, ABCD\}$ is ascending. A class $\mathscr{C} \subseteq 2^V$ is called descending, if for all subsets of the variables S, $S \in \mathscr{C}$ implies that $T \in \mathscr{C}$, for all $T \subseteq S$. For example, $\{A, B, AB, ABD\}$ is not descending, but $\{\emptyset, A, B, C, AB, AC, BC, ABC\}$ is descending. An ascending class

[1] That is, with parameters measuring different characteristics of the distribution.

is said to be generated by its minimal elements, and a descending class is said to be generated by its maximal elements.

Theorem 10.1. *If $\mathscr{C}$ is ascending, then $2^V \setminus \mathscr{C}$ is descending. If $\mathscr{C}$ is descending, then $2^V \setminus \mathscr{C}$ is ascending.*

Proof. Suppose $\mathscr{C}$ is ascending but $2^V \setminus \mathscr{C}$ is not descending. This means that there is an $S \in 2^V \setminus \mathscr{C}$, such that for some $T \subseteq S$, $T \notin 2^V \setminus \mathscr{C}$. But then $T \in \mathscr{C}$, and there is an $S \supseteq T$, which is not in $\mathscr{C}$, so $\mathscr{C}$ is not ascending. The other claim is proved similarly.

□

The division of 2^V into an ascending and a complement descending class is called a cut. For example, if one denotes the members of the ascending class by upper case letters, and the members of the descending class by lower case letters, the following is a cut, which corresponds to the last example before Theorem 10.1:

$$\begin{array}{c} ABCD \\ abc\ \ ABD\ \ ACD\ \ BCD \\ ab\ \ ac\ \ AD\ \ bc\ \ BD\ \ CD \\ a\ \ b\ \ c\ \ D \\ \varnothing \end{array}$$

Clearly, the class of all subsets and also the empty class are both ascending and descending. A cut may also contain the empty class and the class of all subsets, in which case it is improper and is proper otherwise.

Every cut defines a parameterization of all discrete distributions on the contingency table. These parameterizations associate parameters with every subset, and the kind of parameter associated with a subset depends on whether the subset is in the descending or ascending class. Therefore, if the cut is proper, the parameterization is mixed. The parameters used are the marginal distributions of the subsets in the descending class, and the conditional odds ratios, given the categories all other variables, for the subsets in the ascending class. When a subset consists of a single variable, instead of the odds ratio, the odds is used, if it is in the ascending class. When the cut is improper and the empty subset is in the ascending class, there is no parameter (conditional odds or odds ratio) associated with it. If it is in the descending class, the marginal distribution is the number 1, and it does not depend on the actual distribution, so while it may be seen as a parameter of it, its value carries no information with respect to the distribution and is omitted.

The simplest example is a two-way contingency table. Here is a complete list of all possible cuts into an ascending and the complement descending class:

$$\begin{array}{cccccc} AB & AB & AB & AB & AB & ab \\ A\ \ B & A\ \ B & a\ \ B & A\ \ b & a\ \ b & a\ \ b \\ \emptyset & \varnothing & \varnothing & \varnothing & \varnothing & \varnothing \end{array}$$

All the six cuts imply different parameterizations of the probability distribution in the two-way table. For simplicity, these parameterizations will be presented first for the case when both variables are binary, and the generalization needed for arbitrary categorical variables will be discussed after it. A parameterization consists of parameters. For the first cut, these are the following:

$$\frac{p(A=1|B=1)}{p(A=2|B=1)}, \frac{p(A=1|B=2)}{p(A=2|B=2)}, \frac{p(B=1|A=1)}{p(B=2|A=1)}, \frac{p(B=1|A=2)}{p(B=2|A=2)},$$

$$\frac{p(A=1,B=1)p(A=2,B=2)}{p(A=1,B=2)p(A=2,B=1)}.$$

One has the conditional odds for each variable, given all possible categories of the other variables and the odds ratio of the two variables. There is some redundancy among the parameters listed above, if one looks at them as a possible parameterization of the distribution, because (see also Proposition 6.5)

$$\frac{p(A=1|B=1)}{p(A=2|B=1)} \Big/ \frac{p(A=1|B=2)}{p(A=2|B=2)} = \frac{p(B=1|A=1)}{p(B=2|A=1)} \Big/ \frac{p(B=1|A=2)}{p(B=2|A=2)}$$

$$= \frac{p(A=1,B=1)p(A=2,B=2)}{p(A=1,B=2)p(A=2,B=1)}.$$

Therefore, the redundant information may be removed. There are many ways to do this, but the most useful is to keep the following parameters:

$$\frac{p(A=1|B=1)}{p(A=2|B=1)}, \frac{p(B=1|A=1)}{p(B=2|A=1)}, \frac{p(A=1,B=1)p(A=2,B=2)}{p(A=1,B=2)p(A=2,B=1)}. \tag{10.1}$$

It can be seen that the parameters in (10.1) are variation independent and they constitute a parameterization of the 2×2 distributions.

The second cut differs from the first one only in the empty set, and the parameterization is not different. For the third cut, the following parameters are obtained:

$$p(A=1), \, p(A=2), \frac{p(B=1|A=1)}{p(B=2|A=1)}, \frac{p(B=1|A=2)}{p(B=2|A=2)},$$

$$\frac{p(A=1,B=1)p(A=2,B=2)}{p(A=1,B=2)p(A=2,B=1)}.$$

To remove the obvious redundancy, it is sufficient to consider only

$$p(A=1), \frac{p(B=1|A=1)}{p(B=2|A=1)}, \frac{p(B=1|A=2)}{p(B=2|A=2)},$$

because the ratio of the two conditional odds is the odds ratio.

The fourth cut is like the third one, but the roles of A and of B interchanged. The fifth cut leads to the following mixed parameterization (with the redundancies removed):

$$p(A=1),\ p(B=1),\ \frac{p(A=1,B=1)p(A=2,B=2)}{p(A=1,B=2)p(A=2,B=1)}. \tag{10.2}$$

This parameterization consists of the two two-way marginal distributions and the odds ratio.

The sixth, improper, cut implies the cell probabilities in the contingency table, as a parameterization of the joint distribution.

That the parameters listed constitute a parameterization of the joint distribution, that is, the distribution may be reconstructed given the parameters, may be easily seen. In particular, Theorem 6.3 states this for the parameters in (10.2). Further, the same theorem states that the components of the mixed parameterization, the marginal distributions, and the odds ratio are variation independent. These results are going to be generalized in the rest of the section.

For distributions on arbitrary $I \times J$ tables, instead of a single odds ratio, all local or all spanning cell odds ratios may be considered. The concept of odds may also be generalized accordingly. For example, instead of the odds

$$\frac{p(A=1|B=1)}{p(A=2|B=1)},$$

one may use the following parameters:

$$\frac{p(A=2|B=1)}{p(A=1|B=1)},\ \frac{p(A=3|B=1)}{p(A=2|B=1)},\ \dots\ \frac{p(A=I|B=1)}{p(A=I-1|B=1)}$$

or

$$\frac{p(A=2|B=1)}{p(A=1|B=1)},\ \frac{p(A=3|B=1)}{p(A=1|B=1)},\ \dots\ \frac{p(A=I|B=1)}{p(A=1|B=1)}.$$

For three-dimensional tables, there are four kinds of cuts and mixed parameterizations based on them, which will be of central importance. The first such cut is

$$\begin{array}{c} ABC \\ AB\ AC\ BC \\ a\ b\ c \\ \emptyset \end{array} \tag{10.3}$$

and the mixed parameterization, in the binary case, consists of

$$p(A=1),p(B=1),p(C=1),COR(A,B|C=k),COR(A,C|B=j),COR(B,C|A=i), \tag{10.4}$$

where $i,j,k=1,2$. The parameters in (10.4) do not contain the second-order odds ratio, which would be redundant, but do contain further redundant terms. In fact, the redundancy is removed, by assuming, e.g., $i=1,2$, $j,k=1$. That the parameters in (10.4) constitute a parameterization will be proved later.

The second relevant cut is

$$\begin{array}{c} ABC \\ ab\ AC\ BC \\ a\ b\ c \\ \emptyset \end{array} \tag{10.5}$$

and the mixed parameterization, in the binary case, consists of

$$p(A=1,B=1),\ p(A=1,B=2),\ p(A=2,B=1),\ p(C=1), \tag{10.6}$$

$$COR(A,C|B=k), COR(B,C|A=i),$$

where $k = 1,2$, $i = 1$. The parameters in (10.6) do not contain redundant terms.

The third relevant cut is

$$\begin{array}{c} ABC \\ ab\ ac\ BC \\ a\ b\ c \\ \emptyset \end{array} \tag{10.7}$$

and the mixed parameterization, in the binary case, consists of

$$p(A=1,B=1),\ p(A=1,B=2),\ p(A=2,B=1),\ p(A=1,C=1), \tag{10.8}$$

$$p(A=2,C=1), COR(B,C|A=i),$$

where $i = 1,2$. The parameters in (10.8) do not contain redundant terms.

The fourth relevant cut is

$$\begin{array}{c} ABC \\ ab\ ac\ bc \\ a\ b\ c \\ \emptyset \end{array} \tag{10.9}$$

and a nonredundant mixed parameterization, in the binary case, consists of

$$p(A=1,B=1),\ p(A=1,B=2),\ p(A=2,B=1), \tag{10.10}$$

$$p(A=1,C=1),\ p(A=2,C=1),\ p(B=1,C=1),\ OR(A,B,C)$$

In this case, the parameters are the three two-way marginal distributions and the second-order odds ratio of the three variables. The $A \times B$ marginal distribution is given with three cell probabilities (the fourth one is implied by the fact that they sum to 1). The $A \times C$ marginal distribution is given by specifying two cell probabilities, because its A marginal is already given. For the $B \times C$ marginal distribution, both of its one-way marginals are already given; thus a single cell probability is sufficient to specify it.

The parameters in (10.10) are not fully variation independent. For example, $p(A=1,B=1)$ and $p(A=1,B=2)$ are not variation independent. But the marginal probabilities on the descending class and the odds ratio on the complement ascending class are variation independent. A general form of this result is given next for strictly positive probability distributions.

Theorem 10.2. *Let V be a set of categorical variables, let $\mathscr{D} \subseteq 2^V$ a descending class, and let $\mathscr{A} = 2^V \setminus \mathscr{D}$ its complement ascending class. Let $\mathscr{P}$ be the set of strictly positive probability distributions on the contingency table formed by the ranges of the variables V. Then, for any $Q, R \in \mathscr{P}$, there exists a unique $P \in \mathscr{P}$, such that*

$$\textit{for all } D \in \mathscr{D},\ P(D,+) = Q(D,+),$$

where $P(D,+)$ means the marginal distribution of P on D and

$$\textit{for all } A \in \mathscr{A},\ COR_P(A|V \setminus A) = COR_R(A|V \setminus A),$$

where $COR_P(A|V \setminus A)$ means the conditional odds ratios of P for the variables A, given the categories of all other variables.

Note that the conditional odds ratio for a given condition may be a single odds ratio if the conditional table is binary and may be a collection of local or spanning cell odds ratios otherwise. In the latter case, the claim applies to each element in this collection.

The theorem means that if a feasible set of marginal distributions on the descending class and a feasible set of conditional odds ratios on the complement ascending class are considered, then these can always be combined to yield a distribution, and there is only one such distribution. The feasibility of the marginal distributions and of the conditional odds ratios is achieved by taking these parameters from existing distributions. Thus the theorem states that the mixed parameterization is, indeed, a parameterization and that its two components are variation independent.

Proof. The theorem is a consequence of the convergence of the Iterative Proportional Fitting Procedure, to be proved in Sect. 12.2. □

Mixed parameterizations of exponential families were considered in Sect. 7.2. It will be shown next that canonical statistics for the exponential family $\mathscr{P}$ may be defined in such a way that the mean value parameters are (equivalent to) the marginal distributions on the descending class and the canonical parameters are (equivalent to) the conditional odds ratios on the complement ascending class.

The construction will only be presented here for the case, when all variables are binary and the subsequent results in this section are proved in that case only. In the construction, a canonical statistic is associated with every subset of the variables, so the number of cells and the number of canonical statistics are the same. Let $W \subseteq V$ be a subset of the variables and $\mathbf{t}_W$ the canonical statistic associated with it. Let the categories of the variables be denoted as 1 and 2, and let the index of a cell be denoted as (w,w'), where $W = w$ and $W' = V \setminus W = w'$. In every cell, the value of $\mathbf{t}_W$ is

$$\mathbf{t}_W(w,w') = (-1)^{S_w}, \tag{10.11}$$

where S_w is the sum of the indices of the variables which belong to W, in the actual cell. For example, in the case of a 2×2 table, with the lexicographic ordering of the cells and the subsets ordered as $\emptyset$, A, B, and AB, the design matrix is

$$\mathbf{T}_1 = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

and for a $2 \times 2 \times 2$ table, with a similar order of the subsets, it is

$$\mathbf{T}_2 = \begin{pmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

From now on, let $\mathbf{T}$ denote such a design matrix.

Proposition 10.1. *Any two distinct columns of* $\mathbf{T}$ *are orthogonal to each other and so are any two distinct rows of it.*

Proof. Every column of $\mathbf{T}$ is associated with a subset of the variables. For two columns, these subsets may be disjoint, or may intersect.

If the two subsets, say W_1 and W_2, are disjoint, then every combination of the indices of the variables in W_1 occurs with every combinations of the indices of the variables in W_2, and half of the combinations for W_1 yields a 1, and the other half yields a -1 in the relevant position in the design matrix, and the same applies to W_2. Therefore half the rows, where in the column of W_1, there is a 1, contains a 1 in the column of W_2, and half of them contains a -1 and similarly for the rows which have a -1 in the column of W_1.

If $W_1 \cap W_2 \neq \emptyset$, then the same argument applies to the indices of $W_1 \setminus W_2$ and of $W_2 \setminus W_1$. Half of the indices of the latter variables change the sign associated with the indices of W_1.

The proof for the rows is similar. □

Then the inverse of the design matrix is obtained easily.

Proposition 10.2. *The inverse of the design matrix* $\boldsymbol{T}$ *is*

$$\frac{1}{|2^V|}\boldsymbol{T}'.$$

Proof. As $\mathbf{T}$ a $|2^V| \times |2^V|$ matrix, every column multiplied by itself gives $|2^V|$. The rest is implied by the previous proposition. □

As the inverse of a transposed matrix is the transpose of the inverse, the previous results imply the following theorem:

Theorem 10.3. *The mean value parameters*

$$\boldsymbol{T}'\boldsymbol{p} \tag{10.12}$$

are a parameterization of the distribution $\boldsymbol{p}$*, and the distribution can be obtained as*

$$\boldsymbol{p} = \frac{1}{|2^V|}\boldsymbol{T}\left(\boldsymbol{T}'\boldsymbol{p}\right). \tag{10.13}$$

□

If in a mixed parameterization, the mean value parameters are given on a descending class, then the marginal distributions on the subsets in the descending class may be reproduced and vice versa.

Theorem 10.4. *Let* $\mathscr{D}$ *be a descending class. Then, for any* $D \in \mathscr{D}$*, the following sets of parameters may be calculated from each other: the mean value parameters*

$$\boldsymbol{t}'_E\boldsymbol{p}, \textit{ for all } E \in D$$

and the marginal distributions

$$\boldsymbol{p}_E \textit{ for all } E \in D.$$

Proof. The claim in implied by applying Theorem 10.3 to D. □

Therefore, giving the mean value parameters or giving the marginal distributions on a descending class of subsets of variables is equivalent.

In the exponential family representation,

$$\log \mathbf{p} = \mathbf{T}\boldsymbol{\theta}.$$

Therefore,

$$\boldsymbol{\theta} = \frac{1}{|2^V|}\mathbf{T}'\log \mathbf{p}.$$

Thus, each component of $\boldsymbol{\theta}$ is the inner product of a canonical statistic with the vector $\log \mathbf{p}$ and is associated with a subset of the variables. Consider the component associated with $W \subseteq V$. Out of the cells of the contingency table, choose those, where the indices of the variables $V \setminus W$ have specific values, say $V \setminus W = w'$. Introduce the following notation:

$$(\mathbf{t}_W, \log \mathbf{p})_{w'} = \Sigma_w \mathbf{t}_W(w, w') \log \mathbf{p}(w, w'),$$

where Σ_w denotes summation for all possible indices of the variables in W and $\mathbf{t}_W$ is the canonical statistic associated with W. Then one has

Theorem 10.5. *With the notation above,*

$$(\boldsymbol{t}_W, \log \boldsymbol{p})_{w'} = (-1)^{|W|} logCOR(W|V \setminus W = w'),$$

and

$$(\boldsymbol{t}_W, \log \boldsymbol{p}) = \Sigma_{w'}(\boldsymbol{t}_W, \log \boldsymbol{p})_{w'}.$$

Proof. The first statement follows from the definition of the conditional odds ratios, the second follows from the first one. □

For example, in the design matrix $\mathbf{T}_2$, the second column is for the subset consisting of variable A. One possible choice for w' is $(1,1)$. Then

$$(\mathbf{t}_{\{A\}}, \log \mathbf{p})_{(1,1)} = \mathbf{t}_{\{A\}}(1,1,1) \log p(1,1,1) + \mathbf{t}_{\{A\}}(2,1,1) \log p(2,1,1) =$$

$$-\log p(1,1,1) + \log p(2,1,1) = \log \frac{p(2,1,1)}{p(1,1,1)} = (-1)^1 \log \frac{p(1,1,1)}{p(2,1,1)}.$$

For the other choices of w', one obtains, respectively,

$$log \frac{p(2,1,2)}{p(1,1,2)},$$

$$log \frac{p(2,2,1)}{p(1,2,1)},$$

$$log \frac{p(2,2,2)}{p(1,2,2)}.$$

The penultimate column of $\mathbf{T}_2$ is associated with $W = \{B,C\}$. There are two choices for w', 1 and 2. For the first choice,

$$(\mathbf{t}_{\{B,C\}}, \log \mathbf{p})_{(1)} =$$

$$\mathbf{t}_{\{B,C\}}(1,1,1) \log p(1,1,1) + \mathbf{t}_{\{B,C\}}(1,1,2) \log p(1,1,2) +$$

$$\mathbf{t}_{\{B,C\}}(1,2,1) \log p(1,2,1) + \mathbf{t}_{\{B,C\}}(1,2,2) \log p(1,2,2) =$$

$$\log p(1,1,1) - \log p(1,1,2) - \log p(1,2,1) + \log p(1,2,2) =$$

$$\log COR(B,C|A=1).$$

For the other choice of w', one obtains

$$\log COR(B,C|A=2).$$

Then, by the second part of Theorem 10.5,

$$(\mathbf{t}_{\{B,C\}}, \log \mathbf{p}) = \log(COR(B,C|A=1) + \log(COR(B,C|A=2)).$$

Further,

$$(\mathbf{t}_{\{A,B,C\}}, \log \mathbf{p}) = \log OR(A,B,C) = \log \frac{COR(B,C|A=1)}{COR(B,C|A=2)}$$

$$= log(COR(B,C|A=1) - \log(COR(B,C|A=2)).$$

From the last two sets of equations, $COR(B,C|A=1)$ and $COR(B,C|A=2)$ can be determined, so giving the canonical parameters on the ascending class is equivalent to giving the conditional odds ratios on the ascending class. This is not only true in the example considered.

Theorem 10.6. *Let $\mathscr{A}$ be an ascending class. The following sets of parameters may be calculated from each other: the canonical parameters*

$$\theta_A = \frac{1}{|2^V|}\boldsymbol{t}'_A \log \boldsymbol{p} \textit{ for all } A \in \mathscr{A}$$

and the conditional odds ratios

$$COR(A|V \setminus A = a') \textit{ for all categories of } V \setminus A \textit{ and for all } A \in \mathscr{A}.$$

Proof. According to the first formula of Theorem 10.5, the canonical parameters may be calculated from the conditional odds ratios. To see that the conditional odds ratios may also be obtained from the canonical parameters, combine the two formulas in Theorem 10.5 to obtain that

$$(\mathbf{t}_A, \log \mathbf{p}) = \Sigma_{a'}(-1)^{|A|} \log COR(A|V \setminus a = a').$$

For simplicity of presentation, let $V = \{V_1, \dots, V_4\}$, and let A contain the variables V_1, V_2. Then, the previous formula becomes

$$(\mathbf{t}_{\{V_1,V_2\}}, \log \mathbf{p}) = \Sigma_{(v_3,v_4)}(-1)^2 \log COR(V_1V_2|(V_3,V_4) = (v_3,v_4)).$$

$$= \log COR(V_1V_2|(V_3,V_4) = (1,1)) + \log COR(V_1V_2|(V_3,V_4) = (1,2)$$

$$+ \log COR(V_1V_2|(V_3,V_4) = (2,1) + \log COR(V_1V_2|(V_3,V_4) = (2,2).$$

Similarly,

$$(\mathbf{t}_{\{V_1,V_2,V_3\}}, \log \mathbf{p}) = \Sigma_{(v_4)}(-1)^3 logCOR(V_1V_2V_3|V_4 = v_4)$$

$$= -\log COR(V_1V_2V_3|V_4 = 1) - \log COR(V_1V_2V_3|V_4 = 2)$$

$$= \log COR(V_1V_2|(V_3,V_4) = (2,1)) - \log COR(V_1V_2|(V_3,V_4) = (1,1))$$

$$+ \log COR(V_1V_2|(V_3,V_4) = (2,2)) - \log COR(V_1V_2|(V_3,V_4) = (1,2)),$$

using the definition of the higher-order conditional odds ratios. One also has

$$(\mathbf{t}_{\{V_1,V_2,V_4\}}, \log \mathbf{p}) = \Sigma_{(v_3)}(-1)^3 \log COR(V_1V_2V_4|V_3 = v_3)$$

$$= -\log COR(V_1V_2V_4|V_3 = 1) - \log COR(V_1V_2V_4|V_4 = 2)$$

$$= \log COR(V_1V_2|(V_3,V_4) = (1,2)) - \log COR(V_1V_2|(V_3,V_4) = (1,1))$$

$$+ \log COR(V_1V_2|(V_3,V_4) = (2,2)) - \log COR(V_1V_2|(V_3,V_4) = (2,1)).$$

Finally,

$$(\mathbf{t}_{\{V_1,V_2,V_3,V_4\}}, \log \mathbf{p}) = \log OR(V_1V_2V_3V_4)$$

$$= \log COR(V_1V_2|(V_3,V_4) = (1,1)) - \log COR(V_1V_2|(V_3,V_4) = (1,2))$$

$$+ \log COR(V_1V_2|(V_3,V_4) = (2,2)) - \log COR(V_1V_2|(V_3,V_4) = (2,1)).$$

These equations mean that the $\log COR(V_1V_2|(V_3,V_4) = (v_3,v_4))$ quantities, in the lexicographic order, are multiplied by the following vectors to produce the canonical parameters on the ascending class:

$$(1,1,1,1)$$

$$(-1,-1,1,1)$$

$$(-1,1,-1,1)$$

$$(1,-1,-1,1)$$

As these vectors are orthogonal, the system of linear equations, with the conditional odds ratios as unknowns, has a unique solution. □

The results about parameterizations obtained in this section will be used next to define useful statistical models.

10.2 Generalizations of Independence

The various parameterizations discussed in the previous section can be used to define models by fixing some of the parameters in such a way that this defines a meaningful restriction on the probability distributions in the model and the remaining parameters constitute a parameterization of the distributions in the model.

For a 2×2 table, the following cut

$$\begin{array}{c} AB \\ a\ b \\ \emptyset \end{array}$$

defines the mixed parameterization consisting of the odds ratio and the two one-way marginal distributions. If in this parameterization the odds ratio is set to 1, one obtains the 2×2 independence model. Because the odds ratio and the one-way marginal distributions are variation independent (see Sect. 6.1.2), the parameters are not fixed, that is, the one-way marginal distributions parameterize all independent distributions. Log-linear models extend this structure for general tables. The form of independence, which leads to the generalization is

$$p(i,j) = p(i,+)p(+,j), \text{ for all } i,j. \tag{10.14}$$

In this form, independence appears to mean that the joint probability is the product of the marginal or one-way probabilities. However, independence may be seen from a more general perspective.

Theorem 10.7. *For a two-way distribution, (10.14) holds if and only if there exist numbers α_i and β_j, such that*

$$p(i,j) = \alpha_i \beta_j, \textit{ for all } i,j. \tag{10.15}$$

The theorem says that independence is the same as assuming that the joint probability is the product of two effects, one depending on the rows and one on the columns only.

Proof. When (10.14) holds for the distribution, (10.15) holds, too. When (10.15) holds, let A be the sum of the α_i parameters and set $\alpha'_i = \alpha_i / A$ and similarly for the β_j. Then

$$p(i,j) = AB\alpha'_i\beta'_j \tag{10.16}$$

Summation in j, for a fixed i gives that

$$p(i,+) = AB\alpha'_i,$$

and further summation in i yields that $1 = AB$, thus $\alpha'_i = p(i,+)$. Similarly, $\beta'_j = p(+,j)$, and (10.16) is the same as (10.14). □

There is a variant of this parameterization, which will be particularly useful later on. In this,

$$p(i,j) = c\alpha_i\beta_j, \text{ for all } i,j \tag{10.17}$$

In fact, there are many choices of the parameters with the above property, and they may be made unique by assuming that, for an $I \times J$ table,

$$\prod_{i=1}^{I} \alpha_i = 1 \text{ and } \prod_{j=1}^{J} \beta_j = 1 \tag{10.18}$$

For example, the following choice provides such parameters

$$c = \left(\prod_{i=1}^{I} \prod_{j=1}^{J} p(i,j) \right)^{\frac{1}{IJ}}, \tag{10.19}$$

$$\alpha_i = p(i,+)/\sqrt{c}, \text{ and } \beta_j = p(+,j)/\sqrt{c}.$$

For three-way tables, the cut in (10.3) implies the parameterization consisting of the one-way marginal distributions and the conditional odds ratios of any two variables, given the categories of the third one. When all these conditional odds ratios are set to 1, one obtains a model characterized by the following properties.

Theorem 10.8. *The following three properties are identical:*

$$COR(A,B|C=k)=1,\ COR(A,C|B=j)=1,\ COR(B,C|A=i)=1, \textit{ for all } i,j,k$$

$$p(i,j,k) = p(i,+,+)p(+,j,+)p(+,+,k), \textit{ for all } i,j,k$$

$$p(i,j,k) = c\alpha_i\beta_j\gamma_k \textit{ for all } i,j,k$$

Proof. The second statement implies the first and the third ones. The proof that the third one implies the second one is like the proof in Theorem 10.7.

The proof that the first equality implies the third one is presented now for binary variables. In the general case, this argument has to be applied to each local or spanning cell odds ratio.

First define c as $p(2,2,2)$. The fact that the conditional odds ratios are equal to 1 implies the following:

$$\frac{p(1,1,1)}{p(1,1,2)} = \frac{p(1,2,1)}{p(1,2,2)} = \frac{p(2,2,1)}{p(2,2,2)} = \frac{p(2,1,1)}{p(2,1,2)}.$$

Here, the first equality is implied by the third COR being equal to 1, the second is implied by the second one, and the third is implied by the third one. The meaning of these equations is that when C changes from 2 to 1, the proportional change in the cell probabilities is the same, irrespective of the categories of the other variables. Define γ_1 as the common value above, and let $\gamma_2 = 1$.

Similarly, use the common value of

$$\frac{p(i,1,k)}{p(i,2,k)}$$

to define β_1, and let $\beta_2 = 1$. Lastly, the common value of

$$\frac{p(1,j,k)}{p(2,j,k)}$$

defines α_1, and let $\alpha_2 = 1$. Then,

$$p(i,j,k) = c\alpha_i\beta_j\gamma_k.$$

Indeed, when $(i,j,k)=(2,2,2)$, this is true; when i changes from 2 to 1, the factor of proportionality is α_1; and if j changes from 2 to 1, the factor of proportionality is β_1, irrespective, whether i was changed or not, and similarly for k and γ_1.

□

The essential point in the proof above was that the effect of changing one index did not depend on whether the other indices were changed. The structure defined in Theorem 10.8 is called joint or mutual independence of the three variables. The theorem says that in this case, there is a multiplicative representation of the cell probabilities, in which the variables appear separately, not jointly. It is also said that the effects of the variables on the cell probabilities are separable.

Obviously, the parameters c, α_i, β_j, and γ_k parameterize the distribution. This parameterization is called the effect parameterization. A more balanced parameterization is obtained, by choosing c, again, as the geometric mean of the cell probabilities, and replacing α_i with $p(i,j,k)/m$, where m is the geometric mean of $p(1,j,k)$ and $p(2,j,k)$, and similarly for β and γ.

The cut in (10.5) leads to the mixed parameterization consisting of the $A\times B$ two-way and the C one-way marginal distributions and the $COR(A,C|B=j)$ and the $COR(B,C|A=i)$ conditional odds ratios.

Theorem 10.9. *The following statements are identical:*

$$COR(A,C|B=j)=1,\ COR(B,C|A=i)=1,\ \textit{for all } i,j$$

$$p(i,j,k)=p(i,j,+)p(+,+,k),\ \textit{for all } i,j,k$$

$$p(i,j,k)=c\alpha_i\beta_j\gamma_k\delta_{ij}\ \textit{for all } i,j,k$$

Proof. The proof can be told in two different ways. One is to consider the combined AB variable, which has the combinations of the categories of A and of B as its categories, and note that the model considered here is the two-way independence for the variables AB and C.

The other, somewhat more instructive version of the proof, is along the lines of the proof of Theorem 10.8 and starts by establishing that the conditional odds ratios being equal to 1 implies that

$$\frac{p(1,1,1)}{p(1,1,2)}=\frac{p(1,2,1)}{p(1,2,2)}=\frac{p(2,2,1)}{p(2,2,2)}=\frac{p(2,1,1)}{p(2,1,2)},$$

where the first equality follows from the second conditional odds ratio being equal to 1, the second follows from the first one, and the last follows from the second one. Use this to define the effect of C as γ_1, the common value, and $\gamma_2=1$.

Then define c as $p(2,2,2)$, and let

$$\alpha_i=1,\ \text{for } i=1,2,\ \beta_j=1,\ \text{for } j=1,2$$

$$\delta_{ij}=p(i,j,2)/c,\ \text{for } i,j=1,2.$$

These are choices with which the claims of the theorem hold.

In the non-binary case, the argument above applies each local or spanning cell subtable. □

A balanced reparameterization, like the one created after Theorem 10.8, is also possible. The details will be discussed in the next section.

The property formulated in Theorem 10.9 is called multiple independence.

For the cut in (10.7), the parameters in the mixed parameterization were given in (10.8). One has the following result about setting the relevant conditional odds ratios to 1.

Theorem 10.10. *The following statements are identical:*

$$COR(B,C|A=i)=1, \textit{ for all } i$$

$$p(i,j,k)=\frac{p(i,j,+)p(i,+,k)}{p(i,+,+)}, \textit{ for all } i,j,k$$

$$p(i,j,k)=c\alpha_i\beta_j\gamma_k\delta_{ij}\varepsilon_{ik} \textit{ for all } i,j,k$$

Proof. Both the second and the third statements imply the first one. The second statement implies the third one. Proposition 6.6 implies that the first and the second statements are equivalent. □

The models defined in the three last theorems may be seen as generalizations of independence for three-way tables. Multiple independence has three variants and so does conditional independence. These variants do not need to be distinguished as mathematical constructs but need to be seen as different models in a data analytic situation. While the mathematical structures may be seen as identical, it does make a big difference, if gender and income are conditionally independent, given education or income and education are conditionally independent, given gender.

All of these models were obtained by choosing a cut, then taking the implied mixed parameterization, and then setting the conditional odds ratios equal to 1 on the ascending class. These are simple log-linear models, and the general definition is going to be given in the next section. In an ascending class of subsets, it is enough to impose that the conditional odds ratios are equal to 1, on the minimal subsets, and by the definition of higher-order conditional odds ratios, this will hold true for the larger subsets, too. The marginal distributions in the descending class parameterize the distributions in the model, and it is sufficient to give the marginal distributions on the maximal elements of the descending class.

For the cut in (10.9), the mixed parameterization was given in (10.10). For this case, a somewhat weaker result can be given, namely, there is no representation of the model defined by setting the second-order odds ratio to 1 in terms of the marginal probabilities.

Theorem 10.11. *The following statements are identical:*

$$OR(A,B,C)=1$$

and there exist parameters α_i, β_j, γ_k, δ_{ij}, ε_{ik}, and ϕ_{jk} such that

$$p(i,j,k) = c\alpha_i\beta_j\gamma_k\delta_{ij}\varepsilon_{ik}\phi_{jk}\, for\ all\ i,j,k$$

Proof. The theorem is a consequence of the convergence of the Iterative Proportional Fitting Procedure, to be proved in Sect. 10.2. Also, the result is a special case of Theorem 10.15. □

This model is called no second-order interaction, and the joint probabilities cannot be written as a function of marginal probabilities. This is also a log-linear model.

10.3 Log-Linear Models and Parameters

The models discussed in the previous section are all log-linear models. A general definition is obtained by choosing a cut of the class of all subsets of the variables into an ascending and a complement descending class. Then, consider the mixed parameterization induced by the cut, consisting of all conditional odds ratios on the ascending class (with all possible categories of the variables in the complement, as conditions) and the marginal distributions on the descending class. In this mixed parameterization, it is sufficient to consider the conditional odds ratios on the minimal elements of the ascending class and the marginal distributions on the maximal elements of the descending class. A log-linear model is obtained by assuming that all the conditional odds ratios on the ascending class are equal to 1. The marginal distributions on the complement descending class parameterize the distributions in the model. Ultimately, the model is determined by the cut, which need to be selected according to the relevant properties of the substantive problem at hand. Denote the log-linear model belonging to the ascending class $\mathscr{A}$ as $LL(\mathscr{A})$, as a subset of $\mathscr{P}$. That is,

$$\mathbf{p} \in LL(\mathscr{A}) \text{ if and only if } COR_{\mathbf{p}}(A|A' = a') = 1, \text{ for all categories } a' \text{ of } A' = V \setminus A.$$

The meaning of $COR_{\mathbf{p}}(A|A' = a')$ is the conditional odds ratio of the variables in A, given $A' = a'$ in the distribution $\mathbf{p}$. This may be a single number or a collection of local or spanning cell odds ratios, and in that case the equality should hold for all of them.

Log-linear models are very useful tools in analyzing categorical data. They formulate a kind of simplicity, often found in real data: the lack of higher-order interactions, when interaction is measured by the odds ratio. Important examples for three-way tables were shown in the previous section. Further uses and facts relevant for the interpretation of log-linear models will be discussed in the present and the next chapters.

To generalize the multiplicative representations given in the last parts of Theorems 10.8 to 10.11, introduce a general notation for multiplicative parameters. Let

$$\beta_w^W$$

denote a function defined on the W-marginal of the contingency table and, at the same time, its value in the w marginal cell. Further, let

$$(v)_W$$

denote the projection of a cell v to the W marginal, that is, those indices from v, which belong to variables in W. For example, if $V = \{A,B,C,D\}$, $W = \{B,D\}$, then $(i,j,k,l,)_W = (j,l)$. Then, it is easy to define multiplicative parameters, such that

$$p(v) = \prod_{W \subseteq V} \beta^W_{(v)_W}. \tag{10.20}$$

For example, the choice $\beta_v^V = p(v)$ and all other parameters equal to 1 will work. This is, of course, not a very useful reparameterization of the distribution. One way of looking at the meaning of a log-linear model is that if it holds for a distribution, then parameters belonging to lower dimensional subsets are sufficient to specify it. In fact, multiplicative parameters defined on subsets in the descending class are sufficient to reproduce the distribution, that is, they parameterize it. This generalizes the results of the previous section.

Theorem 10.12. *The following two statements are equivalent:*

$$\boldsymbol{p} \in LL(\mathscr{A})$$

and there exist such multiplicative parameters $\beta^W_{(v)_W}$ *that*

$$p(v) = \prod_{W \in \mathscr{D}} \beta^W_{(v)_W}, \textit{ for all cells } v,$$

where $\mathscr{D}$ *is the complement descending class to* $\mathscr{A}$.

Proof. The second statement implies the first one by the definition of the log-linear model and the odds ratio.

The converse is a consequence of the convergence of the Iterative Proportional Fitting Procedure, to be proved later, but there is also a constructive proof implied by Theorem 10.15, to be developed now. □

The definition of the log-linear model has been given in multiplicative form so far. The name of these models comes from the fact that after taking logarithms, which is possible because of the assumed positivity, the models become additive. The choice of the base of the logarithm is irrelevant. More precisely, for every

$W \subseteq V$, let c_W be the number of cells in the W-marginal table, and for every marginal cell v_W, define the following log-linear parameters via recursion.[2]

$$\lambda^{\emptyset} = \frac{1}{c_V} \sum_v \log p(v),$$

$$\lambda_{w^*}^{W^*} = \left(\frac{1}{c_{W^*}} \sum_{v:(v)_w = w^*} \log p(v) \right) - \sum_{Z \subsetneq W^*} \lambda_{(w^*)_Z}^{Z}$$

In the formula above, W^* denotes a specific subset of the variables, and w^* denotes one specific joint index of the variables in W^*. The term between the parentheses is the average log probability of those cells, which project into w^*, and the values of all lower-order terms belonging to this marginal cell are subtracted.

By the definition,

$$\log p(v) = \sum_{W \subseteq V} \lambda_{(v)_W}^{W} \tag{10.21}$$

for every probability distribution, and this is called the log-linear representation of the probability distribution $\mathbf{p}$.

For example, for a probability distribution on an $I \times J \times K$ contingency table, with variables A, B, and C, the log-linear parameters are as follows:

$$\lambda^{\emptyset} = \frac{1}{IJK} \sum_{(i,j,k)} \log p(i,j,k),$$

$$\lambda_i^A = \frac{1}{JK} \sum_{(j,k)} \log p(i,j,k) - \lambda^{\emptyset},$$

$$\lambda_j^B = \frac{1}{IK} \sum_{(i,k)} \log p(i,j,k) - \lambda^{\emptyset},$$

$$\lambda_k^C = \frac{1}{IJ} \sum_{(i,j)} \log p(i,j,k) - \lambda^{\emptyset},$$

$$\lambda_{(i,j)}^{AB} = \frac{1}{K} \sum_k \log p(i,j,k) - \lambda^{\emptyset} - \lambda_i^A - \lambda_j^B,$$

$$\lambda_{(i,k)}^{AC} = \frac{1}{J} \sum_j \log p(i,j,k) - \lambda^{\emptyset} - \lambda_i^A - \lambda_k^C,$$

$$\lambda_{(j,k)}^{BC} = \frac{1}{I} \sum_i \log p(i,j,k) - \lambda^{\emptyset} - \lambda_j^B - \lambda_k^C,$$

$$\lambda_{(i,j,k)}^{ABC} = \log p(i,j,k) - \lambda^{\emptyset} - \lambda_i^A - \lambda_j^B - \lambda_k^C - \lambda_{(i,j)}^{AB} - \lambda_{(i,k)}^{AC} - \lambda_{(j,k)}^{BC}.$$

[2] It is traditional to denote these parameters with λ. They have nothing to do with the parameter of a Poisson distribution.

Another example of the log-linear parameters was given at the beginning of Sect. 7.2.

These log-linear parameters have many interesting properties. They are a parameterization of the joint distribution, and they are balanced in the following sense:

Proposition 10.3. *For every $\emptyset \neq W \subseteq V$ and $A \in W$ with categories i,*

$$\sum_i \lambda_w^W = 0$$

Proof. First, let $W = \{A\}$ with index i, and let J be a joint index of the variables in $V \setminus W$, with ranges $i = 1, \ldots, I$, $j = 1, \ldots, J$. Then,

$$\frac{1}{I}\sum_i \frac{1}{J}\sum_j \log p(i,j) = \lambda^{\emptyset},$$

and by the definition of the log-linear parameters,

$$\sum_i \lambda_i^A = \sum_i \left(\frac{1}{J}\sum_j \log p(i,j) - \lambda^{\emptyset} \right) = I\lambda^{\emptyset} - I\lambda^{\emptyset} = 0.$$

For an induction proof on $|W|$, assume the claim is true for all subsets with fewer than $|W|$ variables. Let $A \in W$, with indices i; let B with indices j the combined variable from all variables in W, except for A; and let C, with indices k, the combined variable from all variables in $V \setminus W$. Then, as seen in the example above,

$$\sum_i \lambda_{ij}^{AB} = \sum_i \left(\frac{1}{K}\sum_k \log p(i,j,k) - \lambda^{\emptyset} - \lambda_i^A - \lambda_j^B \right)$$

$$= I\lambda_j^B + I\lambda^{\emptyset} - I\lambda^{\emptyset} + 0 - I\lambda_j^B = 0,$$

where the first two terms come from the definition of λ_j^B and the fourth term (the zero) comes from the induction assumption. □

To obtain a non-recursive formula for the log-linear parameters in the case of the $I \times J \times K$ table, let (i^*, j^*, k^*) be a fixed cell. Then,

$$\lambda_{i^*}^A = \frac{1}{JK}\sum_{(j,k)} \log p(i^*, j, k) - \lambda^{\emptyset} =$$

$$\frac{1}{JK}\sum_{(j,k)} \log p(i^*, j, k) - \frac{1}{IJK}\sum_{(i,j,k)} \log p(i,j,k) =$$

$$\sum_{(i,j,k)} \frac{\delta(i,i^*)I - 1}{IJK} \log p(i,j,k),$$

where $\delta(i,i^*) = 1$, if $i = i^*$ and is zero otherwise. Similarly,

$$\lambda^{AB}_{(i^*,j^*)} = \frac{1}{K}\sum_k \log p(i^*,j^*,k) - \lambda^{\emptyset} - \lambda^A_i - \lambda^B_j =$$

$$\frac{1}{K}\sum_k \log p(i^*,j^*,k) - \frac{1}{IJK}\sum_{(i,j,k)} \log p(i,j,k) -$$

$$\sum_{(i,j,k)} \frac{\delta(i,i^*)I-1}{IJK} \log p(i,j,k) - \sum_{(i,j,k)} \frac{\delta(j,j^*)J-1}{IJK} \log p(i,j,k) =$$

$$\sum_{(i,j,k)} \frac{(\delta((i,j)(i^*,j^*))IJ) - 1 - (\delta(i,i^*)I-1) - (\delta(j,j^*)J-1)}{IJK} \log p(i,j,k) =$$

$$\sum_{(i,j,k)} \frac{\delta((i,j)(i^*,j^*))IJ - \delta(i,i^*)I - \delta(j,j^*)J + 1}{IJK} \log p(i,j,k).$$

For the three-way (or second-order) interaction term, one obtains that

$$\lambda^{ABC}_{(i^*,j^*,k^*)} = \log p(i^*,j^*,k^*) - \lambda^{\emptyset} - \lambda^A_{i^*} - \lambda^B_{j^*} - \lambda^C_{k^*} - \lambda^{AB}_{(i^*,j^*)} - \lambda^{AC}_{(i^*,k^*)} - \lambda^{BC}_{(j^*,k^*)}.$$

Here, the numerator of the coefficient of $\log p(i,j,k)$ is

$$\delta((i,j,k),(i^*,j^*,k^*))IJK - 1 - (\delta(i,i^*)I-1) - (\delta(j,j^*)J-1) - (\delta(k,k^*)K-1) -$$

$$(\delta((i,j)(i^*,j^*))IJ - \delta(i,i^*)I - \delta(j,j^*)J + 1) -$$

$$(\delta((i,k)(i^*,k^*))IK - \delta(i,i^*)I - \delta(k,k^*)J + 1) -$$

$$(\delta((j,k)(j^*,k^*))JK - \delta(j,j^*)J - \delta(k,k^*)K + 1) =$$

$$\delta((i,j,k),(i^*,j^*,k^*))IJK - \delta((i,j)(i^*,j^*))IJ - \delta((i,k)(i^*,k^*))IK -$$

$$\delta((j,k)(j^*,k^*))JK + \delta(i,i^*)I + \delta(j,j^*)J + \delta(k,k^*)K - 1.$$

The following theorem gives a general variant of the formulas above.

Theorem 10.13. *For every $\mathbf{p} \in \mathscr{P}$ and $W \subseteq V$,*

$$\lambda^W_w = \frac{1}{c_W}\sum_v \sum_{U \subseteq W} (-1)^{|V|-|U|} \delta((v)_U,(w)_U) c_U \log p(v),$$

where $(v)_{\emptyset} = (u)_{\emptyset}$, for all indices u and w, and c_U is the number of joint categories of the variables in U, so that $c_{\emptyset} = 1$. □

This theorem is a special case of the so called Möbius inversion, see, e.g., [14]. A bit more insight into the structure of the log-linear parameters is given by the following result.

Theorem 10.14. *For $W \subseteq V$ and indices w of W and v of V, let $T \subseteq W$ be the largest such set, that $(v)_T = (w)_T$. Then,*

$$\lambda_w^W = \sum_v \frac{e_{w,v}}{c_W} \log p(v),$$

where

$$e_{w,v} = \sum_{U \subseteq T} (-1)^{|V|-|U|} c_U.$$

Proof. First note that T exists, because if v and w are identical on two subsets of the variables, then they are also identical on their union.

The subsets U of V, for which $\delta((v)_U, (w)_U) = 1$ are exactly the subsets of T. The rest follows from Theorem 10.13.

□

To see an example, let $A \times B \times C$ be a $2 \times 3 \times 2$ table. Then the coefficients of the log probabilities in some of the log-linear parameters are given in Table 10.1.

Table 10.1 Coefficients of log probabilities in selected log-linear parameters in a $2 \times 3 \times 2$ table (coefficients multiplied by 12)

	λ_3^B	$\lambda_{2,3}^{AB}$	$\lambda_{1,1,2}^{ABC}$	$\lambda_{1,3,2}^{ABC}$
$\log p(1,1,1)$	-1	-1	-2	1
$\log p(1,1,2)$	-1	-1	2	-1
$\log p(1,2,1)$	-1	-1	1	1
$\log p(1,2,2)$	-1	-1	-1	-1
$\log p(1,3,1)$	2	2	1	-2
$\log p(1,3,2)$	2	2	-1	2
$\log p(2,1,1)$	-1	1	2	-1
$\log p(2,1,2)$	-1	1	-2	1
$\log p(2,2,1)$	-1	1	-1	-1
$\log p(2,2,2)$	-1	1	1	1
$\log p(2,3,1)$	2	-2	-1	2
$\log p(2,3,2)$	2	-2	1	-2

For binary variables, using the notation of Theorem 10.14, one has the following result.

Proposition 10.4. *Let all the variables be binary. Then,*

$$e_{w,v} = (-1)^{|V|+|T|},$$

where $T \subseteq W$ is the largest such set that $(v)_T = (w)_T$.

Proof. For binary variables,

$$\sum_{U \subseteq T} (-1)^{-|U|} C_U = \sum_{U \subseteq T} (-1)^{|U|} c_U$$

$$= 1 - \binom{|T|}{1} 2 + \binom{|T|}{2} 4 \ldots + (-1)^{|T|-1} \binom{|T|}{|T|-1} 2^{|T|-1} + (-1)^{|T|} \binom{|T|}{|T|} 2^{|T|}$$

$$= (1-2)^{|T|} = (-1)^{|T|},$$

therefore

$$e_{w,v} = (-1)^{|V|}(-1)^{|T|}.$$

□

For example, in an $A \times B \times C \times D$ binary table, let us consider the log-linear parameter λ_{121}^{ABC}. Then, for every index l, the coefficient of $\log p(121l) = -1$. If out of the first three indices exactly one is changed, the sign reverses; if two are changed, the sign remains the same; etc. Therefore the coefficient is -1 for the logarithms of $p(121l)$, $p(211l)$, $p(222l)$, and $p(112l)$ and is 1 for the logarithms of $p(221l)$, $p(111l)$, $p(122l)$, and $p(212l)$. Therefore,

$$\lambda_{121}^{ABC} = \frac{1}{8}\log COR(ABC|D=1) + \frac{1}{8}\log COR(ABC|D=2) =$$

$$\frac{1}{8}\log COR(ABC|D=l)^2.$$

To illustrate further, how log-linear parameters and conditional odds ratios are related, consider a 2×3 table. For this case, Table 10.2 gives the log-linear interaction term as a function of the logarithms of the cell probabilities, obtained by the application of Theorem 10.14, and also the logarithms of the spanning cell odds ratios, spanned by the cells $(2,2)$ and $(2,3)$.

Table 10.2 Log-linear interaction parameters multiplied by 6 and odds ratios in a 2×3 table

	$\lambda_{1,1}^{AB}$	$\lambda_{1,2}^{AB}$	$\lambda_{1,3}^{AB}$	$\lambda_{2,1}^{AB}$	$\lambda_{2,2}^{AB}$	$\lambda_{2,3}^{AB}$	$\log OR_{(2,2)}(A,B)$	$\log OR_{(2,3)}(A,B)$
$\log p(1,1)$	2	−1	−1	−2	1	1	1	1
$\log p(1,2)$	−1	2	−1	1	−2	1	−1	
$\log p(1,3)$	−1	−1	2	1	1	−2		−1
$\log p(2,1)$	−2	1	1	2	−1	−1	−1	−1
$\log p(2,2)$	1	−2	1	−1	2	−1	1	
$\log p(2,3)$	1	1	−2	−1	−1	2		1

The coefficients in Table 10.2 also illustrate Proposition 10.3. Because of the property that the log-linear parameters summed in any of their indices give zero, in case the first 2 log-linear parameters in Table 10.2 are zero, then the entire interaction parameter is zero. It is easy to see, that the first interaction parameter may be obtained as the sum of the 2 log odds ratios, and the second interaction term may be obtained, as the second log odds ratio minus twice the first one. Consequently, if the log odds ratios are equal to zero, so are the log-linear interaction terms.

Conversely, the first log odds ratio may be obtained as one-third of the difference between the first and second log-linear parameters, and the second log odds ratio is two-third of the first log-linear parameter plus one-third of the second. This implies that the log odds ratios are zero if and only if the log-linear interaction parameters are zero. An identical result for 2×2 tables was given in Sect. 7.2 in the context of exponential families.

The log-linear parameters may be used to obtain multiplicative parameters with the properties discussed in the present and the previous sections. The multiplicative parameters may be defined as

$$\beta_w^W = exp\lambda_w^W,$$

and this is the definition which will be used from now on. Then, Proposition 10.3 implies that

$$\prod_w \beta_w^W = 1 \text{ for all } W \subseteq V, \tag{10.22}$$

which generalizes (10.18). Parallel to the log-linear representation of the logarithms of the cell probabilities given in (10.21), one has the following multiplicative representation of any probability distribution in $\mathscr{P}$:

$$p(v) = \prod_{W \subseteq V} \beta_{(v)_W}^W.$$

With the current choice of the β parameters, the following result has a constructive proof, instead of the existence given in Theorem 10.12.

Theorem 10.15. *For every $\mathbf{p} \in \mathscr{P}$ and complement descending class of subsets $\mathscr{D}$ and ascending class of subsets $\mathscr{A}$, $\mathbf{p} \in LL(\mathscr{A})$ if and only if*

$$p(v) = \prod_{W \in \mathscr{D}} \beta_{(v)_W}^W. \tag{10.23}$$

Further, $\mathbf{p} \in LL(\mathscr{A})$ if and only if

$$\lambda_a^A = 0 \textit{ for all } A \in \mathscr{A} \textit{ and for all indices } a \textit{ of } A. \tag{10.24}$$

Proof. Condition (10.24) implies (10.23). First, it will be proved that (10.23) implies that the distribution is in the log-linear model; then it will be proved that if the distribution is in the log-linear model, then (10.24) holds.

Assume first that (10.23) holds and consider the conditional odds ratios $COR(A|A' = a')$, which have to be equal to 1, to prove that $\mathbf{p}$ is in the log-linear model. For every binary conditional subtable of the $A' = a'$ conditional table, and for the odds ratio computed in it, it is sufficient to see that, for every subset of variables $W \in \mathscr{D}$ and every joint category w of them, the number of times the β_w^W parameter appears in the numerator and in the denominator of the conditional odds ratio is the same.

Without loss of generality, let the binary subtable of the $A' = a'$ conditional table be defined by the indices 1 and 2. Then, the index of each cell consists of a sequence of 1s and 2s of length $|A|$ (to be called the A part of the index) and of a' (to be called the A' part of the index). Which cell probabilities go into the numerator, and which ones into the denominator of the conditional odds ratio, depends on the parity of the number of 1s in the sequence. Thus, one has to see that the number of times β_w^W appears in the multiplicative representation of cells with even number of 1s and with an odd number of 1s is the same. Whether or not β_w^W appears in the multiplicative representation of a cell v depends on whether or not $(v)_W = w$.

Because W is in the descending class, and A is in the ascending class, A cannot be a subset of W. Thus, $A \setminus W$ is not empty, and thus the A part of every index contains positions, which are not in W. Consider those indices v only, for which $(v)_W = w$ and, therefore, contain β_w^W. The positions in W may be in the A part and may be in the A' part, but there are positions in the A part, which do not belong to W. These positions contain all possible 1–2 sequences of length $|A \setminus W|$, and half of them goes in the denominator and half in the numerator, ultimately canceling out, which concludes the first part of the proof.

Conversely, assume now that $\mathbf{p}$ is in the log-linear model and thus the conditional odds ratios on the ascending class $\mathscr{A}$ are equal to 1. Consider a multiplicative parameter β_w^W, with $W \in \mathscr{A}$. The corresponding log-linear parameter λ_w^W is a linear combination of logarithms of cell probabilities, with coefficients given in Theorem 10.14. The coefficients $e_{w,v}$ do not depend on the part of v, which is outside of w. This is also seen in Table 10.1. Thus, using w^* to denote a specific joint index for the variables in W, and w'^* to denote a specific joint index for the variables in W',

$$\lambda_{w^*}^W = c_{W'} \sum_{v:(v)_{W'}=w'^*} \frac{e_{w^*,v}}{c_W} \log p(v) = \frac{c_{W'}}{c_W} \sum_w e_{w,(w,w'^*)} \log p(w, w'^*).$$

The formula above means that the log-linear parameter is constant times its part, which is computed from those cells that have the same indices on the variables in W'. This is shown for binary variables in the example after Proposition 10.4. Write this as

$$\lambda_{w^*}^W = \frac{c_{W'}}{c_W} \lambda_w^W (w'^*).$$

As seen in Proposition 10.3, λ_w^W, when summed in the index of any variable in W, gives zero. Therefore, the c_V dimensional vector of coefficients of the log probabilities in λ_w^W is in a

$$\prod_{X \in W} (c_X - 1)$$

dimensional subspace. As $\lambda_w^W (w'^*)$ is subject to the same sums being zero, it is, if seen as a vector of coefficients, also in a subspace of the latter dimension, although the size of the vector is c_W.

Consider now the logarithms of the conditional odds ratios of W, given $W' = w'^*$, as the vectors of the coefficients of the logarithms of the odds ratios. These vectors are of the length c_W, and by definition, they also have the property of yielding zero, when summed in any of the indices of the variables in W. Thus, these vectors are in the same subspace, as the $\lambda_w^W (w'^*)$ vectors. Also by the definition of the odds ratio, these vectors are linearly independent.

Whether the conditional odds ratios are based on spanning cells or are local odds ratios, their number is $\prod_{X \in W} (c_X - 1)$. Therefore the vectors of the coefficients in the log odds ratios span the space where the coefficients of the $\lambda_w^W (w'^*)$ vectors are. This was illustrated in the example which yielded Table 10.2.

Thus, the log conditional odds ratios are all zero, and then every vector in the subspace, including $\lambda_w^W(w'^*)$, is zero. This implies that $\lambda_w^W = 0$. □

In this section, log-linear models were introduced for strictly positive discrete probability distributions as the assumption that the conditional odds ratios are all equal to 1 on an ascending class of subsets of the variables. This definition is equivalent to the higher-order terms (log-linear or multiplicative parameters on the ascending class) disappearing from a log-linear or multiplicative representation.

As illustrated in the previous section, these models generalize independence and have an interpretation in terms of a simple interaction structure. Another approach to define and interpret log-linear models starts with the log-linear representation in (10.21), points out that the definition of log-linear parameters is identical to the definition of effects in the analysis of variance (as if the logarithms of the cell probabilities were frequencies), and derives log-linear models from the assumption of no higher-order interaction. Further interpretations of log-linear models will be given in the next chapter of the book.

As discussed above, a log-linear model is essentially determined by the cut, upon which the mixed parameterization is based. This cut can be equally specified through the ascending or the complement descending class. It is usual to apply a shorthand notation for log-linear models, which specifies the maximal elements of the descending class. For example, the model of independence in a two-way table is written as A and B. In a three-way table, A, B, and C are the mutual independence, and AB and BC are the conditional independence of A and C, given B. These maximal elements of the descending class are also called maximal interactions, and together they are called the generating class of the model.

10.4 Things to Do

1. Prove the second claim of Theorem 10.1.
2. What is the parameter associated with the subset containing all variables in a mixed parameterization, if it is in the ascending class, and what is the parameter, if it is in the descending class?
3. What are the ranges of the parameters in (10.1)?
4. Prove that the parameters in (10.1) are a parameterization of the 2×2 distribution. Use Theorem 6.3.
5. Are the parameters in (10.1) variation independent?
6. Show for all parameters derived from the cuts for the 2×2 table that they are parameterizations.
7. Why are the parameters in (10.4) redundant when $i, j, k = 1, 2$?
8. Find three pairs of parameters in (10.10), which are not variation independent from each other.
9. Prove the claim made for the rows of the design matrix in Proposition 10.1.
10. Formulate a general proof of Theorem 10.6.

11. Show that in Theorem 10.8, the third statement implies the second one.
12. Define the parameters in (10.17) from those in (10.15).
13. Choose a $2 \times 2 \times 2$ probability distribution, and calculate the parameters defined in Theorem 10.8.
14. For the probability distribution in the item above, calculate the parameters defined after Theorem 10.8.
15. Explain the meanings of the parameters obtained in the two previous items.
16. Work out the details of the proof of Theorem 10.8 for a $3 \times 3 \times 3$ table. Local or spanning cell odds ratios are more useful for the proof?
17. Define a balanced variant of δ_{ij} in Theorem 10.9.
18. Work out the details of the proof of Theorem 10.9 for a $4 \times 2 \times 2$ and for a $2 \times 2 \times 4$ table. Why are these different?
19. Show that if the conditional odds ratios are all equal to 1 for the minimal elements of an ascending class of subsets of the variables, then they are all equal to 1 on the entire ascending class.
20. Work out the details of the proof of the first part of Theorem 10.12.
21. Write a program to calculate the log-linear parameters in an $I \times J \times K \times L$ contingency table.
22. Write a program to calculate the log-linear parameters in a contingency table with user defined size.
23. In the proof of Theorem 10.15, the possible difference in sign between $e_{a^*,(a,a')}$ and $e_{a^*,a}$ is mentioned. Determine when this change does or does not occur.
24. Illustrate Theorem 10.15 for the model of conditional independence in a $2 \times 2 \times 2$ table.
25. Define canonical statistics on the contingency table, so that the canonical parameters in an exponential family representation of all positive distributions are the λ balanced parameters.

Chapter 11
Log-Linear Models: Interpretation

Abstract This chapter starts with the specification and handling of regression type problems for categorical data. The log-linear parameters can be transformed into multiplicative parameters, and these are useful in dealing with the regression problem for categorical variables, where this approach provides a clear and testable concept of separate effects versus joint effect of the explanatory variables. Further topics related to the use of log-linear models in data analysis are also considered. First, the selection and interpretation of log-linear models are illustrated in regression type and non-regression type problems, using real data sets. Two special classes of log-linear models, decomposable and graphical log-linear models, are presented next. Decomposable log-linear models may be seen as direct generalizations of conditional independence. Graphical log-linear models, which are the basis of many current applications of log-linear models, may also be interpreted using generalized conditional independence statements, called Markov properties. Further, these models admit a representation using graphs, where the nodes are the variables in the model. Next, a representation of every log-linear model as the intersection of several log-linear models is discussed, where all of the latter models belong to one of two classes of simple log-linear models. One is the model of conditional joint independence of a group of variables, given all other variables (and graphical log-linear models may be represented as intersections of such models only) and (in the case of non-graphical models) no highest-order conditional interaction among a group of variables, given all other variables.

In this chapter, further topics related to log-linear models and their application are discussed.

The chapter starts with a section quite different from the rest of the book. The analyses of small real data sets are presented to illustrate how log-linear models may be applied in standard statistical problems.

11.1 The Regression Problem for Categorical Variables

In this section, one of the most important applications of log-linear models is discussed: the handling of the regression problem for categorical variables. Some preliminary remarks with respect to this were made in Chap. 1.

In regression analysis, one wishes to answer the following questions. Which of the potential explanatory variables have effects on the response? How to quantify the strengths of the effects? Are the existing effects separable or are there joint effects?

Out of these questions, the meaning of the third one needs to be clarified. An alternative way to formulate the third question is to ask whether the effects of the explanatory variables are independent or is there effect modification among them. To answer these questions (and to make the third question more precise), the effects of the potential explanatory variables on the conditional odds of the response variable will be investigated under various log-linear models for the joint distribution.

To be more specific, let Y be a response variable and X and Z potential explanatory variables. For simplicity, in this illustration, all variables are assumed to be binary. The conditional odds of response, given the categories of the explanatory variables, may be written as

$$\frac{p(Y=1|X=x,Z=z)}{p(Y=2|X=x,Z=z)} = \frac{p(Y=1,X=x,Z=z)/p(X=x,Z=z)}{p(Y=2,X=x,Z=z)/p(X=x,Z=z)} =$$

$$\frac{p(Y=1,X=x,Z=z)}{p(Y=2,X=x,Z=z)}.$$

Using the multiplicative representations of the numerator and of the denominator, one obtains that

$$\frac{p(Y=1|X=x,Z=z)}{p(Y=2|X=x,Z=z)} = \frac{\beta^{\emptyset}\beta_1^Y\beta_x^X\beta_z^Z\beta_{1x}^{YX}\beta_{1z}^{YZ}\beta_{xz}^{XZ}\beta_{1xz}^{YXZ}}{\beta^{\emptyset}\beta_2^Y\beta_x^X\beta_z^Z\beta_{2x}^{YX}\beta_{2z}^{YZ}\beta_{xz}^{XZ}\beta_{2xz}^{YXZ}}. \tag{11.1}$$

This seemingly complicated formula is simplified as follows:

$$\frac{\beta_1^Y\beta_{1x}^{YX}\beta_{1z}^{YZ}\beta_{1xz}^{YXZ}}{\beta_2^Y\beta_{2x}^{YX}\beta_{2z}^{YZ}\beta_{2xz}^{YXZ}}.$$

In fact, all multiplicative parameters associated with subsets not containing Y cancel. This formula can be rearranged, to see the decomposition of the factors affecting the conditional odds of the response variable:

$$\frac{\beta_1^Y}{\beta_2^Y}\,\frac{\beta_{1x}^{YX}}{\beta_{2x}^{YX}}\,\frac{\beta_{1z}^{YZ}}{\beta_{2z}^{YZ}}\,\frac{\beta_{1xz}^{YXZ}}{\beta_{2xz}^{YXZ}}.$$

Such a representation is always possible. The first term may be interpreted as the overall odds of $Y=1$ versus $Y=2$, which is not affected by X or Z. The second and the third terms are the individual effects of X and Z on the odds. These effects are individual, because the multiplier implied by $X = x$ is the same, irrespective of the category of Z, and similarly for the effect of $Z = z$. In view of this, the last term is the joint or interaction effect of the two explanatory variables on the response variable.

If one wishes to decide which effects are present, various log-linear models may be formulated and tested based on the available data.

The model Y, XZ was called multiple independence so far. Under this model,

$$p(y,x,z) = \beta^{\emptyset}\beta_y^Y\beta_x^X\beta_z^Z\beta_{xz}^{XZ},$$

and (11.1) yields the following:

$$\frac{p(Y=1|X=x,Z=z)}{p(Y=2|X=x,Z=z)} = \frac{\beta_1^Y}{\beta_2^Y},$$

meaning that the conditional odds of $Y = 1$ versus $Y = 2$ does not depend on the categories of the explanatory variables; thus, these have no effect on the response variable. This is the answer to the first question of regression analysis, and in this case, the further questions are irrelevant.

Note, that the overall odds β_1^Y/β_2^Y is, in general, different form the marginal odds $p(1,+,+)/p(2,+,+)$.

The model YX, XZ was interpreted up to now, as the model of conditional independence of Y and Z, given X. In the regression context, a different interpretation is relevant. Under this model,

$$p(y,x,z) = \beta^{\emptyset}\beta_y^Y\beta_x^X\beta_z^Z\beta_{yx}^{YX}\beta_{xz}^{XZ},$$

and (11.1) yields the following:

$$\frac{p(Y=1|X=x,Z=z)}{p(Y=2|X=x,Z=z)} = \frac{\beta_1^Y}{\beta_2^Y}\frac{\beta_{1x}^{YX}}{\beta_{2x}^{YX}}.$$

That is, the conditional odds may be written as the product of two factors. One is the overall odds of $Y = 1$ versus $Y = 2$, and the other one is the effect of $X = x$ on this odds. Because $\beta_{1x}^{YX}\beta_{2x}^{YX} = 1$ for all categories x of X, the effect of $X = x$ on the odds may be written as

$$\frac{\beta_{1x}^{YX}}{\beta_{2x}^{YX}} = (\beta_{1x}^{YX})^2 = (\exp\lambda_{1x}^{YX})^2.$$

This quantifies the strength of the effect of X on (the odds of) Y, answering the second question related to regression analysis for categorical variables.

There is another conditional independence model, YX, YZ, which, because of the asymmetry in the roles of the variables in the regression problem, may be seen as describing a different relationship among the variables. But usually, this model is not considered relevant for the regression problem. As seen after (11.1), all terms not containing Y cancel from the expression for the conditional odds. Therefore, usually, all possible interactions among the explanatory variables are allowed in a regression model. This does not make the expression for the effects on the odds of the response variable more complicated, and in a regression analysis, model simplicity is sought for in terms of the effects on the response, not in terms of the relationship among the explanatory variables.[1]

Another relevant model is the no second-order interaction model YX,YZ,XZ. Under this model,

$$p(y,x,z) = \beta^{\emptyset}\beta_y^Y\beta_x^X\beta_z^Z\beta_{yx}^{YX},\beta_{yz}^{YZ},\beta_{xz}^{XZ},$$

and (11.1) yields the following:

$$\frac{p(Y=1|X=x,Z=z)}{p(Y=2|X=x,Z=z)} = \frac{\beta_1^Y}{\beta_2^Y}\frac{\beta_{1x}^{YX}}{\beta_{2x}^{YX}}\frac{\beta_{1z}^{YZ}}{\beta_{2z}^{YZ}}.$$

According to this model, out of the two potential explanatory variables, both have effects on the odds of the response variable. These effects are independent or, better to say, separable, because the effect of changing X from one of its categories to another one has the same effect on the odds of Y, irrespective of Z, and also the effect of Z is unrelated to the actual category of X. Here, both explanatory variables have effects. As these effects are multiplicative, the stronger effect is the one farther from one. When both X and Z are binary, one obtains that

$$\frac{\beta_{11}^{YX}}{\beta_{21}^{YX}}\frac{\beta_{12}^{YX}}{\beta_{22}^{YX}} = 1,$$

and

$$\frac{\beta_{11}^{YZ}}{\beta_{21}^{YZ}}\frac{\beta_{12}^{YZ}}{\beta_{22}^{YZ}} = 1,$$

thus, one has a single value of the effect of X (the overall odds of Y is either multiplied by this value or by its reciprocal) and a single value for the effect of Z, and the comparison is straightforward. When the explanatory variables are not binary, the strength of the effect has to be evaluated for pairs of categories.

For example, when X has four categories and Z has three categories, Tables 11.1 and 11.2 show hypothetical values of the multiplicative parameters. With these data, the multipliers of the overall odds of $Y = 1$ versus $Y = 2$ for the categories of X are 26.11, 0.31, 0.56, and 0.21, and the same multipliers for the categories of Z are 0.01, 9.42, and 8.77. Overall, the multipliers associated with Z are farther from one than the multipliers associated with X, but it could be the case that one explanatory variable has a strong effect in one of its categories and the other one also has a strong effect in one of its categories.

[1] One might think that in a data analytic situation, model fit cannot be worse by allowing more interactions, that is, a larger model. Indeed, the test statistics measuring the deviation between observed and estimated data are not going to be bigger, but they are evaluated against different reference distributions because of the different degrees of freedom. Therefore, one cannot guarantee that a larger model will always show better fit to the data.

Table 11.1 Hypothetical values of β_{yx}^{YX}

	$X=1$	$X=2$	$X=3$	$X=4$
$Y=1$	5.11	0.56	0.75	0.46
$Y=2$	0.20	1.79	1.33	2.15

Table 11.2 Hypothetical values of β_{yz}^{YZ}

	$Z=1$	$Z=2$	$Z=3$
$Y=1$	0.11	3.07	2.96
$Y=2$	9.09	0.33	0.34

When none of the above models is an acceptable description of the data[2] one has, no simplification to the log-linear representation is possible, and from the multiplicative version of the representation, one obtains that

$$\frac{p(Y=1|X=x,Z=z)}{p(Y=2|X=x,Z=z)} = \frac{\beta_1^Y}{\beta_2^Y}\frac{\beta_{1x}^{YX}}{\beta_{2x}^{YX}}\frac{\beta_{1z}^{YZ}}{\beta_{2z}^{YZ}}\frac{\beta_{1xz}^{YXZ}}{\beta_{2xz}^{YXZ}}.$$

In this case, in addition to the separate effects of the two explanatory variables, one also has a joint effect of them on the odds of the response variable. This joint effect is $\beta_{1xz}^{YXZ}/\beta_{2xz}^{YXZ}$, and it modifies the odds on top of the individual effects. The existence of the joint effect is sometimes referred to by saying that one explanatory variable modifies the effect of the other one or that effect modification is present.

For example, it may be the case that in a graduate-level statistics class, the overall odds of passing the midterm test is 3, that is, the probability of passing is 0.75, and those whose undergraduate major was statistics have an odds three times higher than average. Further, assume that those who study for the midterm at least 7 hours have an odds of passing which is two times higher than the overall odds. If the two explanatory variables had independent effects (no interaction, no effect modification), then the odds of passing versus failing for someone who majored in statistics and studied for at least 7 hours would be $3 \times 3 \times 2$. Similarly, for someone not having majored in statistics but having studied for at least 7 hours, the odds of passing versus failing would be $3 \times 1/3 \times 2$ and similarly for the other combinations of the explanatory variables. If this is not the case, the odds of passing versus failing for someone who had majored in statistics and had studied for more than 7 hours is not 18, rather 14.4; then, in addition to the separate effects, there is also a joint effect, which in this particular category combination is $14.4/18 = 0.8$.

[2] Here the data can be population data, but in the usual setup of statistics, such decisions are done by applying tests of hypotheses to data from a sample.

The log-linear representation, which is always possible, thus is a nonrestrictive assumption and is called by some authors the saturated model. In this book, a model is a restrictive assumption. As the second-order interaction is allowed to be present, the shorthand notation for this situation is YXZ.

Thus, the answer to the question of separable effects only or an additional joint effect depends on the presence of the YXZ interaction term. If it is there, there is a joint effect on top of the individual (or independent or separable) effects. This is the answer to the third question related to the regression problem.

The interpretation of log-linear models for general regression type and non-regression type problems will be discussed in the next section. Among others, the case of non-binary response variables will also be treated.

The logarithm of the conditional odds on the left-hand side of (11.1) is called the logit of Y. Just like the conditional odds was written as the product of multiplicative parameters under the various models, the logit can be written as the sum of log-linear parameters, and the two ways of thinking about log-linear models are equivalent.

The approach described here is closely related to the so-called logistic regression. In one variant, one has continuous explanatory and a categorical response variable. Then, the logit of the response is approximated by a linear function of the explanatory variables. This approach suffers from the problems discussed in Chap. 1. In particular, because of the continuous explanatory variables, the conditional odds are difficult to estimate, and interaction effects are problematic to define. On the other hand, many real data sets, in particular in biology, may be well described with such models. In another variant, the explanatory variables are discrete, and it is essentially identical with the regression type analysis described here. A general reference for these methods is [39].

11.2 Interpretation in Regression Type and Non-regression Type Problems

The fundamentals of the application of log-linear models to deal with regression type problems were presented in the previous section. Here we discuss the extension to several explanatory variables and non-binary response variables. As so often in statistical analysis, one may have a specific model to test, or one may wish to select a model to describe the relevant features of the underlying population. In the case, when one has a model to test, the following model selection steps are omitted.

Usually, a somewhat informal search procedure is applied to find an appropriate model. Such a model has to be simple and has to show acceptable fit to the available data. Before discussing how to find such a model, both of these concepts need clarification.

The desired simplicity of the log-linear model has to be interpreted with respect to the regression type problem. Simplicity, in general, means few and low-order interactions. In a regression type problem, simplicity is sought for in terms of the effects of the explanatory variables on the response variable and not in terms of how the explanatory variables are related to each other according to the model.

Therefore, it is customary to allow the highest-order interaction to be present among the explanatory variables. Limitations of this practice are going to be discussed after an example for a model search procedure. It may happen in practice that it is not straightforward, which one out of two models is simpler. One may contain several lower-order interactions while the other fewer but of higher order. If the fit of both models is satisfactory, one may choose the substantially more relevant or interesting one, as a hypothetical description of the data.

Model selection is often guided by tests of model fit for the various models considered. In this section, the Pearson chi-squared and the likelihood statistics are used to judge model fit. In the case of the example to be discussed, the application of the asymptotic reference distribution seems justified.[3] Simpler applications of these test statistics were presented in Sect. 5.2. Alternative approaches to model testing will be briefly mentioned in Chap. 13.

The data to be used to illustrate model search and interpretation in a regression type problem is available in the *car* package of R; see [29]. The data are from a national survey conducted in Chile in 1988 by FLACSO and concern the willingness of the respondents to cast their votes in the national plebiscite.[4] The sample size was 2700, and although no special sampling information is available, it is assumed that the sampling procedure led to a multinomial distribution, at least, approximately.

For the first illustration, the response variable (V) was recoded with the categories willing to vote and not willing to vote, suppressing the intended vote. The data contain three further categorical variables, gender of the respondent (G), educational level of the respondent (E), and the region (R), where the respondent lived. The question of the analysis is to what extent the explanatory variables G, E, and R affect the response variable V. Table 11.3 contains the generating classes of many relevant log-linear models and information about model fit.

The first model in Table 11.3 may be interpreted as the response variable being independent of the joint distribution of the explanatory variables. This model means that the conditional odds of the response variable do not depend on the categories of the explanatory variables. Equivalently, none of the explanatory variables have an effect on the response. When using the conventional 0.05 threshold, the data provide strong evidence that this model does not fit.

The next three models have the same structure but, of course, different substantial meanings. Each assumes that out of the three explanatory variables, only one has a direct effect. The second model says this variable is G, so the conditional odds of voting does depend on gender but does not depend on E or R. This does not mean that E or R would be independent of V. The model may also be interpreted as conditional independence of V form the joint distribution of ER, given G. Therefore, V may be not independent from E and R, so they may be seen as having some effect

[3] See [70] for simulation results about using the asymptotic chi-squared distributions as reference distributions for finite samples.

[4] In 1988, Chile had a referendum, where voters were asked to decide, whether the then de facto leader of the country should or should not remain in power for another 8 years. The leader was Augusto Pinochet, who had assumed power in 1973 as a result of a military coup d'état.

Table 11.3 Fit of selected regression type log-linear models to the Chile data (see text for explanation)

Generating class	Df	Pearson	p-value	likelihood ratio	p-value
V, GER	29	101.41	0.00	106.14	0.00
VG, GER	28	68.04	0.00	72.94	0.00
VE, GER	27	70.34	0.00	70.82	0.00
VR, GER	25	87.07	0.00	93.43	0.00
VG, VE, GER	26	40.37	0.03	41.85	0.03
VG, VR, GER	24	55.55	0.00	60.28	0.00
VE, VR, GER	23	55.55	0.00	56.29	0.00
VG, VE, VR, GER	22	27.98	0.18	27.63	0.19
VGE, VR, GER	20	25.54	0.18	24.86	0.21
VGR, VE, GER	18	16.23	0.58	16.75	0.54
VER, VG, GER	20	25.54	0.18	24.86	0.21

on V,[5]; although this effect is not direct, it rather occurs through G. The conditional odds of V is only influenced by the category of G and not by the category of E or R. These three models are also rejected, as the data provide clear evidence against them.

The next three models assume that two of the explanatory variables have direct and separable effects on response. These models are also rejected.

The next model, VG, VE, VE, GER, assumes that all three explanatory variables have effects on response, and these effects may be separated.[6] Using the conventional level of 0.05, the data provide no evidence against this model. Tables 11.4, 11.5, 11.6, 11.7 give the values of the multiplicative parameters used in Sect. 11.1.

Table 11.4 Multiplicative parameters for Vote in the separable effect log-linear model for the Chile data

Vote	Vote	No vote
β_v^V	1.670	0.599

Table 11.5 Multiplicative parameters for Gender in the separable effect log-linear model for the Chile data

Gender	Female	Male
β_g^G	1.079	0.927

[5] Because this is a regression type analysis, the association between the response and some of the explanatory variables is interpreted as the effect of the latter.

[6] This situation is often called independent effects, but the word independence is used in so many meanings in statistics that separable may be a clearer description.

Table 11.6 Multiplicative parameters for Education in the separable effect log-linear model for the Chile data

Education	Primary	Secondary	Post-secondary
β_e^E	1.593	1.444	0.435

Table 11.7 Multiplicative parameters for Region in the separable effect log-linear model for the Chile data

Region	North	Central	South	Metro Santiago	Central Santiago
β_r^R	0.823	1.514	1.773	0.172	2.629

The multiplicative parameters are calculated from the maximum likelihood estimates[7] of the true distribution and are maximum likelihood estimates of the true multiplicative parameters; see Proposition 4.2.

According to the estimates, the overall odds for planning to vote versus not to vote is

$$\frac{1.670}{0.599} = 2.799.$$

This overall odds is modified by the separate effects of the explanatory variables. For a female, this odds is multiplied by, see Table 11.8,

$$\frac{0.889}{1.125} = 0.790,$$

and this is the multiplicative effect of considering a female, irrespective of the level of education or the region where this person lives. So for a female, the odds of voting versus not voting is

$$2.799 \times 0.790 = 2.211.$$

The effect of someone being a male is the reciprocal of the effect of being female, so the odds of a man to vote is

$$2.799 \times \frac{1}{0.790} = 3.543.$$

Thus, the odds for a male to vote versus not to vote is about $3.543/2.211 = 1.602$ times higher than for a female. To see the effect of education on the odds of reported voting versus nonvoting, the overall odds for someone with primary education is modified (see Table 11.9), by the multiplier

$$\frac{0.872}{1.146} = 0.761,$$

[7] These will be discussed in Sect. 12.1.

while the multiplier is 0.817 for secondary education and 1.607 for post-secondary education. These numbers indicate that higher levels of education go with higher odds of voting versus not voting, but the difference between primary and secondary education is not big.

Similar calculations are possible to quantify the effect attributable to the variable Region. According to the data, the highest odds of voting versus not voting is in the region North, and the lowest is in Central Santiago; see Table 11.10.

Taking all effects into account, the odds of voting versus not voting for a man with post-secondary education living in the region North is

$$\frac{1.670}{0.599}\frac{1.125}{0.889}\frac{1.268}{0.789}\frac{1.136}{0.880} = 7.319.$$

The estimated frequencies, which were compared in the tests with the observed ones, can be fully obtained from the multiplicative parameters given in Tables 11.4, 11.5, 11.6, 11.7, 11.8, 11.9, 11.10 (and the overall effect, which is 24.187) using (10.23). The tests carried out suggest that there is no reason to assume that further effects than those of the individual variables exist, and these act independently from each other. Thus, the parameters estimated contain all information, which may be seen as relevant.

Of course, the word “relevant” in the previous paragraph refers to the lack of statistical significance, that is, that one sees no reason to believe that additional effects exist, because the deviation between observed and estimated data does not belong to the most unlikely deviations, when the current model is assumed. Substantive significance, that is, whether the effects allowed in the model are of a meaningful magnitude given the research problem, the mode of data collection, and any other relevant information and, at the same time, effects not allowed to exist by the model are below such a threshold, may be different from this. Also, this example shows that one may not give a straightforward causal interpretation to a regression type analysis. It is very problematic to think, based solely on the data, that living in a particular region makes people more or less willing to vote. It seems more reasonable to think that certain factors, which make people to be more or less willing to vote, are present in the different regions to different extents.

Table 11.8 Multiplicative parameters for the Vote by Gender interaction in the separable effect log-linear model for the Chile data

		Female	Male
β_{vg}^{VG}	Vote	0.889	1.125
	No vote	1.125	0.889

To illustrate regression type analyses, when the response variable is not binary, the original version of the Vote variable is used, with categories will vote yes, will vote no, will abstain, undecided. The model of separable effects shows a good fit, with a p-value of about 0.5. When the response variable has more than two cate-

Table 11.9 Multiplicative parameters for the Vote by Education interaction in the separable effect log-linear model for the Chile data

		Primary	Secondary	Post-secondary
β_{ve}^{VE}	Vote	0.872	0.904	1.268
	No vote	1.146	1.106	0.789

Table 11.10 Multiplicative parameters for the Vote by Region interaction in the separable effect log-linear model for the Chile data

		North	Central	South	Metro Santiago	Central Santiago
β_{vr}^{VR}	Vote	1.136	0.938	1.042	1.008	0.893
	No vote	0.880	1.066	0.960	0.992	1.120

gories, there are several conditional odds which may be of interest. When there is an order among the categories, the odds of being in a higher versus the next lower category is a useful comparison. In the present case, the odds of choosing to vote yes, or no, or abstain, relative to being undecided, will be studied. The estimated multiplicative parameters for the variable Vote are given in Table 11.11. The three odds mentioned are estimated to be

$$\frac{1.867}{0.971} = 1.923, \ \frac{1.754}{0.971} = 1.806, \ \frac{0.314}{0.971} = 0.323,$$

respectively. The multiplicative effects of the categories of Education on the odds can be determined from the Vote by Education interaction term, given in Table 11.12. It is interesting to see that people with primary education were much more undecided as to how to vote than people with post-secondary education and that in the former group, the yes vote was more popular, than the no vote, while in the latter group the preferences went in the opposite direction. The odds of a yes vote versus being undecided for the three levels of education may be estimated as

$$1.923\frac{1.236}{1.501} = 1.583, \ 1.923\frac{0.826}{1.027}1.547, \ 1.923\frac{0.979}{0.649} = 2.901,$$

indicating that in this respect, people with primary and secondary education are quite similar, but those having a post-secondary education are different from the others. The analysis may be continued similarly to the case of a binary response variable.

The foregoing example illustrated the general principles of handling a regression type problem with log-linear analysis. If there are higher-order interaction terms in the model, their effects take places on top of the lower-order terms. For example, if, in addition to the individual effects of the three explanatory variables, there was a VGE interaction, the odds for the vote being yes versus undecided conditioned on categories g, e, and r of the explanatory variables would be obtained as

$$\frac{\beta_y^V}{\beta_u^V}\frac{\beta_{yg}^{VG}}{\beta_{ug}^{VG}}\frac{\beta_{ye}^{VE}}{\beta_{ue}^{VE}}\frac{\beta_{yr}^{VR}}{\beta_{ur}^{VR}}\frac{\beta_{yge}^{VGE}}{\beta_{uge}^{VGE}}.$$

Table 11.11 Multiplicative parameters for the Vote with four categories in the separable effects log-linear model for the Chile data

	Yes	No	Abstain	Undecided
β_v^V	1.867	1.754	0.314	0.971

Table 11.12 Multiplicative parameters for the Vote with four categories by Education in the separable effects log-linear model for the Chile data

		Primary	Secondary	Post-secondary
β_{ve}^{VE}	Yes	1.236	0.826	0.979
	No	0.736	0.937	1.449
	Abstain	0.732	1.257	1.087
	Undecided	1.501	1.027	0.649

To illustrate a non-regression type analysis, data from the World Values Survey will be used from 1995 to 1997, for the countries Australia, Norway, Sweden, and the USA. In addition to this Country variable, there is a Religion, a Degree (university degree), and a Gender variable. This data set is also available in R, in the package called *effects*; see [28]. As one is interested in the relationships among these variables, without having preference for any particular type of structure, there are a large number of log-linear models which may be tried in a model search procedure. The fit of some of these models is reported in Table 11.13. The model search starts with the model containing all four second-order interactions. This model shows good fit; thus, the next four models omit one of the interactions.

When the *R* by *D* by *G* interaction is omitted, the *p*-values go up, and this may be seen as somewhat counterintuitive. Although the *p*-value is not a measure of model fit, it appears that the data provide one with stronger evidence against a model which is more flexible in reproducing the observed data, because it allows more interactions. Indeed, the values of the statistics go up, as one interaction is omitted from the model, so one might expect a smaller *p*-value. This would be correct if the number of degrees of freedom did not change.[8]

Because of the highest *p*-value, the model search continues with the model *CRD*, *CRG*, *CDG* and sees if further interactions may be omitted. Given that the *CRD*, *CDG*, *RDG* model also shows good fit, omitting the interaction not present in the

[8] Similarly, it is possible that in a regression type analysis, one allows the highest-order interaction among the explanatory variables but would obtain a higher *p*-value if this interaction was omitted.

latter model is tried. This leads to the model *CRD*, *CDG*, *RG*, which also shows good fit, so further interactions will be removed.

At this point, there does not seem to be a unique choice as to which interaction to remove. The decision should combine substantive (e.g., conformity with existing theory or relevance of certain interactions) and statistical (e.g., structural simplicity) considerations. In this illustration, only the latter aspect is used, and the next model to try is *CRD*, *CDG*. This model is particularly attractive, because it has a simple interpretation: *R* and *G* are conditionally independent, given the combined *CD* variable. Unfortunately, the data provide strong evidence against this model, and, it seems, the *RG* interaction needs to be kept in the model.

In the next model, the *CRD* interaction is omitted, leading to *CR*, *RD*, *CDG*, *RG*. The *p*-values of 0.03 and 0.04 mean that the data do provide some evidence against this model at the conventional level of 0.05. It seems interesting to try, what happens, if the other second-order interaction, *CDG*, is omitted. This leads to the model of *CRD*, *CG*, *DG*, *RG*. With *p*-values of 0.06, this hypothesis is not rejected, but one sees little difference between the goodness of fit of these models.

The two latter models do show a kind of similarity. They both have a variable playing a central role. In one, this is *R*; in the other this is *G*. This central variable is in the first-order interaction with all three remaining variables, and these remaining three variables are in a second-order interaction with each other.

While the *CRD*, *CDG* model had a conditional independence interpretation, the latter two models do not have one. This issue will be discussed in the next section.

Further models obtained by omitting some of the interactions may also be tested, but the results (not shown here) suggest rejecting those models. The interpretations of the last two models given above may be completed with the inspection of the estimates of the multiplicative parameters. The multiplicative parameters in this context increase or decrease the cell probabilities. This is not going to be shown here.

Table 11.13 Fit of selected log-linear models to the World Values Survey data (see text for explanation)

Generating class	Df	Pearson	*p*-value	likelihood ratio	*p*-value
CRD, CRG, CDG, RDG	3	4.21	0.24	4.20	0.24
CRD, CRG, CDG	4	4.23	0.38	4.22	0.38
CRD, CRG, RDG	6	13.54	0.04	13.51	0.04
CRD, CDG, RDG	6	8.25	0.22	8.00	0.24
CRG, CDG, RDG,	6	15.43	0.02	15.34	0.02
CRD, CDG, RG	7	8.27	0.31	8.03	0.33
CRD, CDG	8	48.74	0.00	48.85	0.00
CR, RD, CDG, RG	10	20.30	0.03	18.71	0.04
CRD, CG, DG, RG	10	17.74	0.06	17.54	0.06

A general strategy for exploratory analyses, that is, analyses without a previously specified model, is stepwise selection, illustrated in the example above. Stepwise selection may be performed in the forward or in the backward way. The former means starting with joint independence, and adding interactions one by one, to find a model which does not need to be rejected. The latter means starting with the model including the highest-order interactions and leaving out interactions to find the simplest which does not need to be rejected; see [25]. Such search strategies may be combined with substantive preferences about meaningful models.

The model search procedures illustrated here essentially assume that models which allow for more interactions do not exhibit a worse fit to the data, than models allowing fewer interactions. The reliance on this assumption is clear in regression type problems, when one allows all interactions among the explanatory variables, which is correct, if the fit of such a model cannot be worse than the fit of a model which does not allow all interactions among the explanatory variables. But also in non-regression type problems, if one sees acceptable model fit, an interaction is removed, expecting that fit may be worse but still acceptable. It is true that the maximum likelihood estimates under a model cannot be further away from the data, than estimates under a model allowing for fewer interactions. Indeed, models allowing for more interactions always can produce the same estimates as models with fewer interactions, by choosing one as the values of the multiplicative parameters present in the first model but not present in the second one. But the numbers of degrees of freedom are also different for the two models, see Sect. 12.1, and the reduction in the deviation between data and estimates may be smaller than expected based on the reduction in the number of degrees of freedom. It may happen, therefore, that the model allowing more interactions is rejected, while the model allowing fewer interactions is not rejected at the same level. A formal testing procedure based on the differences of the likelihood ratio statistics is possible.

An issue to be taken into account in model search procedures is the so-called multiple testing problem. When several tests of hypotheses are carried out, each with error type I probability α, then if all of them are true, the probability that there will be at least one to be rejected is not α. The situation is easily understood when the tests are independent from each other. Then, if one hypothesis is rejected with probability α, then the probability that out of m test there will be at least one leading to the rejection of the hypothesis is

$$1-(1-\alpha)^m,$$

which for $\alpha = 0.05$ and $m = 10$ is about 0.4. One possible strategy is to reduce α, so that the overall rejection probability becomes smaller. For example, with $\alpha = 0.005$, the overall rejection probability is about 0.049, fairly close to 0.05. The strategy of using α/m as the individual error type I probability, when m tests are to be carried out and the overall rejection rate is desired to be around α, is called the Bonferroni correction. Obviously, the test of hypotheses for different log-linear models, based on the same data, cannot be seen as independent; thus, the real problem is more involved than the description given here. See [58] as an introduction to the problem.

A closely related formulation which sheds further light on the issue is that in a model search procedure, one often chooses the model with the largest p-value, just like it was done when the model *CRD*, *CRG*, *CDG* was chosen above; see Table 11.13. It is easily seen that if one tests a single hypothesis, then the achieved

p-value, which is a random variable, so will be denoted as P_a, has a uniform distribution, at least in the continuous case.

Proposition 11.1. *If the hypothesis is true, and the test statistic has a continuous distribution, then the distribution of the achieved p-value is uniform on the interval* $[0,1]$.

Proof. For an arbitrary continuous random variable X, with distribution function F_X, $F_X(X)$ has a uniform distribution. Indeed, for an arbitrary value $0 \leq u \leq 1$,

$$P(F_X(X) < u) = P(X < F_X^{-1}(u)) = F_X(F_X^{-1}(u)) = u.$$

But if X is the test statistic, $1 - P_a = F_X(X)$, so $1 - P_a$ is uniform on $[0,1]$; thus P_a also has this distribution. □

As the Pearson chi-squared and the likelihood ratio statistics have an asymptotic chi-squared distribution, for large samples they may be considered fairly close to being continuous.

Of course, Proposition 11.1 only means that the proportion of the values of X, which are characterized by an achieved probability level of α or more, is exactly α, which is simply the definition of the achieved probability level.

The choice of the log-linear models, however, was not governed by the p-values, rather by the maximal out of a number of p-values. If X is a continuous random variable, with distribution function F_X, and one has k independent copies of X, that the maximal one is smaller than u, means all of them are smaller than u, which has probability

$$P(max_{i=1,\ldots,k} X_i < u) = F_X^k(u),$$

thus, the maximum has F_X^k as its distribution function; thus what is a small or large p-value should be judged relative to a power of a uniform distribution.

The same principles of model search may be used for larger tables, that is, more variables, too. But for very large problems, the asymptotic testing methods used here may not be applicable, because of insufficient sample size. The comments made at the end of Sect. 5.4.3 apply in this case, too. For some of the approaches that may be used there, see Chap. 13.

11.3 Decomposable and Graphical Log-Linear Models

The interpretation and estimation of log-linear models are further facilitated by considering two special classes of them, decomposable and graphical log-linear models. The latter class also leads to very broad generalizations, some of which are discussed later in the book, which come up in a wide variety of applied fields.

A log-linear model, or its generating class $\mathscr{C} = (C_1, \ldots, C_k)$, is said to be decomposable, if it has less than three maximal interactions or if there is an order of the subsets of variables which are the maximal interactions in $\mathscr{C}$ such that

$$C_j \cap (\cup_{i=1}^{j-1} C_i) = C_j \cap C_{i(j)}, \text{ for some } 1 \leq i(j) \leq j, \text{ for all } j = 1, \dots k.$$

The meaning of the definition is that in the decomposable order, for every subset, its intersection with the union of all of the previous subsets is the same as its intersection with one of them. For example, $(ABC, ABD, BDEF, ACG, BDH)$ is a decomposable order, but (ABC, ABD, ACD) is not one, and there is no decomposable ordering of this generating class. The order (AD, BE, ABC) is not decomposable, but the order (ABC, AD, BE) is decomposable. From now on, decomposable generating classes will always be considered in a decomposable order, unless said otherwise.

Decomposable log-linear models have a very simple structure. To understand it, first a result about the structure of distributions in certain log-linear models is presented.

Theorem 11.1. *Let A be a variable, perhaps a combined one, which appears in only one maximal interaction in the generating class of a log-linear model. Let B be the other variable or combination of other variables which appear in the same maximal interaction, and also in others, and let C be the collection of all other variables. Then for all distributions belonging to the log-linear model,*

$$A \perp\!\!\!\perp C | B,$$

where $\perp\!\!\!\perp$ *means independent.*

Proof. In this case $COR(A, B|C) = 1$, and the claim is directly implied by Theorem 10.10. □

This result is also useful in interpreting log-linear models, especially for non-regression type problems. If a variable appears only in a single maximal interaction, then it is conditionally independent from the rest of the variables, given the variables with which it is together in the maximal interaction. The variable may be omitted, and only the rest of the model needs to be interpreted.

A generalization of the previous result describes the structure of distributions in decomposable log-linear models.

Theorem 11.2. *Let the decomposable generating class* $\mathscr{C}$ *consist of the maximal interactions* $C_1, \dots, C_k$ *and denote* $LL(\mathscr{C})$*, the log-linear model generated by it. Then, if for a probability distribution* $\boldsymbol{p}$*,*

$$\boldsymbol{p} \in LL(\mathscr{C})$$

then for all $j = 3, \dots, k$*, in the*

$$\cup_{i=1}^{j} C_i$$

marginal,

$$(C_j \setminus \cup_{i=1}^{j-1} C_i) \perp\!\!\!\perp (\cup_{i=1}^{j-1} C_i \setminus C_j) \,|\, (C_j \cap C_{i(j)}) \text{ under } \boldsymbol{p}. \quad (11.2)$$

Proof. The proof is by induction on the number of interactions j, first showing it is true for $j = k$ and then reducing the number of maximal interactions to $k-1$. When the number of maximal interactions goes below three, the proof is completed.

Let be $j = k$. Then, define

$$A = C_k \setminus \cup_{i=1}^{k-1} C_i.$$

The set of variables A is not empty, because in that case C_k would contain no variable not contained by the union of the previous interactions, but then, by decomposability, there would be an earlier interaction containing C_k, and then C_k would not be maximal. Further, let

$$B = C_k \setminus A$$

$$C = \cup_{i=1}^{k-1} C_i \setminus C_k.$$

Theorem 11.1 implies that (11.2) holds for $j = k$.

Then, marginalize over the variables A. The remaining variables are

$$\cup_{i=1}^{k-1} C_i,$$

and as

$$C_k \setminus A \subseteq C_{i(j)},$$

the maximal interactions on this marginal are $C_1, \ldots, C_{k-1}$. Thus, the number of interactions gets reduced, and the induction proof is completed. □

In fact, the proof showed a collapsibility property of decomposable log-linear models. On the marginal defined by the union of the first j interactions, the model is specified by the first j interactions. The smallest nondecomposable log-linear model is AB, AC, BC, the model of no second-order interaction. Here, the union of the first two interactions is the whole table, but the first two interactions suggest conditional independence holds on the table, which is not true, as the third interaction also affects the table. The order (AD, BE, ABC) is not decomposable, and the first two interactions suggest multiple independence holds on the $A \times B \times D \times E$ marginal. However, the AB interaction implied by the third maximal interaction changes this. In the decomposable order (ABC, AD, BE), the first two maximal interactions suggest that $BC \perp\!\!\!\perp D | A$ and the third one does not change this.

The interpretation through conditional independences applies to a larger class of log-linear models than the decomposable log-linear models. The rest of this section is devoted to the study of this class, called graphical models.

The name graphical model refers to the fact that every graph on the set of variables defines such a model. In a graph, a subgraph is obtained by considering a subset of the nodes (the variables, in this case) and restricting the edges to this subset. A graph is called complete, if it contains all possible edges among the nodes. A complete subgraph, which is maximal with respect to this property, that is, any graph containing this is not complete, or shortly a maximal complete subgraph, is called a clique of the graph.

Graphs on the set of variables define log-linear models by choosing the cliques of the graph as the maximal allowed interactions or, in other words, using the smallest descending class containing all cliques to define the log-linear model.

In understanding the structure of graphical log-linear models, forbidden configurations play a central role. A subset of the variables is a forbidden configuration of a generating class (or of the log-linear model), if it contains three variables at least, and it is not contained in any of the interactions, but all its proper subsets are contained in some of the interactions. For example, the generating class with maximal elements AB, AC, BC does have a forbidden configuration: ABC. The generating class with maximal elements ABD, ACE, BCF also has a forbidden configuration, ABC.

In fact, the lack of forbidden configurations characterizes graphical log-linear models.

Theorem 11.3. *A generating class $\mathscr{C}$ is graphical with respect to some graph, if and only if it does not have a forbidden configuration.*

Proof. Consider the graph on the variables as nodes, with the maximal allowed interactions as complete subgraphs. One has to see that these subgraphs are also maximal, that is, cliques, if there is no forbidden configuration in $\mathscr{C}$. If, to the contrary, there is a maximal allowed interaction $C_j \in \mathscr{C}$, such that the C_j subgraph is not a maximal complete subgraph, then there is, at least, one more node, say A, that there is no $C_j \cup \{A\}$ interaction in the generating class, but the $C_j \cup \{A\}$ subgraph is complete, and the cardinality of $C_j \cup \{A\}$ is $m \geq 3$. As $C_j \cup \{A\}$ is complete, every pair of nodes in it is contained in some maximal complete subgraph. If there is a triplet of variables in $C_j \cup \{A\}$, which is not contained in any interaction, then it is a forbidden configuration. If there is no such triplet, then every triplet within $C_j \cup \{A\}$ is contained in some interaction. Then see if there is any quadruple in $C_j \cup \{A\}$, which is not contained in any interaction. If such a quadruple exists, it is a forbidden configuration. Otherwise, all quadruples in $C_j \cup \{A\}$ are contained in some interaction. Continue this argument. As there is an m-tuple in $C_j \cup \{A\}$ which is not contained in any interaction (itself), the procedure will find an n-tuple, $n \leq m$, of which every proper subset is contained in some interaction, but itself is not, and is, thus, a forbidden configuration.

To see the converse, assume the maximal allowed interactions are the cliques of a graph. If there was a forbidden configuration, it would be a complete graph not contained in any clique; thus, it would be part of an additional maximal complete subgraph. □

One can prove now that decomposable log-linear models are always graphical.

Theorem 11.4. *Let $\mathscr{C}$ be a decomposable generating class. Then $\mathscr{C}$ is also graphical.*

Proof. Suppose, to the contrary, that $\mathscr{C} = (C_1, \ldots, C_k)$ has a forbidden configuration, say S. If $|S| = s$, it has s maximal proper subsets, say $S_1, \ldots, S_s$, and $|S_j| = s - 1$, $j = 1, \ldots, s$. By assumption, $S \nsubseteq C_i$, for all $i = 1, \ldots, k$, but for every S_j, there is a first C_{s_j}, such that $S_j \subseteq C_{s_j}$. Suppose the last one out of these in the decomposable order is C_{s_s}, which is the first interaction in the decomposable order, which contains S_s.

The subset S_s has a nonempty intersection with all S_j, $j = 1,\dots,s-1$, because these are all sets of cardinality $s-1$ and are all subsets of a set of cardinality s. Further,

$$S_s \cap (\cup_{j=1}^{s-1} S_j) = S_s,$$

Therefore,

$$C_{s_s} \cap (\cup_{j=1}^{s-1} C_{s_j}) \supseteq S_s.$$

Then, because of decomposability, the intersection above is the same as the intersection of C_{s_s} with one of the earlier interactions. This earlier interaction contains S_s, and then C_{s_s} is not the first interaction containing S_s, which contradicts the construction. □

Graphical log-linear models may be interpreted by certain conditional independence statements, and this simplicity makes them the first choice in many applied problems. These conditional independence properties are often referred to as Markov properties. There are several such Markov properties used in various fields of science. For example, certain characteristics of the weather are measured and recorded every day. A simple and useful model is that "yesterday's weather affects tomorrow's weather only through today's weather" and may be formulated more precisely by assuming that

$$X_{t-1} \perp\!\!\!\perp X_{t+1} \,|\, X_t,\; t = 1,2,\dots$$

where X is the measurement in different time points. Such a property is called a Markov chain and may be represented with a graph where X_t is linked to X_{t-1} and X_{t+1}. This expresses intuitively that there is no direct effect between X_{t-1} and X_{t+1}, only an effect through X_t. Indeed, (6.30) is that

$$P(X_{t+1}|X_t, X_{t-1}) = P(X_{t+1}|X_t),$$

that is, once X_t is given, X_{t-1} provides no additional information with respect to the distribution of X_{t+1}. One might say that the Markov chain has a one-step memory, because it remembers (i.e., depends on) the previous stage only. A Markov chain with k-step memory is defined as

$$X_{t-k} \perp\!\!\!\perp X_{t+1} \,|\, X_t, X_{t-1}, \dots X_{t-k+1}, \tag{11.3}$$

more precisely, the conditional independence is assumed not only for X_{t-k} but also for all previous time points.

The one-dimensional temporal structure of dependence of the Markov chain is generalized to a two-dimensional spatial dependence in the Ising model. This approximates ferromagnetism and consists of binary variables representing magnetic dipole moments. According to the model, a variable in a given location is conditionally independent from all other variables, which are not in neighboring locations, given the neighbors.

As will be shown, graphical log-linear models generalize the conditional independence mentioned last, to abstract neighborhood concepts described by graphs, where neighbor means being linked by an edge. The conditional independence above means that

$$A \perp\!\!\!\perp V \setminus nb(A) \setminus \{A\} \,|\, nb(A),$$

where $nb(A)$ is the set of neighbors of A. This is called the local Markov property and is meaningful with respect to a given graph.

Log-linear models based on graphical generating classes are said, in certain fields of applications, to have a Gibbs structure with respect to a graph. This means that the maximal interactions are the cliques of the graph.

The fundamental property, which makes graphical log-linear models so useful, may be formulated as Markov with respect to a graph if and only if it has Gibbs structure with respect to the graph. Before this is formulated more precisely, a general variant of the Möbius inversion, stated for a special case as Theorem 10.13, is needed.

Proposition 11.2. *Let V be a set of categorical variables, such that for $A \in V$, the categories of A are $1, 2, \ldots, c_A$. For every subset of variables $W \subseteq V$, let $(v)^W = ((v)_W, 1_{V \setminus W})$, that is, v and $(v)^W$ are identical on the indices of the variables in W, and $(v)^W$ has indices 1 on the other variables. Let f be a real function on the contingency table formed by the ranges of the variables V, and define*

$$f_W(v) = f((v)^W)$$

Further, let

$$g_W(v) = \sum_{U \subseteq W} (-1)^{|W \setminus U|} f((v)^U).$$

Then, for every $W \neq \emptyset$,

$$g_W(v) = 0, \textit{ if there is a } A \in W, \textit{ such that } (v)_A = 1.$$

Further,

$$f(v) = \sum_{W \subseteq V} g_W(v).$$

Proof. To see the first claim, note that for every $U \subseteq W$, such that $A \in U$, there is another subset of W, $U \setminus \{A\}$, which does not contain A, and half of the sets $U \subseteq W$ is of the first type, and half is of the second type. Because the index of A in v is 1,

$$f((v)^U) = f((v)^{U \setminus \{A\}}),$$

and these cancel out each other, because out of $|W \setminus (U \setminus \{A\})|$ and $|W \setminus U|$, one is even, the other one is odd.

To see the second claim, consider

$$\sum_{W \subseteq V} g_W(v) = \sum_{W \subseteq V} \sum_{U \subseteq W} (-1)^{|W \setminus U|} f((v)^U). \tag{11.4}$$

In this sum, $f((v)^U)$ for a fixed $U \subseteq V$ occurs for each W of which U is a subset, and the number of such sets W is $2^{|V \setminus U|}$. In fact, for each $E \subseteq V \setminus U$, the corresponding

W can be written as $W = U \cup E$, and $|W \setminus U| = |E|$. Thus, the contribution of $f((v)^U)$ in (11.4) is

$$\sum_{E \subseteq V \setminus U} (-1)^{|E|} f((v)^U). \tag{11.5}$$

For $U \neq V$, the value of (11.5) is zero. This is because half of the sets E contain an odd and half of them contain an even number of variables.

If $U = V$, the complement contains the empty set only with cardinality zero, and the coefficient is 1, which completes the proof, because

$$f_V(v) = f((v)^V) = f(v).$$

□

To see the relevance of this result, consider a 2 × 2 contingency table formed by the variables A and B, and define

$$f(i,j) = \log p(i,j).$$

Then,

$$g_\emptyset(i,j) = \log p(1,1),$$

$$g_A(1,j) = 0, \; g_A(2,j) = \log p(2,1) - \log p(1,1),$$

$$g_B(i,1) = 0, \; g_B(i,2) = \log p(1,2) - \log p(1,1),$$

$$g_{AB}(1,1) = 0, g_{AB}(1,2) = 0, g_{AB}(2,1) = 0,$$

$$g_{AB}(2,2) = \log p(2,2) - \log p(1,2) - \log p(2,1) + \log p(2,2)$$

Further,

$$\log p(1,1) = g_\emptyset(1,1) + g_A(1,1) + g_B(1,1) + g_{AB}(1,1),$$

$$\log p(1,2) = g_\emptyset(1,2) + g_A(1,2) + g_B(1,2) + g_{AB}(1,2),$$

$$\log p(2,1) = g_\emptyset(2,1) + g_A(2,1) + g_B(2,1) + g_{AB}(2,1),$$

$$\log p(2,2) = g_\emptyset(2,2) + g_A(2,2) + g_B(2,2) + g_{AB}(2,2).$$

Or, omitting the zero terms,

$$\log p(1,1) = \log p(1,1),$$

$$\log p(1,2) = \log p(1,1) + (\log p(1,2) - \log p(1,1)),$$

$$\log p(2,1) = \log p(1,1) + (\log p(2,1) - \log p(1,1)),$$

$$\log p(2,2) = \log p(1,1) + (\log p(2,1) - \log p(1,1)) + (\log p(1,2) - \log p(1,1))$$
$$+ (\log p(2,2) - \log p(1,2) - \log p(2,1) + \log p(1,1))$$

This is clearly a parameterization of the joint distribution, called the corner parameterization,[9] which is different from the balanced parameterization developed in Chap. 10.

[9] Or effect parameterization, see Sect. 10.2.

In general, let $f(v) = \log p(v)$. Then denote $f_W(v) = f((v)^W)$ as f_w^W. These are log-linear parameters, and let the corresponding multiplicative parameters be $\gamma_w^W = \exp(f_w^W)$. The two sets of parameters, the βs and the γs, can be converted into each other directly, that is, without first calculating the cell probabilities; see [53]. The most important thing, however, is that the definitions of the log-linear model in terms of the βs and in terms of the γs coincide. More precisely,

Theorem 11.5. *For a probability distribution* $\mathbf{p} \in \mathscr{P}$ *and ascending class* $\mathscr{A}$*,*

$$\beta_w^W = 1, \textit{ for all } W \in \mathscr{A} \textit{ and index } w, \tag{11.6}$$

if and only if

$$\gamma_w^W = 1, \textit{ for all } W \in \mathscr{A} \textit{ and index } w. \tag{11.7}$$

Proof. Theorem 10.15 states that a $\mathbf{p} \in \mathscr{P}$ is in $LL(\mathscr{A})$, that is,

$$COR(W|W' = w') = 1, \text{ for all } W \in \mathscr{A} \text{ and index } w' \text{ of the variables } W' = 2^V \setminus W, \tag{11.8}$$

if and only if (11.6) holds. Therefore, it is enough to show that (11.8) and (11.7) are equivalent.

The conditional odds ratios in (11.8) are meant to include all local odds ratios, if the variables are not all binary. Without loss of generality, the proof is formulated for the subtable containing the indices 1 and 2 for all variables.

Let W be a minimal element in $\mathscr{A}$, and consider the conditional odds ratio in (11.8) for a fixed w'. Write the probabilities in the conditional odds ratio, as the products of all γ_u^U terms, for all $U \subseteq V$. Relative to W, the $U \neq \emptyset$ subsets can be one of the following types:

$U \cap W \neq W$ or

$U = W$ or

$U \supset W$.

In the first case, U may contain variables from W, but not all of them, and possibly other variables. The indices of the latter variables are constant, and the other indices, if any, appear in all possible combinations in the numerator and in the denominator of the conditional odds ratio; thus these γ_u^U terms cancel.

In the second case, $\gamma_w^W = 1$, unless all indices of W are equal to 2, which happens in one cell.

In the third case, $\gamma_u^U = 1$, if not all indices of W are equal to 2, and still may be 1, if there are indices equal to 1 among the conditioning variables.

Thus, if all conditioning variables are 1, the conditional odds ratio is equal to γ_2^W, or its reciprocal, and thus is equal to 1, if and only if $\gamma_2^W = 1$, that is, if and only if $\gamma_w^W = 1$.

Let now A be a variable in $V \setminus W$. If $A = 2$ and the other conditioning variables are all 1, then the conditional odds ratio is $\gamma_2^W \gamma_2^A$ or its reciprocal and can be equal to 1, if and only if $\gamma_2^A = 1$, because $\gamma_w^W = 1$ holds in this case. Thus, $\gamma_a^A = 1$.

A similar argument can be used for the other variables in the complement of W, then for pairs of variables in the complement, etc., till the result is obtained for $W = V$. □

Now the main result about graphical log-linear models may be proven.

Theorem 11.6. *For every $\mathbf{p} \in \mathscr{P}$, $\mathbf{p}$ is Markov with respect to a graph with the variables as nodes, if and only if it has Gibbs structure with respect to that graph.*

Proof. Let first $\mathbf{p}$ be Gibbs; thus

$$p(v) = \prod_{W \in \mathscr{D}} \beta_w^W((v)_w),$$

and for any $A \in V$ and $v = (a, a')$,

$$p(a,a') = \left(\prod_{W \in \mathscr{D}, A \in W} \beta_w^W((a,a')_w) \right) \left(\prod_{W \in \mathscr{D}, A \notin W} \beta_w^W((a,a')_w) \right).$$

In the expression above, the second term does not depend on a. The conditional probability

$$p(A = a | V \setminus A = a')$$

can be written as

$$\frac{p(a,a')}{p(a')} = \frac{\left(\prod_{W \in \mathscr{D}, A \in W} \beta_w^W((a,a')_w)\right) \left(\prod_{W \in \mathscr{D}, A \notin W} \beta_w^W((a,a')_w)\right)}{\sum_a \left(\prod_{W \in \mathscr{D}, A \in W} \beta_w^W((a,a')_w)\right) \left(\prod_{W \in \mathscr{D}, A \notin W} \beta_w^W((a,a')_w)\right)} =$$

$$\frac{\prod_{W \in \mathscr{D}, A \in W} \beta_w^W((a,a')_w)}{\sum_a \prod_{W \in \mathscr{D}, A \in W} \beta_w^W((a,a')_w)}.$$

The last expression depends on a and out of the indices of the other variables, only on those which are neighbors of A in the graph. Thus, the conditional distribution of A, given its neighbors and non-neighbors, depends on the neighbors only, proving the desired conditional independence.

To see the converse, let $\mathbf{p}$ be Markov with respect to a graph. Apply Proposition 11.2 with $f(v) = \log p(v)$, to obtain that

$$\log p(v) = \sum_{W \subseteq V} g_W(v) = \sum_{W \subseteq V} \sum_{U \subseteq W} (-1)^{|W \setminus U|} f_U((v)^U).$$

Choose $W \subseteq V$, $W \notin \mathscr{D}$. One needs to see that

$$g_W(v) = 0, \text{ for all } v.$$

Further, choose variables A and B, such that both are in W but are not connected in the graph. Such variables exist, because the W is not complete in the graph.

Let $U \subseteq W$ be any subset such that $A \notin U$. Then $U \cup \{A\} \subseteq W$. As the cardinalities of these sets differ by 1, their joint contribution to $g_W(v)$ is

$$\pm \left(f((v)^{U \cup \{A\}}) - f((v)^U) \right). \tag{11.9}$$

By definition, this quantity is equal to

$$\pm \log \frac{p((v)^{U \cup \{A\}})}{p((v)^U)}.$$

The probabilities in the numerator and denominator differ only in the index of the variable A. It is $(v)_A = a$ in the numerator and 1 in the denominator. Using this, the ratio of the two probabilities is also equal to the ratio of two conditional probabilities, conditioned on the same event:

$$\pm \log \frac{p(a|((v)_{A'})^U)}{p(1|((v)_{A'})^U)}.$$

Because $\mathbf{p}$ is Markov, both of the conditional probabilities remain unchanged, if $(v)_B$ is changed to 1 (if it is not already 1) and so does the quantity and (11.9).

When U goes through all subsets of W, which do not contain A, the pairs U and $U \cup \{A\}$ go through all subsets of W. Therefore, the value of $g_W(v)$ does not change, if the index of B is set to 1. But by Proposition 11.2, $g_W(v)$ is zero in that case. This implies that $g_W(v) = 0$ for all v. □

This result says that the graphical log-linear models are exactly those which can be interpreted with the Markov property. This property is very useful, and it has found many applications and generalizations in current statistical practice. An authoritative account of the theory of graphical log-linear models and some of their generalizations is [48].

11.4 Canonical Representation of Log-Linear Models

The material in this section uses graphical models and forbidden configurations to characterize any log-linear model and is largely based on [74].

Every generating class $\mathscr{D}$ defines a graph, in which two nodes are connected if and only if the two variables appear in the same interaction in $\mathscr{D}$. Denote this graph by $\mathscr{G}(\mathscr{D})$. Two log-linear models generated by the descending classes, $\mathscr{D}_1$ and $\mathscr{D}_2$, are said to be graphical equivalent, if

$$\mathscr{G}(\mathscr{D}_1) = \mathscr{G}(\mathscr{D}_2).$$

For example, AB, BC, ACE and ABC, ACE are graphical equivalent. It is easy to see that graphical equivalence is an equivalence relation, that is, it is reflexive, symmetric, and transitive; thus, its equivalence classes define a partition of all log-linear models for a given set of variables. Then it follows readily that

Proposition 11.3. *Every graphical equivalence class contains exactly one graphical model, namely, the one defined by the cliques of the common graph.* □

Out of the two generating classes above, ABC, ACE is graphical, and AB, BC, ACE is not.

If one considers $LL(\mathscr{D}) \subseteq \mathscr{P}$ a set of distributions, and $\mathscr{G}(\mathscr{D})$ also as a generating class, which contains all subsets of V, which are complete in $\mathscr{G}(\mathscr{D})$, then

$$LL(\mathscr{G}(\mathscr{D})) \supseteq LL(\mathscr{D}),$$

because $\mathscr{D}$ may not contain interactions present in $\mathscr{G}(\mathscr{D})$. Note that while on the right-hand side the descending class containing all interactions is used to define a log-linear model, on the left-hand side, a graph is given, and its cliques are the maximal interactions. For example, the model defined by ABC, ACE also contains the ABC interaction, not allowed by AB, BC, ACE, which is graphical equivalent to it.

This suggests that non-graphical log-linear models may be described by applying additional restrictions on graphical log-linear models. The additional restrictions will be formulated as intersections of several models. Let $\mathscr{D}_1$ and $\mathscr{D}_2$ be two generating classes. Let the ascending class complement to $\mathscr{D}$ be denoted as $\mathscr{A}(\mathscr{D})$. The intersection of two log-linear models contains those distributions which have both representations according to (10.23). Then, it is easy to see that

$$LL(\mathscr{D}_1) \cap LL(\mathscr{D}_2) = LL(\mathscr{A}(\mathscr{D}_1)) \cap LL(\mathscr{A}(\mathscr{D}_2)) = LL(\mathscr{A}(\mathscr{D}_1) \cup \mathscr{A}(\mathscr{D}_2)).$$

This leads to the following result.

Theorem 11.7. *For two generating classes $\mathscr{D}_1$ and $\mathscr{D}_2$,*

$$LL(\mathscr{D}_1) \cap LL(\mathscr{D}_2) = LL(\mathscr{D}_1 \cap \mathscr{D}_2).$$

Proof. In view of the formula before the theorem, it is enough to see that

$$\mathscr{A}(\mathscr{D}_1) \cup \mathscr{A}(\mathscr{D}_2) = \mathscr{A}(\mathscr{D}_1 \cap \mathscr{D}_2).$$

Indeed, $\mathscr{A}(\mathscr{D}) = 2^V \setminus \mathscr{D}$, and the claim is implied by the properties of complements.
□

The next theorem sheds light on how graphical and non-graphical models differ.

Theorem 11.8. *If there is a $W \subseteq V$, which is an interaction in $LL(\mathscr{G}(\mathscr{D}))$, but W is not an interaction in $LL(\mathscr{D})$, then $\mathscr{D}$ contains a forbidden configuration.*

Proof. The situation described cannot occur for $|W| = 2$, and if it does occur for $|W| = 3$, then W is a forbidden configuration.

The rest of the proof is by induction on $|W|$. If the claim is true for all W with fewer than $k+1$ variables, let W with $|W| = k+1$ be an interaction in $\mathscr{G}(\mathscr{D})$ but not in $\mathscr{D}$. Either all subsets of W with cardinality k are interactions in $\mathscr{D}$, in which case W is a forbidden configuration, or there is a subset of cardinality k, which is not an interaction in $\mathscr{D}$. But as W is an interaction in $\mathscr{G}(\mathscr{D})$, so is this subset, and by the induction assumption, it contains a forbidden configuration. □

Obviously, if the interactions in $LL(\mathscr{D})$ and in $LL(\mathscr{G}(\mathscr{D}))$ are the same, then $\mathscr{D}$ is graphical. The following theorem shows how any generating class $\mathscr{D}$ can be characterized by $\mathscr{G}(\mathscr{D})$ and by its forbidden configurations.

Theorem 11.9. *Let $\mathscr{D}$ be a generating class and let $F_1,\ldots,F_k$ its forbidden configurations, if any. Let the forbidden configuration F_i be*

$$F_i = (F_{i,1}, F_{i,2}, \ldots, F_{i,|F_i|}).$$

Define the generating classes

$$\mathscr{F}_i = \mathscr{D}(V \setminus F_{i,1}, V \setminus F_{i,2}, \ldots, V \setminus F_{i,|F_i|}),$$

for $i = 1,\ldots,k$, where $\mathscr{D}(W)$ means the descending class generated by W.

Then,

$$LL(\mathscr{D}) = LL(\mathscr{G}(\mathscr{D})) \cap \left(\cap_{i=1}^{k} LL(\mathscr{F}_i)\right). \tag{11.10}$$

Proof. Let $W \in \mathscr{D}$ be an interaction in $LL(\mathscr{D})$. For any of the forbidden configurations F_i, $W \cap F_i = F_i$ is not possible, because in this case W would contain F_i, but a forbidden configuration cannot be a subset of an interaction. Thus, $W \cap F_i$ is a proper subset of F_i, and there is a variable, say A_i, which is in F_i but not in W. Therefore, $W \subseteq V \setminus A_i$; thus, every interaction in $\mathscr{D}$ is also an interaction in $\mathscr{F}_i$; thus

$$LL(\mathscr{D}) \subseteq LL(\mathscr{F}_i),\ i = 1,\ldots,k.$$

Further,

$$LL(\mathscr{D}) \subseteq LL(\mathscr{G}(\mathscr{D}))$$

and, thus, the left-hand side of (11.10) is contained in the right-hand side.

To see the converse, consider a $\mathbf{p} \in \mathscr{P}$, which is contained in the right-hand side of (11.10). Then, $\mathbf{p} \in LL(\mathscr{G}(\mathscr{D}))$ and has a representation as the product of multiplicative parameters belonging to the interactions in $\mathscr{G}(\mathscr{D})$. Let W be any subset, which is an interaction in $\mathscr{G}(\mathscr{D})$, but not in $\mathscr{D}$. If there is no such subset, the proof is complete. If there is one, it will be shown in the rest of the proof that the parameter associated with W can be omitted from the multiplicative representation.

In this case, by Theorem 11.8, $\mathscr{D}$ has a forbidden configuration. Let this be F_i, and, as seen in the proof of Theorem 11.8, $F_i \subseteq W$. As $\mathbf{p}$ belongs to the right-hand side of (11.10),

$$\mathbf{p} \in LL(\mathscr{G}(\mathscr{D})) \cap LL(\mathscr{F}_i),$$

and does have a multiplicative representation according to the generating class $\mathscr{F}_i$. But $F_i \notin \mathscr{F}_i$ by definition; thus, $W \notin \mathscr{F}_i$, so there is a multiplicative representation without the multiplicative parameter belonging to W. □

As an example, consider the generating class with maximal interactions AB, AC, BC, and AD. This generating class is not graphical, as ABC is a forbidden configuration. The graphical model in its graphical equivalence class has maximal interactions ABC and AD. The theorem says that

$$LL(AB,AC,BC,AD) = LL(ABC,AD) \cap LL(ABD,ACD,BCD).$$

The meaning of this and the insight it provides into the structure of the original model will be discussed after one more result is given. This concerns the structure of graphical models.

Theorem 11.10. *Let $\mathscr{D}$ be graphical, and let $\mathscr{C}_1,\ldots,\mathscr{C}_l$ be the cliques of the complement graph*[10] *of $\mathscr{G}(\mathscr{D})$. For every $\mathscr{C}_j$, define a generating class as follows:*

$$\mathscr{D}_j = \mathscr{D}(\{A\} \cup \overline{\mathscr{C}_j} : A \in \mathscr{C}_j),$$

that is, the descending class generated by the set consisting of the complement of $\mathscr{C}_j$ and one variable from $\mathscr{C}_j$. Then

$$LL(\mathscr{D}) = \cap_{j=1}^{l} LL(\mathscr{D}_j). \tag{11.11}$$

Proof. The cliques of the complement graphs are maximal empty subgraphs of $\mathscr{G}(\mathscr{D})$, so the structure of the interactions in $\mathscr{D}_j$ is such that each consists of the complement of a maximal empty subgraph and one variable from the maximal empty subgraph.

It is enough to show, on the one hand, that if two variables, say A and B, are not contained in any interaction of $\mathscr{D}$, then there is a j that they are not contained in any interaction of $\mathscr{D}_j$. And on the other hand, if two variables, say A and B, are not contained together in any of the interactions appearing in at least one of the models on the right-hand side, then they do not appear together in any of the interactions in $\mathscr{D}$.

The first claim is true, because as A and B are not connected in $\mathscr{D}$, they are an empty subgraph of $\mathscr{G}(\mathscr{D})$. Let $\mathscr{C}_j$ be the maximal empty subgraph containing A and B. Then $\mathscr{D}_j$ does not contain A and B in the same interaction.

To see the second claim, note that if there exists a $\mathscr{D}_j$, such that A and B are not contained together in any of its interactions, then they are in $\mathscr{C}_j$, which is empty in $\mathscr{D}$, so they are not connected in $\mathscr{D}$. □

The last two theorems can be combined to give the canonical representation of every log-linear model.

Theorem 11.11. *For any generating class $\mathscr{D}$,*

$$LL(\mathscr{D}) = \left(\cap_{j=1}^{l} LL(\mathscr{D}_j)\right) \cap \left(\cap_{i=1}^{k} LL(\mathscr{F}_i)\right),$$

where $\mathscr{D}_j$ and $\mathscr{F}_i$ are defined as follows.

Let $\mathscr{G}(\mathscr{D})$ be the graphical model in the graphical equivalence class of $\mathscr{D}$ and let $\mathscr{C}_1,\ldots,\mathscr{C}_l$ be the cliques of its complement graph. For every $\mathscr{C}_j$, define a generating class as

$$\mathscr{D}_j = \mathscr{D}(\{A\} \cup \overline{\mathscr{C}_j} : A \in \mathscr{C}_j).$$

[10] The complement graph has the same nodes as the original, and two nodes are connected if and only if they are not connected in the original graph.

Further, let $F_1, \ldots, F_k$ be the forbidden configurations of $\mathscr{D}$, and let the forbidden configuration F_i be

$$F_i = (F_{i,1}, F_{i,2}, \ldots, F_{i,|F_i|}).$$

Then define

$$\mathscr{F}_i = \mathscr{D}(V \setminus F_{i,1}, V \setminus F_{i,2}, \ldots, V \setminus F_{i,|F_i|}),$$

for $i = 1, \ldots, k$, where $\mathscr{D}(W)$ means the descending class generated by W.

The theorem gives a representation of any log-linear model using only two kinds of log-linear models. The graphical model in the graphical equivalence class is characterized as intersection of models, each stating that a certain group of variables has joint independence, if conditioned on all other variables.

For example, for the generating class $\mathscr{D} = (AB, BC, CD)$, the cliques of the complement of $\mathscr{G}(\mathscr{D})$ are AD, AC, BD. These are clearly empty in $\mathscr{G}(\mathscr{D})$, and if a third variable was added, they would not be empty anymore. There are three models, the intersection of which is the original model. These are defined by the following generating classes: (ABC, BCD), (ABD, CBD), and (ABC, ACD). These are all conditional independence models, implying

$$A \perp\!\!\!\perp D|BC,\ A \perp\!\!\!\perp C|BD,\ B \perp\!\!\!\perp D|AC,$$

respectively. As $\mathscr{D}$ is graphical, there are no forbidden configurations.

Another example is the log-linear model with $\mathscr{D} = (AB, AC, AD, BC, BD, CD)$. Then, $\mathscr{G}(\mathscr{D})$ is a full graph, so the graphical model in the equivalence class does not contain any restrictions; therefore, there are no $\mathscr{D}_j$ generating classes. On the other hand, the model itself is not graphical, and its forbidden configurations are ABC, ABD, ACD, BCD, that is, all possible triplets. Then the maximal elements of the $\mathscr{F}_i$ are

$$(BCD, ACD, ABD),\ (BCD, ACD, ABC),\ (BCD, ABD, ABC),\ (ACD, ABD, ABC).$$

These all say that conditioned on the fourth variable, there is no second-order interaction among three variables. For example, the second one says that conditioned on C, there is not highest-order association among the variables ABD.

Thus, every log-linear model is the intersection of several models, which all belong to either one of two types of log-linear models. One type is the joint independence of some of the variables, given the other variables, and the other one has no highest-order interaction among a group of variables, given all other variables.

11.5 Things to Do

1. Explain why the overall odds in a log-linear model and the marginal odds from the same distribution are different in general.

2. In the midterm test example in Sect. 11.1, if the interaction term in those who majored in statistics and studied for at least 7 hours combination is 0.8, then how much is it in the combination of statistics major but did study for less than 7 hours?
3. Check that the multiplicative parameters given in Tables 11.4, 11.5, 11.6, 11.7, 11.8, 11.9, 11.10 multiply to 1 as stated in (10.22).
4. Obtain the Chile data set. Use R or other software to fit the log-linear models in Table 11.3.
5. Continue the previous exercise by generating the estimated data using the multiplicative parameters and compare them to the estimates from the software.
6. Using the Chile data set, complete the analysis of the regression type problem with the response variable having four categories.
7. Check how model fit changes if the highest-order interaction among the explanatory variables is not allowed in the models for the vote data.
8. Test the fit of the three log-linear models obtained from the last one in Table 11.13 by removing one of the first-order interactions.
9. Explain how may it happen that allowing more interactions increases model fit, but not to the extent expected based on the reduction of the number of degrees of freedom.
10. Is $(AB, BC, ABCD, DE, EFG, DGH)$ decomposable?
11. Prove that the definition of decomposability applies both to generating classes containing the maximal interactions only and to generating classes containing the entire descending class and that decomposability does not depend on the type of the generating class to which the definition is applied.
12. Prove that if a variable appears only in one of the maximal interactions, then whether or not the generating class is decomposable is not affected if this variable is removed.
13. Prove that if a new variable is added to one of the maximal interactions, it does not affect whether the generating class is decomposable.
14. Prove that if two variables always do or do not appear together in the maximal interactions, replacing them by a single variable (combining them) does not affect whether or not the generating class is decomposable.
15. Design an algorithm to decide whether a generating class is decomposable, irrespective of the order in which it is given.
16. Draw the graph of the model (11.3).
17. Prove the claim made after (11.5).
18. Develop the corner parameterization for a $3 \times 2 \times 4$ contingency table.
19. Find a conversion procedure between corner and balanced log-linear parameters.
20. Show that all values of the balanced log-linear parameter belonging to a subset of variables are zero, if and only if all values of the corner log-linear parameter belonging to the same subset are zero.
21. Prove that graphical equivalence is an equivalence relation.
22. Find the canonical representation of LL(AB, AC, BC).

Chapter 12
Log-Linear Models: Estimation

Abstract Maximum likelihood estimation of log-linear models is considered as a special case of maximum likelihood estimation in exponential families, with the mixed parameterization used in the definition of log-linear models playing a central role: the canonical parameters on the ascending class are specified by the model, while the marginal distributions on the descending class are taken from the observed data in the maximum likelihood estimates. Then, the main tool of computing maximum likelihood estimates, the Iterative Proportional Fitting Procedure, is described, and its convergence is proved.

Before the convergence of the procedure to find maximum likelihood estimates under log-linear models is proved, the main properties of maximum likelihood estimates are discussed.

12.1 Maximum Likelihood Estimation

The theory of maximum likelihood estimation of log-linear models builds largely on the results about the maximum likelihood estimation in exponential families given in Sect. 7.2.

For all distributions in $\mathscr{P}$, (10.21) was that

$$\log p(v) = \sum_{W \subseteq V} \lambda^{W}_{(v)_W}.$$

This is clearly an exponential family representation, with the canonical statistics being indicators of $(v)_W$, for all choices of $W \in \mathscr{D}$ and combinations of indices w of W. For example, in a 3×2, $A \times B$ table, with $\mathscr{D} = \{\emptyset, A, B\}$, one may set up the canonical statistics as follows:

$t_1 = 1$ in every cell
$t_2 = 1$ in cells $(1, j)$

$t_3 = 1$ in cells $(2, j)$
$t_4 = 1$ in cells $(3, j)$
$t_5 = 1$ in cells $(i, 1)$
$t_6 = 1$ in cells $(i, 2)$

These canonical statistics are not linearly independent. One choice for describing the same exponential family by linearly independent canonical statistics in this example is the following:

$t_1 = 1$ in every cell
$t_2 = 1$ in cells $(2, j)$
$t_3 = 1$ in cells $(3, j)$
$t_4 = 1$ in cells $(i, 2)$

These canonical statistics are linearly independent, and they lead to the corner parameterization (introduced in Sect. 11.3) of the distributions in the independence model. The general strategy to define log-linear models in a minimal representation as exponential families is based on the same idea.

Theorems 10.15 and 11.5 together state that a distribution $\mathbf{p} \in \mathscr{P}$ is in $LL(\mathscr{A})$, if and only if all the corner log-linear parameters defined in the paragraph before Theorem 11.5 are zero on all $W \in \mathscr{A}$. A representation of $\mathscr{P}$ as an exponential family, using the corner parameters, is obtained, if the design matrix has, for every $\emptyset \neq W \subseteq V$

$$\prod_{A \in W} (c_A - 1)$$

columns. (For binary tables, this leads to 2^V columns in total.) Each of the columns is associated with indices w of W, such that none of the indices in w is equal to 1. The column associated with a particular w^* is the indicator of those cells v, for which

$$(v)_W = w^*,$$

and the column associated with $\emptyset \subseteq V$ contains 1s.

For the case of the 3×2 independence, $\mathbf{p}$ has six components, which is the number of rows of the design matrix. The six components in lexicographic order are

$$(p(1,1), p(1,2), p(2,1), p(2,2), p(3,1), p(3,2))'.$$

In the design matrix, the columns are ordered as $\emptyset$, $A = 2$, $A = 3$, $B = 2$.

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{pmatrix} \tag{12.1}$$

and, accordingly, the corner log-linear parameters are arranged as

$$\mathbf{f} = (f^{\emptyset}, f_2^A, f_3^A, f_2^B)'.$$

With these notations,

$$\log \mathbf{p} = \mathbf{Tf}.$$

The columns of the design matrix are linearly independent, not only in the example but also generally, so one has a minimal representation of the log-linear model as regular exponential family.[1]

Then, Theorem 7.6 directly translates into the following result about maximum likelihood estimates under log-linear models.

Theorem 12.1. *Let $LL(\mathscr{D})$ be a log-linear model for the variables V, and let $\boldsymbol{X}$ be the observed frequencies on the contingency table formed by the variables V, under multinomial or Poisson sampling. Then, there may be at most 1 such $\boldsymbol{p}$ that*

$$\boldsymbol{p} \ \in LL(\mathscr{D}) \text{ and}$$

$$p(w,+) = \frac{1}{n}x(w,+) \text{ for all } W \in \mathscr{D} \text{ and indices } w \text{ of } W, \tag{12.2}$$

where $p(w,+)$ means the marginal probability of $W = w$ and n is the sample size of the multinomial distribution or the observed total of the Poisson distribution.

Further, if such a $\boldsymbol{p}$ exists, it is the MLE for all distributions for which (12.2) holds, under the model $LL(\mathscr{D})$. □

Obviously, it is sufficient to require (12.2) for the maximal elements in $\mathscr{D}$. In addition to the characterization of the MLE in the previous theorem, the next result provides sufficient conditions for the existence of the MLE.

Theorem 12.2. *Under multinomial or Poisson sampling, if there exists a $\boldsymbol{p} \in \mathscr{P}$, such that (12.2) holds, then there exists an MLE in the regular exponential family, which is uniquely characterized by the following mixed parameters:*

$$COR(A|A' = a') = 1 \text{ for all } A \in \mathscr{A} \text{ and all } a' \text{ index of } V \setminus A,$$

$$\frac{1}{n}x(d) \text{ marginal distributions on } D \in \mathscr{D}.$$

Proof. The characterization is implied by Theorem 12.1 and the existence by Theorem 10.2. For the latter, the convergence of the Iterative Proportional Fitting Procedure, to be proved in the next section, is needed. □

When the observed frequencies are all positive, $\mathbf{x}/n$ is a distribution in $\mathscr{P}$ with (12.2). If the observed frequencies are not all positive, such a distribution may or may not exist. Perhaps surprisingly, even if the sufficient statistics are all positive, such a distribution may not exist. For example, consider a $2 \times 2 \times 2$ table and the model of no second-order association. Let the observed frequencies be positive in all cells, except for the cells $(1,1,1)$ and $(2,2,2)$, which are empty in the data. Then, the sufficient statistics are the three two-way marginal distributions (see Proposition

[1] Because $\mathbf{p} \in \mathscr{P}$, the domain of the parameters $\mathbf{f}$ is open.

7.2), which are positive and so are all the marginal distributions of the subsets of variables in $\mathscr{D}$. In spite of this, in this case, there exists no positive $\mathbf{p}$, for which (12.2) holds. Indeed, if with $p(i,j,k) = x(i,j,k)/n$ and if $p(1,1,1) = 0$ is increased to be positive, then, as $p(1,1,+)$ has to be kept constant, $p(1,1,2)$ has to be reduced. But this means that, in order to keep $p(1,+,2)$ unchanged, $p(1,2,2)$ has to be increased, and finally, to keep $p(+,2,2)$ constant, $p(2,2,2)$ needs to be reduced. That would make the originally zero probability negative, and $\mathbf{p}$ would not be positive. This example is from [36]. For a current review and deeper results concerning the issue of the existence of the MLE in log-linear models when there are zero observed frequencies, see [26].

One structural issue related to log-linear models is the number of degrees of freedom associated with the model. This issue comes up, most often, in relation to certain approaches to testing model fit. Very often, the Pearson or likelihood ratio statistics (see Sect. 5.2) are calculated to compare the observed distribution $\mathbf{x}/n$ to the maximum likelihood estimate under the model. As a rough approximation to the real distributions of these statistics, which for finite sample sizes do depend on the size of the problem and on the true distribution, their asymptotic distribution is used, which, in addition to the size, only depends on whether or not the true distribution is in the model. The size referred to in the previous sentence, which depends on the number of variables, the numbers of their categories, and the model, is characterized by the degrees of freedom.

In the simplest classical case, one wishes to test independence in an $I \times J$ table. Here, as described in Sect. 5.4.3, the number of parameters to estimate from the data, without any model assumed, is $IJ-1$. The model of independence restricts $(I-1)(J-1)$ of these (the local or spanning cell odds ratios), and one has $IJ-1-IJ+I+J-1 = (I-1)+(J-1)$ parameters to estimate. Sometimes, $(I-1)(J-1)$ is called the number of degrees of freedom of the model (more precisely, the number of degrees of freedom used by the model) and $(I-1)+(J-1)$ the residual number of degrees of freedom). The number of degrees of freedom associated with the testing problem (and with the asymptotic chi-squared distributions) is $(I-1)(J-1)$. In the 3×2 example discussed above, the total number of free parameters is 5, the number of model degrees of freedom is 2, and the number of columns of the design matrix (12.1), which is the number of parameters to be estimated, is 4. But this latter concept requires a bit more care. Although there are four parameters in $\mathbf{f}$ to be estimated, they have to be estimated in a way that the sum of the estimated $p(v)$ values is 1. Thus, 3 may be seen as free to be estimated and 1 may be seen as implied. Thus, the total number of degrees of freedom is 5, the number of model parameters is 2, and the number of free parameters is 3. The latter number is also obtained by the formula $(I-1)+(J-1)$.

12.2 Iterative Proportional Fitting

This section discusses the Iterative Proportional Fitting Procedure (or Iterative Scaling Procedure). This algorithm is used widely in statistics, from small area estimation in official statistics (see, e.g., [3]) to poststratification in survey research (see, e.g., [55]). In this section, it will be presented as an algorithm to obtain maximum likelihood estimates under log-linear models.

IPFP is based on the simple observation that in a 2 × 2 table, if both entries in the first row are multiplied by the same number, say α, and both entries in the second row are multiplied by the same number, say, β, the odds ratio remains unchanged, because the multipliers cancel out. Indeed, for the transformed table

$$\frac{(\alpha p(1,1))(\beta p(2,2))}{(\alpha p(1,2))(\beta p(2,1))} = \frac{p(1,1)p(2,2)}{p(1,2)p(2,1)}.$$

Moreover, with an appropriate choice of the multipliers α and β, the row marginal of the new table can be made equal to any marginal distribution. If the new marginal distribution is to be $(q(1,+),q(2,+))'$, then

$$\alpha = \frac{q(1,+)}{p(1,+)},\ \beta = \frac{q(2,+)}{p(2,+)}$$

changes the row marginal probabilities to the desired values.

This is true more generally.

Theorem 12.3. *Let $U \subset W \subseteq V$, and for every index u of the variables U, let $q(u)$ be arbitrary positive numbers. Then, the value of $COR(W|W' = w')$ is the same, whether computed from a probability distribution $\mathbf{p} \in \mathscr{P}$ or from the quantities $q(u)p(u,u')$, where w' and u' mean indices for the complements of W and of U, respectively.*

Proof. Without loss of generality, assume that 1 is one of the indices for all variables in the subtable, in which the conditional odds ratio is computed. The quantity $q(u)p(u,u')$, for fixed u, appears in the numerator or in the denominator of $COR(W|W' = w')$, depending on the parity of the number of indices 1 in the joint index of $W \setminus U \neq \emptyset$. The number of indices $w \cap u'$ with an even and with an odd number of 1s is the same; thus $q(u)$ appears in the numerator and in the denominator the same number of times and, thus, cancels. □

The purpose of the IPFP is to create a distribution, the existence of which was stated in Theorem 10.2. Such a distribution combines the conditional odds ratios of one distribution on an ascending class, with the marginal distributions of another distribution on the complement descending class. The reason for using the marginals of an existing distribution is that marginal distributions given on a descending class of subsets are not necessarily compatible. To illustrate this, consider the three two-way distributions presented in Table 12.1. Can these be the $A \times B$, $A \times C$, and $B \times C$

marginals of an $A \times B \times C$ three-dimensional distribution? The three two-way distributions are weakly compatible, which means that the one-way marginal distributions computed from any of the two-way distributions for any of these three variables are identical. For example, the marginal distribution of A can be derived from the first and second tables, and from both, one gets a uniform distribution for A. In this example, all three one-way marginals are uniform, but this is not an essential feature. In spite of being weakly compatible, these distributions are not strongly compatible in the sense that there is no three-way distribution of which these are the two-way marginals. Indeed, if such a distribution existed, $p(1,1,+) = 0.05$ would imply that $p(1,1,1) < 0.05$ and $p(1,1,2) < 0.05$, $p(2,2,+) = 0.05$ would imply that $p(2,2,1) < 0.05$ and $p(2,2,2) < 0.05$, $p(1,+,1) = 0.05$ would imply that $p(1,2,1) < 0.05$, $p(2,+2) = 0.05$ would imply that $p(2,1,2) < 0.05$, $p(+,1,1) = 0.05$ would imply that $p(2,1,1) < 0.05$, and $p(+,2,2) = 0.05$ would imply that $p(1,2,2) < 0.05$. That is, all 8 probabilities would be less than 0.05, and their sum could not be equal to 1.

Table 12.1 Three 2 × 2 distributions which are not strongly compatible

0.05	0.45	0.05	0.45	0.05	0.45
0.45	0.05	0.45	0.05	0.45	0.05

The fact that for a $2 \times 2 \times 2$ table weak compatibility of the given marginal distributions on all three two-way marginal tables does not imply their strong compatibility is related to the model of no second-order interaction not being decomposable. Weak compatibility of marginal distributions on a decomposable class of marginals implies their strong compatibility.

Theorem 12.4. *Let $\mathscr{C} = \{C_1, \dots, C_k\}$ be the maximal interactions of a decomposable class of subsets of the variables V. Let $\mathbf{q}$ be a probability distribution on the contingency table formed by the ranges of the variables in V. Then,*

$$p(v) = \frac{\prod_{j=1}^{k} q((v)_{C_j}) q((v)_{\cup_{i=1}^{j-1} C_i}))}{\prod_{i=1}^{j} q((v)_{C_j \cap C_{i(j)}})} \tag{12.3}$$

is a probability distribution, such that

$$p((v)_{C_j}) = q((v)_{C_j}), \textit{ for all } v \textit{ and } j = 1, \dots, k.$$

Further, if $\mathbf{q}$ is the observed distribution from a multinomial or Poisson sampling procedure, then $\mathbf{p}$ is the maximum likelihood estimate in $LL(\mathscr{C})$.

Proof. When the $q(i,j,+)$ and $q(+,j,k)$ marginal distributions are extended in a conditionally independent way to

$$q(i,j,k) = \frac{q(i,j,+)q(+,j,k)}{q(+,j,+)},$$

the original two-way marginals are preserved. Then, repeated application of this fact yields the first claim.

The first claim of this theorem is one of the conditions of Theorem 12.2. The second condition of Theorem 12.2 is implied by noting the fact that for a uniform distribution, all conditional odds ratios are equal to 1, then considering (12.3) as a series of modification for $j = 1, \ldots, k$ of a uniform distribution as described in this fact, and then applying the latter theorem.

Then, Theorem 12.2 yields the second claim. □

The formula in (12.3) is the closed form MLE for decomposable log-linear models. In the case of log-linear models based on nondecomposable generating classes, such a closed form estimate does not exist, and the IPFP needs to be applied.

Let V be given categorical variables, and let $\mathbf{q}$ and $\mathbf{r} \in \mathscr{P}$ be two distributions[2] on the contingency table formed by the ranges of the variables in V. Further, let $\mathscr{D}$ be a descending class of subsets of the variables V, with maximal elements $W_1, \ldots, W_k$. Then, the IPFP is defined as follows. First let

$$p_0(v) = r(v),$$

for all joint indices v. For $i = 1, 2, \ldots$, define

$$p_i(v) = p_{i-1}(v)\frac{q((v)_{W_j}, +)}{p_{i-1}((v)_{W_j}, +)}, \tag{12.4}$$

where $i = lk + j$. The algorithm starts with the distribution $\mathbf{r}$ and cyclically adjusts the current W_j marginals to the W_j marginals of $\mathbf{q}$. Indeed, it is immediate that

$$p_{lk+j}((v)_{W_j}, +) = q((v)_{W_j}, +),$$

for all l and j and v, and Theorem 12.3 implies that with $\mathscr{A} = 2^V \setminus \mathscr{D}$, for all $A \in \mathscr{A}$,

$$COR_{\mathbf{p}_i}(A|A' = a') = COR_{\mathbf{r}}(A|A' = a'),$$

for all i and a'. The summary of the previous results is the following:

Theorem 12.5. *If the IPFP converges, then the limit is a probability distribution, with the properties described in Theorem 10.2.* □

When, in particular, $\mathbf{q}$ is an observed distribution, then, with reference to Theorem 12.2, one has

Theorem 12.6. *If $\boldsymbol{q} = \boldsymbol{x}/n$, where $\boldsymbol{X}$ has a multinomial or Poisson distribution, and the IPFP, if started with $\boldsymbol{r} = \boldsymbol{u}$, the uniform distribution, converges to a $\boldsymbol{p} \in \mathscr{P}$, then $\boldsymbol{p}$ is the maximum likelihood estimate of the true distribution in $LL(\mathscr{D})$.*

In order to prove the convergence of the IPFP, the first step is to consider (12.4), as a projection of $\mathbf{p}_{i-1}$ into a the linear family

[2] The distribution $\mathbf{q}$ does not have to be strictly positive.

$$L(W_i, \mathbf{q}) = L(W_{lk+j}, \mathbf{q}) = \{\mathbf{p} : p((v)_{W_j}, +) = q((v)_{W_j}, +)\}.$$

This is the family of distributions which have the same W_j marginal distribution as $\mathbf{q}$. The family is linear, because a linear combination of two distributions in it, if it is a probability distribution, is also in the family. While $\mathbf{p}_{i-1}$ is not necessarily in $L(W_i, \mathbf{q})$, $\mathbf{p}_i \in L(W_i, \mathbf{q})$. The sense, in which $\mathbf{p}_i$ is the projection of $\mathbf{p}_{i-1}$ in $L(W_i, \mathbf{q})$, is the information divergence (I-divergence) or Kullback-Leibler divergence. The information divergence between two probability distributions $\mathbf{p}$ and $\mathbf{q}$ is defined as

$$I(\mathbf{p}||\mathbf{q}) = \sum_v p(v) \log \frac{p(v)}{q(v)},$$

with the understanding that $\log 0 = -\infty$, $\log(a/0) = \infty$, for positive a, and $0 \cdot \pm\infty = 0$. The choice of the base of the logarithm is important in information theory, and it is usually 2, but it is irrelevant for our purposes.

The Kullback-Leibler divergence is a measure of the expected amount of information one has, when taking observations from $\mathbf{p}$, to distinguish it from $\mathbf{q}$ (see [17]). The relevant meaning here is a measure of deviation. $I(\mathbf{p}||\mathbf{q})$ is asymmetric, and if $\mathbf{p} = \mathbf{q}$, then $I(\mathbf{p}||\mathbf{q}) = 0$.

Proposition 12.1. *The following holds for the Kullback-Leibler divergence:*

$$I(\boldsymbol{p}||\boldsymbol{q}) \geq 0,$$

and

$$I(\boldsymbol{p}||\boldsymbol{q}) = 0 \textit{ if and only if } \boldsymbol{p} = \boldsymbol{q}.$$

Proof. As $\log_2 a = \ln a / \ln 2$, it is sufficient to prove the claim for natural logarithms. For positive x, $\ln x \leq x - 1$, with equality only at $x = 1$. This is seen most easily by noting that the derivative of $\ln x$ is $1/x$ and of $x - 1$ it is 1. Thus, before the two functions are equal at $x = 1$, $\ln x$ grows faster and after it $x - 1$ does.

Let I_p be the set of cells where $\mathbf{p}$ is not zero, called the support of $\mathbf{p}$. As these do not contribute to $I(\mathbf{p}||\mathbf{q}$ by the convention made after the definition, they may be omitted from consideration. Then

$$-I(\mathbf{p}||\mathbf{q}) = \sum_{v \in I_p} p(v) \ln \frac{q(v)}{p(v)} \leq \sum_{v \in I_p} p(v) \left(\frac{q(v)}{p(v)} - 1 \right)$$

$$= \sum_{v \in I_p} q(v) - \sum_{v \in I_p} p(v) \leq 0.$$

For equality to hold throughout the above derivation, the supports I_p and I_q have to be related as $I_q \supseteq I_p$, so if $\mathbf{p}$ is positive, $\mathbf{q}$ is positive too. Further, $q(v)/p(v) = 1$ should hold for all $v \in I_q$; otherwise a strict inequality would occur for that summand. This also implies the first condition and completes the proof.

When $\mathbf{p}$ is an observed distribution, and

$$I(\mathbf{p}||\mathbf{q}) = \sum_v p(v) \log p(v) - \sum_v p(v) \log q(v)$$

is minimized in $\mathbf{q}$ belonging to a statistical model, then the minimizer is the maximum likelihood estimate under multinomial sampling. In the sequel, however, minimization will be done in the first argument. More precisely, let $\mathscr{H}$ be a set of probability distributions. Then

$$\arg \min_{\mathbf{p} \in \mathscr{H}} I(\mathbf{p}||\mathbf{q})$$

is called the I-projection of $\mathbf{q}$ in $\mathscr{H}$. For the proof of convergence of the IPFP, the following results are needed, which are stated without proof.

Theorem 12.7.

(i) For an arbitrary distribution in $\mathscr{P}$, its I-projection in a family of distributions defined by linear restrictions always exists.

(ii) In particular, for the distributions in the updating step (12.4) of the IPFP, $\boldsymbol{p}_i$ is the I-projection of $\boldsymbol{p}_{i-1}$ in $L(W_i, \boldsymbol{q})$.

(iii) Further, for any distribution $\boldsymbol{s} \in L(W_i, \boldsymbol{q})$,

$$I(\boldsymbol{s}||\boldsymbol{p}_{i-1}) = I(\boldsymbol{s}||\boldsymbol{p}_i) + I(\boldsymbol{p}_i||\boldsymbol{p}_{i-1}). \tag{12.5}$$

(iv) If $\mathscr{F} \subseteq \mathscr{G}$ are families of distributions defined by linear constraints, then the I-projection of a distribution in $\mathscr{F}$ is the same as the I-projection in $\mathscr{F}$ of its I-projection in $\mathscr{G}$.

Proof. For a proof, see [18]. □

The fact in (12.5) is an important property of I-projections, showing the information divergence behaving similarly to the square of the Euclidean distance. The deviation, as measured by the I-divergence between any distribution in the linear family $\mathbf{s} \in L(W_i, \mathbf{q})$, and the probability distribution $\mathbf{p}_{i-1}$ outside of the linear family, can be decomposed into the sum of the deviation from $\mathbf{s}$ to the I-projection of $\mathbf{p}_{i-1}$, $\mathbf{p}_i$ and the deviation between the latter distribution and $\mathbf{p}_{i-1}$.

Now we are ready to prove the convergence of the IPFP. The proof given here is essentially the one given by Csiszár [18].

Theorem 12.8. *Suppose there exists a distribution $\boldsymbol{s} \in \mathscr{P}$ with the same W_j marginal distributions, $j = 1, \ldots, k$, as $\boldsymbol{q}$. Then, the sequence of distributions generated by (12.4) converges to a probability distribution $\boldsymbol{p}$, which is the I-projection of* $\mathbf{r}$ *on*

$$L(\mathscr{D}, \mathbf{q}) = \cap_{j=1}^{k} L(W_j, \mathbf{q}).$$

Proof. Obviously, $\mathbf{s} \in L(W_j, \mathbf{q})$, for all j. Then, (12.5) applies and gives

$$I(\mathbf{s}||\mathbf{p}_{i-1}) = I(\mathbf{s}||\mathbf{p}_i) + I(\mathbf{p}_i||\mathbf{p}_{i-1}).$$

This relationship, written for $i = 1$, is

$$I(\mathbf{s}||\mathbf{p}_0) = I(\mathbf{s}||\mathbf{p}_1) + I(\mathbf{p}_1||\mathbf{p}_0).$$

and for $i = 2$, it is

$$I(\mathbf{s}||\mathbf{p}_1) = I(\mathbf{s}||\mathbf{p}_2) + I(\mathbf{p}_2||\mathbf{p}_1).$$

From these equations, one obtains that

$$I(\mathbf{s}||\mathbf{p}_0) = I(\mathbf{s}||\mathbf{p}_2) + I(\mathbf{p}_2||\mathbf{p}_1) + I(\mathbf{p}_1||\mathbf{p}_0),$$

or

$$I(\mathbf{s}||\mathbf{p}_0) = I(\mathbf{s}||\mathbf{p}_2) + \sum_{m=1}^{2} I(\mathbf{p}_m||\mathbf{p}_{m-1}).$$

One sees by induction that for every $i = 1, 2, \ldots$,

$$I(\mathbf{s}||\mathbf{p}_0) = I(\mathbf{s}||\mathbf{p}_i) + \sum_{m=1}^{i} I(\mathbf{p}_m||\mathbf{p}_{m-1}).$$

Thus, for every $i = 1, 2, \ldots$,

$$0 \leq \sum_{m=1}^{i} I(\mathbf{p}_m||\mathbf{p}_{m-1}) \leq I(\mathbf{s}||\mathbf{p}_0),$$

then also

$$0 \leq \sum_{m=1}^{\infty} I(\mathbf{p}_m||\mathbf{p}_{m-1}) \leq I(\mathbf{s}||\mathbf{p}_0).$$

As $\mathbf{p}_0 = \mathbf{r} \in \mathscr{P}$, the right-hand side of the formula above is finite, so one can conclude that

$$\sum_{m=1}^{\infty} I(\mathbf{p}_m||\mathbf{p}_{m-1})$$

converges. Therefore,

$$I(\mathbf{p}_m||\mathbf{p}_{m-1}) \to 0,$$

and by Pinsker's inequality (see [19]),

$$|\mathbf{p}_m - \mathbf{p}_{m-1}| \to 0, \tag{12.6}$$

where $|\mathbf{c}|$ is the component of $\mathbf{c}$ with the highest absolute value, called the total variation norm.

Because the sequence $(\mathbf{p}_i)$ consists of probability distributions, which are finite vectors, there exists a subsequence $(\mathbf{p}_{i_m})$, which converges to a probability distribution $\mathbf{p}'$. It will be shown in the rest of the proof that $\mathbf{p}' \in L(\mathscr{D}, \mathbf{q})$ and that $(\mathbf{p}_i) \to \mathbf{p}'$; thus the latter is the desired $\mathbf{p}$.

Indeed, if $(\mathbf{p}_{i_m}) \to \mathbf{p}'$, then

$$(\mathbf{p}_{i_m+1}) \to \mathbf{p}',$$

because of (12.6), and, similarly,

$$(\mathbf{p}_{i_m+2}) \to \mathbf{p}',$$

and so on till

$$(\mathbf{p}_{i_m+k}) \to \mathbf{p}',$$

implying that $\mathbf{p}' \in L(\mathscr{D},\mathbf{q})$, because each of these subsequences visits infinitely many times different ones out of the k $L(W_j,\mathbf{q})$ sets. If $(\mathbf{p}_{i_m})$ visits $L(W_j,\mathbf{q})$ infinitely many times for some j, then $(\mathbf{p}_{i_m+1})$ visits $L(W_{j+1},\mathbf{q})$ infinitely many times, etc.

Repeated application of parts (ii) and (iv) of Theorem 12.7 shows that the desired $\mathbf{p}$ is the I-projection of $\mathbf{p}_i$ in $L(\mathscr{D},\mathbf{q})$, as the latter set is smaller than any of the $L(W_j,\mathbf{q})$ sets, for $i = 1,2,\ldots$. Therefore, as implied by part (iii) of Theorem 12.7, for any $\mathbf{u} \in L(\mathscr{D},\mathbf{q})$,

$$I(\mathbf{u}||\mathbf{p}_i) = I(\mathbf{u}||\mathbf{p}) + I(\mathbf{p}||\mathbf{p}_i).$$

In particular, for $\mathbf{u} = \mathbf{p}'$, one has

$$I(\mathbf{p}'||\mathbf{p}_i) = I(\mathbf{p}'||\mathbf{p}) + I(\mathbf{p}||\mathbf{p}_i).$$

The left-hand side of the above expression converges to zero by continuity, which implies that $I(\mathbf{p}'||\mathbf{p})$=0, and this means (see Proposition 12.1) that $\mathbf{p}' = \mathbf{p}$. □

The most important implication of this result is that when the condition of the existence of the MLE in a log-linear model (see Theorem 12.2) holds, then a limit of the IPFP, initiated with the uniform distribution, say $\mathbf{p}$, exists, it possesses the unique characteristics defining the MLE and, thus, is the MLE itself (see Theorem 12.6).

As Theorem 12.8 shows, the MLE in a log-linear model $LL(\mathscr{D})$ is the I-projection of the uniform distribution in the linear family $L(\mathscr{D},\mathbf{q})$. Thus, the MLE may be interpreted as the most uniform (i.e., closest to the uniform in the sense of I-divergence) distribution which has the observed marginal distributions on the descending class $\mathscr{D}$. Because of the variation independence of the two components of a mixed parameterization of all distributions on the contingency table, the exponential family $LL(\mathscr{D})$ is parameterized by the marginal distributions of the subsets of variables in $\mathscr{D}$. The unique distribution in

$$LL(\mathscr{D}) \cap L(\mathscr{D},\mathbf{q})$$

is the MLE of any distribution which is in $L(\mathscr{D},\mathbf{q})$ in the model $LL(\mathscr{D})$ (see Theorem 7.6) and thus is the most uniform distribution in $LL(\mathscr{D})$. Thus, the log-linear model $LL(\mathscr{D})$ is the collection of the most uniform distributions subject to their $\mathscr{D}$ marginal distributions being prescribed to any strongly compatible set of values. This is yet another interpretation of log-linear models. Further, as being closest to uniform may be interpreted as containing the least amount of additional information, log-linear models may be considered as families of distributions containing no information, additional to their marginal distributions on the defining descending classes. This is the information theoretical definition of log-linear models. The same models, in many applications, are called maximum entropy models, meaning distributions having maximum entropy (most uniform distribution) subject to linear assumptions.

Maximum likelihood estimates of the log-linear or of the multiplicative parameters may be obtained by calculating them from the maximum likelihood estimate of the true distribution (see Proposition 4.2). However, as shown in a more general setting in [42], they also may be obtained as the products of the multipliers applied during the course of the IPFP.

The IPFP may be used to obtain maximum likelihood estimates for models more general than the log-linear models discussed so far. In some cases, log-linear models assuming values for the conditional odds ratios different from 1 may be relevant. For example, the conditional odds ratio may be obtained from census data, and data from a sample in a later year may be used to test the hypothesis that the conditional odds ratio remained unchanged. Such models were described in [71] and [84].

The IPFP has many generalizations. For a review and comparison in a general setting, see [42].

12.3 Things to Do

1. Develop the design matrix for the corner parameterization defined in Sect. 12.1 for a $3 \times 3 \times 2$ no second-order association model.
2. Prove that the design matrix defined for the corner parameterization in Sect. 12.1 has linearly independent columns.
3. Can the example given after Theorem 12.2 be generalized to $2 \times 2 \times 2 \times 2$ tables?
4. Prove that in Theorem 12.5, the limit is a probability distribution.
5. Use a continuity argument to show that in Theorem 12.5, the limit preserves the conditional odds ratios of the ascending class $\mathscr{A}$.
6. Prove that if the IPFP is applied to a decomposable model in the decomposable ordering of the maximal interactions, it converges after as many steps were performed, as the number of maximal interactions and the resulting distribution is the same as the one given in (12.3).
7. Complete the previous result by illustrating with an example that if the order is not the decomposable one, then convergence may not occur after as many adjustments, as the number of maximal interactions.

Chapter 13
What's Next?

Abstract Readers who have followed through with studying the material presented in the book are now ready to read the literature leading to current research in the field. This brief chapter contains summaries of and references to interesting and useful topics.

Undirected Graphical Models The theory of graphical log-linear models is much more developed than the treatment given in Sect. 11.3. In addition to the local Markov property, there are two more interpretations of a graphical log-linear model. The pairwise Markov property is that if two variables are not connected in the graph, then they are conditionally independent given all other variables. The global Markov property is that if two groups of variables are separated by a third group of variables, then the joint distributions in the two first groups are conditionally independent, given the third group. Separation here means that every path that connects a variable from the first group with one in the second goes through one variable in the third group. When a probability distribution is strictly positive, the three Markov properties are equivalent but not in general. A very good reference for the theory of graphical log-linear models is [48].

Directed Graphical Models The term graphical model does not only mean graphical log-linear models but also models which are related to other types of graphs and are not necessarily log-linear in the sense discussed in this book. The most important such class is directed graphical models. These models are intended to describe effects among variables. One group of such models is associated with directed acyclic graphs or DAGs. In such a graph, the nodes are the variables entering the analysis, and any two nodes may be linked by an arrow. Two arrows between the same two nodes or an arrow pointing to the same node where it started are not allowed. To be acyclic, the existence of a directed circle is excluded. A directed circle is a path along the arrows which ends where it started. Figure 8.1 contains all directed graphs between three variables. These are all DAGs, with the exception of [s] and [z]. The model associated with a DAG is also defined by Markov properties.

To describe the directed Markov property, one needs some concepts related to DAGs. The descendants of a node A, say $de(A)$, are those nodes, into which a directed path leads from A. The nondescendants, $nd(A)$, are the nodes into which no directed path leads from A. The parents, $pa(A)$, are those nodes from which an arrow points to A. For a DAG, $pa(A) \subseteq nd(A)$, because if a parent was a descendant, then one would have a directed cycle. The directed Markov property is defined as

$$A \perp\!\!\!\perp nd(A) \setminus pa(A) | pa(A),$$

see [48]. The intuitive interpretation associated with this definition is that A may affect those variables which are its descendants, but as it cannot affect its nondescendants, if these do not affect A, then they will be conditionally independent of A, given the variables affecting A directly. Effect here may or may not mean causal effect; see Chap. 8.

Much of the applications of DAG models rely on the fact that there exists an ordering of the variables, called well numbering, such that all variables are preceded by their respective parents; see [49]. The possibility of such an ordering is very attractive from the perspective of causal interpretations of directed graphical models.

Directed graphical models are applied in practice under different names, including Bayesian networks, Bayes nets, and belief networks; see [47]. A great part of the applications appears in the machine learning field, where the goal may be to learn the graph or to learn the parameters. Directed graphical models play an important role in expert systems, in risk management, or, more generally, in artificial intelligence, too; see, e.g., [63].

Chain Graph Models Chain graph models are further generalizations, which combine some of the properties of directed and undirected graphical models. A chain graph consists of components. The nodes (variables) within each component have an undirected graph among them, representing associations. In one interpretation, variables in the same component are contemporaneous. In the graph itself, this means edges or missing edges between variables in the same component. There are arrows or missing arrows between nodes (variables) in different components, with the restriction, that all arrows between variables in two different components go from the variable in one component, to the variable in the other component. These arrows represent effects, perhaps according to a time order or for other reasons. Different authors associate somewhat different Markov properties with such a chain graph; see [24] and [31]. The existence of different interpretations shows clearly that the Markov property is not an implication of the graph, which only illustrates the model, rather the Markov property is associated with the graph based on different data analytical needs and the experiences of statisticians.

To define the various Markov properties considered in the literature, the concepts of parents and nondescendants have to be extended to apply to groups of variables, including components. For groups W, all within one component, other than a whole component, the parents $pa(W)$ consists of variables, which are parents of any of the variables in W. The neighbors $nb(W)$ consists of variables which are connected to any of the variables in W by an edge. For a whole component, $\mathscr{K}$, the parents $PA(\mathscr{K})$ are all the variables in those components, from which at least one arrow goes to $\mathscr{K}$. The nondescendants of a component, $ND(\mathscr{K})$ is the union of those components, into which no semi-directed path leads from any variable in $\mathscr{K}$. A semi-directed path is a path through arrows and edges, so that the direction of the arrows is observed. With these notations, the components of the different chain graph Markov properties may be written as (see [82])

P1: For all components $\mathscr{K}$, $\quad \mathscr{K} \perp\!\!\!\perp ND(\mathscr{K}) \setminus PA(\mathscr{K}) | PA(\mathscr{K})$
P2a: For all $\mathscr{K}$ and $W \subseteq \mathscr{K}$, $\quad W \perp\!\!\!\perp \mathscr{K} \setminus W \setminus nb(W) | PA(\mathscr{K}) \cup nb(W)$
P2b: For all $\mathscr{K}$ and $W \subseteq \mathscr{K}$, $\quad W \perp\!\!\!\perp \mathscr{K} \setminus W \setminus nb(W) | PA(\mathscr{K})$
P3a: For all $\mathscr{K}$ and $W \subseteq \mathscr{K}$, $\quad W \perp\!\!\!\perp PA(\mathscr{K}) \setminus pa(W) | pa(W) \cup nb(W)$
P3b: For all $\mathscr{K}$ and $W \subseteq \mathscr{K}$, $\quad W \perp\!\!\!\perp PA(\mathscr{K}) \setminus pa(W) | pa(W)$

P1 is the directed Markov property applied to the components. P2 means that a subset of variables W in a component is conditionally independent from the other variables in the component to which they are not linked, given the parents of the component. The two variants differ in whether or not the condition also involves the neighbors of W. P3 is that W is conditionally independent from the parents of the other variables in the component, given its own parents. Again, the two variants differ in whether or not one also has to condition on the neighbors of W.

Variants of the chain graph Markov property are obtained by assuming P1, one out of P2a and P2b, and one out of P3a and P3b. All four types have been considered in the literature. The combination P1-P2a-P3a is called the Lauritzen-Wermuth-Frydenberg block-recursive Markov property; see [31, 50]. The combination P1-P2a-P3b is called the Andersson-Madigan-Perlman block-recursive Markov property; see [4].

There are many further types of graphs, which are used to represent various independence structures, among others bidirected graphs to illustrate marginal independences; see [66].

Marginal Models Directed graphical models are not log-linear models. Although they are defined by conditional independence assumptions, like graphical log-linear models, the conditional independences defining the former do not always involve all variables, as is the case with the latter class of models. Indeed, the pairwise Markov property says two variables if not connected are conditionally independent given all other variables, and such a conditional independence statement involves all variables. On the other hand, the directed Markov property only involves a variable and its nondescendant. That is, the conditional independences defining a directed graphical model are specified on marginals of the type $\{A\} \cup nd(A)$. Marginal models, as introduced by Bergsma and Rudas [10], provide a general framework to define, parameterize, and interpret such models.

Marginal models are applicable in many other statistical problems, including repeated measurements or observations being dependent in some other ways, panel data, data fusion, and others; see [9] and [80].

A marginal model, or marginal log-linear model, is defined by a marginal log-linear parameterization, just like log-linear models were defined using a special kind of parameterization of all distributions on the contingency table. This marginal log-linear parameterization uses marginal log-linear parameters, which are log-linear parameters defined in marginal tables. To be more specific, consider a sequence of subsets of the variables, $W_1, W_2, \dots W_k = V$, such that if $i < j$, then $W_j \nsubseteq W_i$. The choice of the marginals on which the marginal log-linear parameterization is built is determined by the problem to which the marginal log-linear model is to be applied. Then, in every W_j marginal distribution, the log-linear parameters are considered for all subsets of variables, which are contained in W_j and are not contained in any W_i, for $i < j$. These marginal log-linear parameters are also a parameterization of all distributions on the full table, and marginal log-linear models are obtained by setting some of these equal to zero.

Every log-linear model is also a marginal log-linear model, when only one marginal, $W_1 = V$, is used. But directed graphical models are also marginal models, with the marginals being of the form $\{A\} \cup nd(A)$, for the variables involved. For a detailed discussion, see [60].

Certain types of chain graph models are also marginal log-linear models (see [82]), and this fact implies, among others, that the maximum likelihood estimates under such models have standard asymptotic properties.

Bidirected graph models, representing marginal independences, are also marginal models; see [56].

Relational Models Relational models may be seen as coordinate free generalizations of log-linear models in the following sense. In standard multivariate statistics, the sample space is the Cartesian product of the ranges of the variables involved. When the setup is categorical, this Cartesian product is the contingency table. The effects which are allowed in a log-linear models are effects associated with subsets of variables. The collection of the cells of the table which project into the same marginal cell are called cylinder sets, and every effect, that is, every multiplier is present on such a cylinder set. For example, the multiplicative parameter $\beta_{i^*}^{A}$, for a fixed i^*, is present in every cell of the form (i^*, j), for every j. The collection of these is a cylinder set.

Relational models do not assume such a product structure. The sample space has finitely many cells, which may or may not be structured. Quite often, the sample space is a proper subset of a Cartesian product. For example, [40] reported an experiment, where swimming crabs were caught in traps containing as bait either sugarcane or fish or a combination of sugarcane and fish. This design, from the perspective of the variables sugarcane present or not and fish present or not, is not a 2×2 contingency table, which is the Cartesian product of the ranges of the two variables, because not all four combinations were observed, only three of them. The structure of the sample space is shown in Table 13.1. For several other examples of similar designs and also process produced data of such a structure, and also for examples where it is logically impossible to have all combinations of the individual categories, see [42] and [44].

Table 13.1 The design in [40], an incomplete Cartesian product

	Fish	
Sugarcane	Yes	No
Yes	Observed	Observed
No	Observed	Not observed

If one was interested in finding out, whether the effects of the two bait types, sugarcane and fish, are independent on the crabs, then the situation would depend a lot on whether crabs could or could not be caught in an empty trap. If they could be, then the not observed cell is not empty in the population, just was left out from the design, and one could assume that a positive fraction of the population is there. But this hypothetical fraction could always be determined in such a way, that the odds ratio in the 2×2 contingency table is 1. Indeed, if the observed frequencies are X_{11}, X_{12}, X_{21}, then assuming $X_{12}X_{21}/X_{11}$ for the hypothetical value makes the odds ratio equal to 1. Thus, in this case independence cannot be tested from the data. If, on the other hand, one believed that no crabs could be caught with a trap with no bait, then assuming independence of the effects of the two bait types makes no sense, because independence would assign a positive probability to every cell, including the one which is assumed to be empty in the population.

Thus, with the given design, independence either cannot be tested or does not make sense at all. An alternative definition of independence, proposed by Aitchinson and Silvey in [2], however, may be applied. This assumes that

$$p_{12}p_{21} = p_{11}, \tag{13.1}$$

that is, the probability that a crab ends up in the trap with both bait types is the product of the probabilities that it ends up in the traps with either one of the baits. Or, the effect of the two baits together is the product of the individual effects. This model is an exponential family, but it has no normalizing constant or overall effect. It cannot have one, because if a parameter present in all three cells was introduced, the number of parameter would be three, which is the number of cells, and the model would not be restrictive anymore and could not be tested.

As illustrated in the previous example, another important way in which relational models generalize log-linear models is that the overall effect is not necessarily present. If the overall effect is not present, normalization cannot be achieved by dividing by the total, rather it is a restriction on the parameter space. Such exponential families are called curved. When the overall effect is not present, the odds ratio specification of the model (parallel to defining a log-linear model by saying that the conditional odds ratios on an ascending class are equal to 1) always contains a nonhomogeneous odds ratio. In such an odds ratio, the number of cells in the numerator is different from the number of cells in the denominator. For example, the Aitchison-Silvey independence in (13.1) may be defined by setting the value of a nonhomogeneous odds ratio equal to 1:

$$\frac{p_{12}p_{21}}{p_{11}} = 1.$$

The final component of generalization is that the effects need not be associated with cylinder sets. In the Aitchison-Silvey independence, there is an effect, say α, associated with sugarcane being present, and an effect, say β associated with fish being present. The parametric version of the model is illustrated in Table 13.2. The parameters α and β apply to cylinder sets in this case.

Table 13.2 Aitchison-Silvey independence

	Fish	
Sugarcane	Yes	No
Yes	$\alpha\beta$	α
No	β	Not observed

To illustrate a model, where the parameters are not associated with cylinder sets, assume a researcher had the hypothesis that it is irrelevant, what is the bait, what only counts is if there is only one bait or two baits in the trap. That model may be parameterized as shown in Table 13.3. In this model, α is the effect of two baits together, and β is the relative multiplicative effect of having one bait only. Here, the effect β is not associated with cylinder sets, but α is an overall effect. The odds ratio representation is $p_{12}/p_{21} = 1$. This is a homogeneous odds ratio, because this model contains the overall effect.

Table 13.3 A model where only the number of baits matters

	Fish	
Sugarcane	Yes	No
Yes	α	$\alpha\beta$
No	$\alpha\beta$	Not observed

Relational models were introduced in [44], and they have many interesting properties. When the overall effect is not present, the equivalence of the multinomial and Poisson likelihoods (Theorems 4.8 and 12.2) does not hold. In fact, maximum likelihood estimates for relational models without the overall effect under Poisson sampling do not preserve the observed total and under multinomial sampling do not preserve the subset sums. The subset sums are the observed totals of those sets of cells where a particular parameter is present, e.g., in the model of Table 13.2, the subset of α is $\{(1,1),(1,2)\}$. These are the generalizations of the marginals of the interactions allowed in a log-linear model. To fit these models, one needs a generalized version of IPFP, described in [42], and testing also requires modified statistics; see [43].

Path Models A path model is a directed graphical model, where arrows represent direct effects and directed paths represent indirect effects. This idea assumes that there are no effects in the joint distribution, which could not be represented by an arrow or a sequence of arrows. This may be true for data with multivariate normal distribution (see Proposition 1.2), where effects are between pairs of variables only. But with categorical variables, there may be effects not attributable to a pair of variables; see Proposition 1.3 and [45]. Motivated by [33], [81] defined path models starting with a directed graphical model, and considering it as a marginal model, using the marginal log-linear parameterization. Then, the definition proceeds by setting all marginal log-linear parameters of all such interactions to zero, which pertain to more than two variables. Thus, only main effects and first-order interactions are allowed. The latter are interpreted as the effect of one variable on the other.

To illustrate a simple model, suppose the researcher thinks that out of three variables, C is response to A and to B, and A and B are independent. The graph illustrating this structure looks like $[o]$ in Fig. 8.1. To set up the marginal modeling framework, one needs the AB marginal to be able to impose the marginal independence of these variables. The marginal log-linear parameterization using the marginals AB and ABC is

$$\lambda_{\emptyset}^{AB}, \lambda_{A}^{AB}, \lambda_{B}^{AB}, \lambda_{AB}^{AB}, \lambda_{C}^{ABC}, \lambda_{AC}^{ABC}, \lambda_{BC}^{ABC}, \lambda_{ABC}^{ABC}.$$

The notation here is that the upper index shows the marginal, in which the log-linear parameter is computed, the lower index shows the effect, to which it pertains. The individual cell indices are not shown. These parameters provide a parameterization of all distributions on the three-way table. To impose the marginal independence of A and B, one needs to set

$$\lambda_{AB}^{AB} = 0.$$

The remaining parameters parameterize all three-way distributions, in which A and B are marginally independent. To impose the path model, one has to set the second-order interaction to zero

$$\lambda_{ABC}^{ABC} = 0.$$

The marginal log-linear parameters not set to zero are

$$\lambda_{\emptyset}^{AB}, \lambda_{A}^{AB}, \lambda_{B}^{AB}, \lambda_{C}^{ABC}, \lambda_{AC}^{ABC}, \lambda_{BC}^{ABC}.$$

In addition to the normalizing constant and the main effects of the three variables, one has

$$\lambda_{AC}^{ABC}, \lambda_{BC}^{ABC},$$

which may be seen as quantifying the individual effects of A and B on C. For more details and examples, see [60].

Model Testing This book has largely neglected issues related to model testing and, instead, has concentrated on structural issues. One of the reasons for this is that the quickly expanding machine learning-data mining approach, which does use the models presented here, does not emphasize model testing. Although this practice cannot be endorsed in general, when the data available for the analysis is the entire population, which is often the case with big or organic data, results of exploratory analyses are often believed in, without any attempt to test the findings. It is true in such cases that the traditional statistical testing approach does not apply.

When the now standard asymptotic testing is applied, the test statistic is most often a member of the power divergence family (see [64])

$$\frac{2}{\lambda(\lambda+1)} \sum_{v} X_v \left(\left(\frac{X_v}{np_v} \right)^{\lambda} - 1 \right),$$

where λ is a real parameter. This class includes the Pearson chi-squared statistic and the likelihood ratio statistic. The advantage of the asymptotic testing procedures is that while the distributions of the test statistics depend on the true population distribution, asymptotically, they only depend on whether or not the hypothesis holds. Detailed proofs can be found in [13] or in the generality of relational models in [43].

There is a large body of literature offering advice, whether or not in a given situation, the asymptotic distribution could be used to assess the actual values of the Pearson or of the likelihood ratio statistics. Such suggestions are usually based on simulations. The published results do diverge in many aspects and are not easy to summarize. In an earlier simulation study, [70], this author found that the distribution of the Pearson statistic converges to its limit somewhat faster than that of the likelihood ratio statistic does. For small tables and simple log-linear models, already two or three times higher sample sizes than the number of cells of the sample space produced acceptable results, when the Pearson statistic was used, while the likelihood ratio statistic required somewhat larger sample sizes for the actual critical values to get close enough to the asymptotic critical values. See also the comments at the end of Sect. 5.4.3.

When one does not wish to rely on the asymptotic distributions of test statistics, Monte Carlo simulations [67] may be applied to approximate a critical value to which the actual value of the test statistic may be compared. This procedure first estimates the true distribution based on the assumed model and the available data and then generates samples from the estimated distribution. Then, from each estimated distribution, the value of the test statistic is determined, by assuming the same model. Finally, the estimated values of the test statistic are used to determine an approximate critical value, to which the value of the test statistic based on the real data is compared.

An alternative procedure is related to the volume interpretation of the chi-squared statistic; see [23]. Here, every possible data set having the same sufficient statistic values as the actual data set has, where, of course, the sufficient statistics depend on the model of interest, is assumed to have the same probability of being observed. For each such data set, the value of the chi-squared statistic is determined, and the actual value is compared to this distribution. To approximate the relevant critical value, one needs to generate many data sets with fixed values of the sufficient statistics. This process of data generation, misleadingly in a statistical context, is often called sampling.

For example, in a 2×2 table to test independence this way, one needs to generate 2×2 frequency distributions, with the same one-way marginal distributions, as observed. This is not a difficult exercise, but for larger tables and more complex models, it is difficult to generate data, so that the assumed equiprobability holds. Often, this problem is described as a random walk on the space of all relevant data sets, where the points (data sets) visited form a Markov chain. This means, that where the walk goes from a point, does depend on this point, but not on the previous points visited. In this case, the procedure is called Markov chain Monte Carlo; see [8]. The steps in this procedure may be thought of as repeatedly adding or subtracting from the data in the given point, in such a way, that the values of the sufficient statistics remain unchanged. These incremental components are called a Markov basis for the problem. For example, in the case of 2 × 2 independence, the Markov basis consists of a single incremental step shown in Table 13.4.

Table 13.4 The Markov basis for 2 × 2 independence

+1	−1
−1	+1

A fully unrelated alternative to the above testing procedures is the mixture index of fit, proposed in [83]; see also [27]. This is formulated in a nonrestrictive framework, which assumes that a distribution from the model of interest may describe a part of the population and asks how large this part may be. The framework is nonrestrictive, because the relative size of the fraction, where a distribution from the model applies, may even be zero. Therefore, the true population distribution P is seen as a mixture:

$$P = (1 - \pi)G + \pi H,$$

where G belongs to the model of interest, and H is unspecified. The mixture index of (mis)fit π^* is defined as

$$\pi^* = \inf\{\pi : P = (1 - \pi)G + \pi H, G \in \mathscr{M}\},$$

where $\mathscr{M}$ is the model of interest. Then, $1 - \pi^*$ is the largest possible fraction of the population, where a distribution from $\mathscr{M}$ holds true. The larger is $1 - \pi^*$, the better is model fit. The level of model fit, in this approach, is defined as a parameter and may be estimated using the EM algorithm; see Sect. 7.1. Advantages include that the mixture index of fit also applies to situations, when one analyses population data, while the standard statistical testing procedures are limited to data from a sample. A further advantage is that the residual distribution H does have a direct meaning and it helps assessing the substantive relevance of the model to describe the population distribution; see [20]. In contrast, residuals in standard testing approaches have a somewhat dubious interpretation. When the hypothesis is rejected, they are calculated based on a model which was deemed not relevant, and when the hypothesis is not rejected, the residuals were deemed just a random fluctuation.

The mixture index of fit may be applied to any statistical hypothesis (see, e.g., [73]), and it also has applications in handling missing data, [75].

References

1. Agresti, A.: *Categorical Data Analysis*, 2nd ed. Wiley, New York (2002)
2. Aitchison, J., Silvey, S.D.: Maximum-likelihood estimation procedures and associated tests of significance. *Journal of the Royal Statistical Society, Ser. B*, **22**, 154–171 (1960)
3. Anderson, B.: Estimating small-area income deprivation: An iterative proportional fitting approach. In Tanton, R., Edwards, K. (ed.) *Spatial Microsimulation: A Reference Guide for Users*, pp. 49–67. Springer, New York (2012)
4. Andersson, S.A, Madigan, D., Perlman, M.: Alternative Markov properties for chain graphs. *Scandinavian Journal of Statistics*, **28**, 33–85 (2001)
5. Barndorff-Nielsen, O.E.: *Information and Exponential Families in Statistical Theory*. Wiley, New York (1978)
6. Bartolucci, F., Forcina, A.: Extended RC Association Models Allowing for Order Restrictions and Marginal Modeling. *Journal of the American Statistical Association*, **97**, 1192–1199 (2002)
7. Becker, M.P., Clogg, C.C.: An alysis of Sets of Two-way Contingency Tables Using Association Models. *Journal of the American Statistical Association*, 84, 142–151 (1989)
8. Berg, B.A.: *Markov Chain Monte Carlo Simulations and Their Statistical Analysis*. World Scientific (2004)
9. Bergsma, W, Croon, M., Hagenaars, J.A.: *Marginal Models For Dependent, Clustered and Longitudinal Categorical Data*. Springer, New York (2009)
10. Bergsma, W.P., Rudas, T.: Marginal models for categorical data. *Annals of Statistics*, **30**, 140–159 (2002)
11. Bertsekas, D.P.: *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, New York (1982)
12. Bickel, P.J., Hammel, E.A., O'Connell, J.W.: Sex bias in graduate admissions: Data from Berkeley, *Science* **187**, 398–404 (1975)
13. Bishop, Y.M.M., Fienberg, S.E., Holland, P.W.: *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Boston (1975)
14. Cameron, P.J.: *Combinatorics: Topics, Techniques, Algorithms*. Cambridge University Press (1994)
15. Clogg, C.C., Shihadeh, E.S.: *Statistical Models for Ordinal Variables*. Sage Publications, Thousand Oaks (1994)
16. Cohn, D.L.: *Measure Theory*. Birkhauser, Boston (1980)
17. Cover, T.M., Thomas, J.A.: *Elements of Information Theory, 2nd ed*. Wiley, New York (2006)
18. Csiszár, I.: I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, **3**, 146–158 (1975)
19. Csiszár, I., Körner, J.: *Information Theory: Coding Theorems for Discrete Memoryless Channels*. Cambridge University Press, (2011)

20. Clogg, C.C, Rudas, T., Matthews, S.: Analysis of model misfit, structure, and local structure in contingency tables using graphical displays based on the mixture index of fit. In Blajius, J., Greenacre, M. (ed.) *Visualization of Categorical Data*, pp. 425–439. Academic Press, New York (1997)
21. Curley, S.P., Browne, G.J.: Normative and Descriptive Analyses of Simpson's Paradox in Decision Making. *Organizational Behavior and Human Decision Process*, **84**, 308–333 (2001)
22. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Scociety (Ser B)*, Vol 39, 1–38 (1977)
23. Diaconis, P., Efron, B.: Testing for independence in a two-way table: new interpretations of the chi-square statistic. *Annals of Statistics*, **13**, 845–874 (1985)
24. Drton, M.: Discrete chain graph models. *Bernoulli* **15**, 736–753 (2009)
25. Edwards, D., Havranek, T.: A fast procedure for model search in contingency tables. *Biometrika*, **72**, 339–351 (1985)
26. Fienberg, S.E, Rinaldo, A.: Maximum likelihood estimation in log-linear models. *Annals of Statistics*, **40**, 996–1023 (2012)
27. Formann, A.K.: Latent class model diagnostics – a review and some proposals. *Computational Statistics and Data Analysis* **41**, 549–559 (2003)
28. Fox, J., Andersen, R.: Effect displays for multinomial and proportional-odds logit models. *Sociological Methodology* **36**, 225–255 (2006)
29. Fox, J., Weisberg, S.: An R Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion (2011)
30. Friedl, H, Hormann, S.: Frequentist Probability Theory. In Rudas, T. (ed.) *Handbook of Probability: Theory and Applications*, pp. 15–34. Sage Publications, Thousand Oaks (2008)
31. Frydenberg, M.: The chain graph Markov property. *Scandinavian Journal of Statistics* **17**, 333–353 (1990)
32. Greenacre, M.J.: *Theory and Applications of Correspondence Analysis*. Academic Press, London (1984)
33. Goodman, L.A.: The analysis of multidimensional contingency tables, when some variables are posterior to others: a modified path analysis approach. *Biometrika* **60**, 179–192 (1973)
34. Goodman, L.A.: Simple Models for the Analysis of Association in Cross- Classifications Having Ordered Categories. *Journal of the American Statistical Association* **74**, 37–52 (1979)
35. Goodman, L. A. & Kruskal, W.H.: Measures of association for cross classifications. *Journal of the American Statistical Association* **49**, 732–764 (1954)
36. Haberman, S.J.: *The Analysis of Frequency Data*. Univ. Chicago Press, Chicago, IL. (1974)
37. Haberman, S.J.: Iterative scaling procedures for log-linear models for frequency tables derived by indirect observation. *Proc. Amer. Statist. Assoc. (Statist. Comp. Sect., 1975)*, 45–50 (1976)
38. Hansen, M.H., Hurwitz, W.N., Madow, W.G.: *Sample Survey Methods and Theory, Volumes I and II*. Wiley, New York (1993)
39. Hosmer, D.W., Lemeshow, S.: *Applied Logistic Regression, 2nd ed*. Wiley, New York (2000)
40. Kawamura, G., Matsouka, T., Taijri, T., Nishida, M., Hayashi, M.: Effectiveness of a sugarcane-fish combination as bait in trapping swimming crabs. *Fisheries Research*, **22**, 155–160 (1995)
41. Kish, L.: *Survey Sampling*. Wiley, New York (1995)
42. Klimova, A., Rudas, T.: Iterative scaling in curved exponential families. *Scandinavian Journal of Statistics*, **42**, 832–847 (2015)
43. Klimova, A., Rudas, T.: Testing the fit of relational models. *arxiv:* 1612.02416v1 (2016)
44. Klimova, A., Rudas, T., Dobra, A.: Relational models for contingency tables. *Journal of Multivariate Statistics*, **104**, 159–173 (2012)
45. Klimova, A., Uhler, C., Rudas, T.: Faithfulness and learning of hypergraphs from discrete distributions. *Computational Statistics and Data Analysis*, **87**, 57–72 (2015)
46. Kopylov, I.: Subjective Probability. In Rudas, T. (ed.) *Handbook of Probability: Theory and Applications*, pp. 35–48. Sage Publications, Thousand Oaks (2008)
47. Koski, T, Noble, J.: *Bayesian Networks: An Introduction*. Wiley, New York (2009)

48. Lauritzen, S.L.: *Graphical Models*. Clarendon Press, Oxford (1996)
49. Lauritzen, S.L:, Dawid, A.P., Larsen, B.N., Leimer, H.-G.: Independence properties of directed Markov fields. *Networks*, **20**, 491–505 (1990)
50. Lauritzen, S.L, Wemuth, N.: Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, **17**, 31–57 (1989)
51. Lazarsfeld P.F., Henry, N.W.: *Latent structure analysis*. Houghton Mifflin, Boston (1968)
52. Lehmann, E.L., D'Abera, H.J.M.: *Nonparametrics: Statistical Methods Based on Ranks*. Springer, New York (2006)
53. Leimer, H.-G., Rudas, T.: Conversion between GLIM- and BMDP-type log-linear parameters. *GLIM Newsletter*, **19**, 47 (1989)
54. Lohr, S.L.: *Sampling: Design and Analysis*. Brooks/Cole, Boston (2009)
55. Lumley, T.: *Complex Surveys: A Guide to Analysis Using R*, Wiley, New York (2010)
56. Lupparelli, M., Marchetti, G.M., Bergsma, W.: Parameterization and fitting of bi-directed graph models to categorical data. *Scandinavian Journal of Statistics*, **36**, 559–576 (2008)
57. Meek, C., Glymour, C.: Conditioning and intervening. *The British Journal for the Philosophy of Science*, **45**, 1001–1021 (1994)
58. Miller, R: *Simultaneous Statistical Significance, 2nd ed.*. Springer, New York (1981)
59. Neutel, C. I.: The Potential for Simpson's Paradox in Drug Utilization Studies. *Annals of Epidemiology*, **7**, 517–521 (1997)
60. Németh, R., Rudas, T: On the application of discrete marginal graphical models. *Sociological Methofology*, **43**, 70–100 (2013)
61. Neyman, J.: *Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes*. Master's Thesis (1923). Excerpts reprinted in English, Statistical Science, Vol. 5, pp. 463–472. (D. M. Dabrowska, and T. P. Speed, Translators)
62. Pearl, J. *Causality*. Cambridge University Press (2000)
63. Pourret, O, Naim, P., Marcot, B.: *Bayesian Networks: A Practical Guide to Applications*. Wiley, New York (2008)
64. Read, T.R.C., Cressie, N.A.C.: *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer, New York (1988)
65. Reintjes, R, de Boer, A., van Pelt, W., Mintjes-de Groot, J.: Simpson's Paradox: An Example from Hospital Epidemiology. *Epidemiology*, **11**, 81–83 (2000)
66. Richardson, T.S., Spirtes, P.: Ancestral graph Markov models. *Annals of Statistics*, **30**, 962–1030 (2002)
67. Roberts, C.P., Casella, G.: *Monte Carlo Statistical Methods, 2nd ed.* Springer, New York (2013)
68. Rosenbaum, P., Rubin, D.: The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55 (1983)
69. Rubin, D.: Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, **100**, 322–331 (2005)
70. Rudas, T.: A Monte Carlo comparison of the small sample behaviour of the Pearson, the likelihood ratio and the Cressie-Read statistics. *Journal of Statistical Computation and Simulation*, **24**, 107–120 (1986)
71. Rudas, T.: Prescribed conditional interaction structure models with application to the analysis of mobility tables. *Quality & Quantity* **25**, 345–358 (1991)
72. Rudas, T.: *Odds Ratios in the Analysis of Contingency Tables*. Sage Publications, Thousand Oaks (1998)
73. Rudas, T.: The mixture ndex of fit and minimax regression. *Metrika*, **50**, 163–172 (1999)
74. Rudas, T.: Canonical representation of log-linear models. *Communications in Statistics – Theory and Methods*, **31**, 2311–2323 (2002)
75. Rudas, T.: Mixture models of missing data. *Quality & Quantity*, **39**, 19–36 (2005)
76. Rudas, T.: Informative allocation and consistent treatment selection. *Statistical Methodology*, **7**, 323–337 (2010)
77. Rudas, T.: Effects and interactions. *Methodology*, **11**, 142–149 (2015)

78. Rudas, T.: Directionally collapsible parameterizations of multivariate binary distributions. *Statistical Methodology*, **27**, 132–145 (2015)
79. Rudas, T., Bergsma, W.: Letter to the Editor. *Statistics in Medicine*, **23**, 3545–3547 (2004)
80. Rudas, T., Bergsma, W.: On applications of marginal models to categorical data. *Metron*, **42**, 15–37 (2004)
81. Rudas, T., Bergsma, W., Németh, R.: Parameterization and estimation of path models for categorical data. in Rizzi, A., Vichi, M., eds. *COMPSTAT 2006*, 383–394, Physica Verlag, Heidelberg (2006)
82. Rudas, T., Bergsma, W., Németh, R.: Marginal log-linear parameterization of conditional independence models. *Biometrika*, **97**, 1006–1012 (2010)
83. Rudas, T., Clogg, C.C., Lindsay, B.G.: A new index of fit based on mixture methods for the analysis of contingency tables. *Journal of the Royal Statistical Society, Ser B*, **56**, 623–639 (1994)
84. Rudas, T., Leimer, H.-G.: Analysis of contingency tables with known conditional odds ratios or known log-linear parameters. In: Francis, B., Seeberg, G. U. H., van der Heijden, P. G. M., Jansen, W. (eds.) *Statistical Modeling*, pp. 313–322, Elsevier, North-Holland, Amsterdam (1992)
85. Shackel, N.: Paradoxes in Probability Theory. In Rudas, T. (ed.) *Handbook of Probability: Theory and Applications*, pp. 49–66. Sage Publications, Thousand Oaks (2008)
86. Shao, J.: *Mathematical Statistics*, 2nd ed. Springer, New York (2003)
87. Spirtes, P., Glymour, C, Scheines, R.: *Causation, Prediction and Search*. Springer, New York (1993)
88. Tyree, A.: Mobility ratios and association in mobility tables. *Population Studies*, **27**, 577–588 (1973)
89. Uebersax, J.: The tetrachoric correlation. http://www.john-uebersax.com/stat/tetra.htm 13 July 2012
90. Vargha, A., Rudas, T. Delaney, H.D., Maxwell, S.E.: Dichotomization, partial correlation, and conditional independence. *Journal of Educational and Behavioral Statistics* **21**, 264–282 (1996)
91. Wainer, H., Brown, L.M.: Two statistical paradoxes in the interpretation of group differences illustrated with medical school admission and licensing data, *The American Statistician* **58**, 117–123 (2004)
92. Wu, C.F.J.: On the convergence properties of the EM algorithm. *Annals of Statistics* **11**, 95–103 (1983)

Index