CHAPTER 1

Introduction: Distributions and Inference for Categorical Data

From helping to assess the value of new medical treatments to evaluating the factors that affect our opinions and behaviors, analysts today are finding myriad uses for categorical data methods. In this book we introduce these methods and the theory behind them.

Statistical methods for categorical responses were late in gaining the level of sophistication achieved early in the twentieth century by methods for continuous responses. Despite influential work around 1900 by the British statistician Karl Pearson, relatively little development of models for categorical responses occurred until the 1960s. In this book we describe the early fundamental work that still has importance today but place primary emphasis on more recent modeling approaches.

1.1 CATEGORICAL RESPONSE DATA

A *categorical variable* has a measurement scale consisting of a set of categories. For instance, political philosophy is often measured as liberal, moderate, or conservative. Diagnoses regarding breast cancer based on a mammogram use the categories normal, benign, probably benign, suspicious, and malignant.

The development of methods for categorical variables was stimulated by the need to analyze data generated in research studies in both the social and biomedical sciences. Categorical scales are pervasive in the social sciences for measuring attitudes and opinions. Categorical scales in biomedical sciences measure outcomes such as whether a medical treatment is successful.

Categorical data are by no means restricted to the social and biomedical sciences. They frequently occur in the behavioral sciences (e.g., type of mental illness, with the categories schizophrenia, depression, neurosis), epidemiology and public health (e.g., contraceptive method at last sexual intercourse, with the categories none, condom, pill, IUD, other), genetics (type of allele inherited by an offspring), botany and zoology (e.g., whether or not a particular organism is observed in a sampled quadrat), education (e.g., whether a student response to an exam question is correct or incorrect), and marketing (e.g., consumer

Categorical Data Analysis, Third Edition. Alan Agresti.

preference among the three leading brands of a product). They even occur in highly quantitative fields such as engineering sciences and industrial quality control. Examples are the classification of items according to whether they conform to certain standards, and subjective evaluation of some characteristic: how soft to the touch a certain fabric is, how good a particular food product tastes, or how easy a worker finds it to perform a certain task.

Categorical variables are of many types. In this section we provide ways of classifying them.

1.1.1 Response–Explanatory Variable Distinction

Statistical analyses distinguish between *response* (or *dependent*) *variables* and *explanatory* (or *independent*) *variables*. This book focuses on methods for categorical response variables. As in ordinary regression modeling, explanatory variables can be any type. For instance, a study might analyze how opinion about whether same-sex marriages should be legal (yes or no) changes according to values of explanatory variables, such as religious affiliation, political ideology, number of years of education, annual income, age, gender, and race.

1.1.2 Binary–Nominal–Ordinal Scale Distinction

Many categorical variables have only two categories. Such variables, for which the two categories are often given the generic labels "success" and "failure," are called *binary variables*. A major topic of this book is the modeling of binary response variables.

When a categorical variable has more than two categories, we distinguish between two types of categorical scales. Variables having categories without a natural ordering are said to be measured on a *nominal scale* and are called *nominal variables*. Examples are mode of transportation to get to work (automobile, bicycle, bus, subway, walk), favorite type of music (classical, country, folk, jazz, rock), and choice of residence (apartment, condominium, house, other). For nominal variables, the order of listing the categories is irrelevant to the statistical analysis.

Many categorical variables *do* have ordered categories. Such variables are said to be measured on an *ordinal scale* and are called *ordinal variables*. Examples are social class (upper, middle, lower), political philosophy (very liberal, slightly liberal, moderate, slightly conservative, very conservative), patient condition (good, fair, serious, critical), and rating of a movie for Netflix (1 to 5 stars, representing hated it, didn't like it, liked it, really liked it, loved it). For ordinal variables, distances between categories are unknown. Although a person categorized as very liberal is more liberal than a person categorized as slightly liberal, no numerical value describes *how much more* liberal that person is.

An *interval variable* is one that *does* have numerical distances between any two values. For example, systolic blood pressure level, length of prison term, and annual income are interval variables. For most such variables, it is also possible to compare two values by their ratio, in which case the variable is also called a *ratio variable*.

The way that a variable is measured determines its classification. For example, "education" is only nominal when measured as (public school, private school, home schooling); it is ordinal when measured by highest degree attained, using the categories (none, high school, bachelor's, master's, doctorate); it is interval when measured by number of years of education completed, using the integers 0, 1, 2, 3, ...

A variable's measurement scale determines which statistical methods are appropriate. It is usually best to apply methods appropriate for the actual scale. In the measurement hierarchy, interval variables are highest, ordinal variables are next, and nominal variables are lowest. Statistical methods for variables of one type can also be used with variables at higher levels but not at lower levels. For instance, statistical methods for nominal variables can be used with ordinal variables by ignoring the ordering of categories. Methods for ordinal variables cannot, however, be used with nominal variables, since their categories have no meaningful ordering. The distinction between ordered and unordered categories is not important for binary variables, because ordinal methods and nominal methods then typically reduce to equivalent methods.

In this book, we present methods for the analysis of binary, nominal, and ordinal variables. The methods also apply to interval variables having a small number of distinct values (e.g., number of times married, number of distinct side effects experienced in taking some drug) or for which the values are grouped into ordered categories (e.g., education measured as $\leq$12 years, $>$12 but $<$16 years, $\geq$16 years).

1.1.3 Discrete–Continuous Variable Distinction

Variables are classified as *discrete* or *continuous*, according to whether the number of values they can take is countable. Actual measurement of all variables occurs in a discrete manner, due to precision limitations in measuring instruments. The discrete–continuous classification, in practice, distinguishes between variables that take few values and variables that take lots of values. For instance, statisticians often treat discrete interval variables having a large number of values (such as test scores) as continuous, using them in methods for continuous responses.

This book deals with certain types of discretely measured responses: (1) binary variables, (2) nominal variables, (3) ordinal variables, (4) discrete interval variables having relatively few values, and (5) continuous variables grouped into a small number of categories.

1.1.4 Quantitative–Qualitative Variable Distinction

Nominal variables are *qualitative*—distinct categories differ in quality, not in quantity. Interval variables are *quantitative*—distinct levels have differing amounts of the characteristic of interest. The position of ordinal variables in the qualitative–quantitative classification is fuzzy. Analysts often treat them as qualitative, using methods for nominal variables. But in many respects, ordinal variables more closely resemble interval variables than they resemble nominal variables. They possess important quantitative features: Each category has a *greater* or *smaller* magnitude of the characteristic than another category; and although not possible to measure, an underlying continuous variable is often present. The political ideology classification (very liberal, slightly liberal, moderate, slightly conservative, very conservative) crudely measures an inherently continuous characteristic.

Analysts often utilize the quantitative nature of ordinal variables by assigning numerical scores to the categories or assuming an underlying continuous distribution. This requires good judgment and guidance from researchers who use the scale, but it provides benefits in the variety of methods available for data analysis.

1.1.5 Organization of Book and Online Computing Appendix

The models for categorical response variables discussed in this book resemble regression models for continuous response variables; however, they assume binomial or multinomial response distributions instead of normality. One type of model receives special attention—*logistic regression*. Ordinary logistic regression models apply with *binary* responses and assume a binomial distribution. Generalizations of logistic regression apply with multicategory responses and assume a multinomial distribution.

The book has four main units. In the first, Chapters 1 through 3, we summarize descriptive and inferential methods for univariate and bivariate categorical data. These chapters cover discrete distributions, methods of inference, and measures of association for contingency tables. They summarize the non-model-based methods developed prior to about 1960.

In the second and primary unit, Chapters 4 through 10, we introduce models for categorical responses. In Chapter 4 we describe a class of *generalized linear models* having models of this text as special cases. Chapters 5 and 6 cover the most important model for binary responses, logistic regression. Chapter 7 presents alternative methods for binary data, including the probit, Bayesian fitting, and smoothing methods. In Chapter 8 we present generalizations of the logistic regression model for nominal and ordinal multicategory response variables. In Chapters 9 and 10 we introduce the modeling of multivariate categorical response data, in terms of association and interaction patterns among the variables. The models, called *loglinear models*, apply to counts in the table that cross-classifies those responses.

In the third unit, Chapters 11 through 14, we discuss models for handling repeated measurement and other forms of clustered data. In Chapter 11 we present models for a categorical response with matched pairs; these apply, for instance, with a categorical response measured for the same subjects at two times. Chapter 12 covers models for more general types of repeated categorical data, such as longitudinal data from several times with explanatory variables. In Chapter 13 we present a broad class of models, *generalized linear mixed models*, that use random effects to account for dependence with such data. In Chapter 14 further extensions of the models from Chapters 11 through 13 are described, unified by treating the response as having a mixture distribution of some type.

The fourth and final unit has a different nature than the others. In Chapter 15 we consider non-model-based classification and clustering methods. In Chapter 16 we summarize large-sample and small-sample theory for categorical data models. This theory is the basis for behavior of model parameter estimators and goodness-of-fit statistics. Chapter 17 presents a historical overview of the development of categorical data methods.

Maximum likelihood methods receive primary attention throughout the book. Many chapters, however, contain a section presenting corresponding Bayesian methods.

In Appendix A we review software that can perform the analyses in this book. The website `www.stat.ufl.edu/~aa/cda/cda.html` for this book contains an appendix that gives more information about using R, SAS, Stata, and other software, with sample programs for text examples. In addition, that site has complete data sets for many text examples and exercises, solutions to some exercises, extra exercises, corrections, and links to other useful sites. For instance, a manual prepared by Dr. Laura Thompson provides examples of how to use R and S-Plus for all examples in the second edition of this text, many of which (or very similar ones) are also in this edition.

In the rest of this chapter, we provide background material. In Section 1.2 we review the key distributions for categorical data: the binomial and multinomial, as well as another that

is important for discrete data, the Poisson. In Section 1.3 we review the primary mechanisms for statistical inference using maximum likelihood. In Sections 1.4 and 1.5 we illustrate these by presenting significance tests and confidence intervals for binomial and multinomial parameters. In Section 1.6 we introduce Bayesian inference for these parameters.

1.2 DISTRIBUTIONS FOR CATEGORICAL DATA

Inferential data analyses require assumptions about the random mechanism that generated the data. For regression models with continuous responses, the normal distribution plays the central role. In this section we review the three key distributions for categorical responses: *binomial*, *multinomial*, and *Poisson*.

1.2.1 Binomial Distribution

Many applications refer to a fixed number n of binary observations. Let $y_1, y_2, \ldots, y_n$ denote observations from n independent and identical trials such that $P(Y_i = 1) = \pi$ and $P(Y_i = 0) = 1 - \pi$. We refer to outcome 1 as "success" and outcome 0 as "failure." *Identical trials* means that the probability of success π is the same for each trial. *Independent trials* means that the $\{Y_i\}$ are independent random variables. These are often called *Bernoulli trials*. The total number of successes, $Y = \sum_{i=1}^{n} Y_i$, has the *binomial distribution* with index n and parameter π, denoted by bin(n, π).

The probability mass function for the possible outcomes y for Y is

$$p(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \quad y = 0, 1, 2, \ldots, n, \tag{1.1}$$

where the binomial coefficient $\binom{n}{y} = n!/[y!(n-y)!]$. Since $E(Y_i) = E(Y_i^2) = 1 \times \pi + 0 \times (1-\pi) = \pi$,

$$E(Y_i) = \pi \quad \text{and} \quad \text{var}(Y_i) = \pi(1-\pi).$$

The binomial distribution for $Y = \sum_i Y_i$ has mean and variance

$$\mu = E(Y) = n\pi \quad \text{and} \quad \sigma^2 = \text{var}(Y) = n\pi(1-\pi).$$

The skewness is described by $E(Y-\mu)^3/\sigma^3 = (1-2\pi)/\sqrt{n\pi(1-\pi)}$. The distribution is symmetric when $\pi = 0.50$ but becomes increasingly skewed as π moves toward either boundary. The binomial distribution converges to normality as n increases, for fixed π, the approximation being reasonable[1] when $n[\min(\pi, 1-\pi)]$ is as small as about 5.

There is no guarantee that successive binary observations are independent or identical. Thus, occasionally, we will utilize other distributions. One such case is sampling binary outcomes without replacement from a finite population, such as observations on whether a homework assignment was completed for 10 students sampled from a class of size 20. The

[1] See `www.stat.tamu.edu/~west/applets/binomialdemo2.html`.

hypergeometric distribution, studied in Section 3.5.1, is then relevant. In Section 1.2.4 we discuss another case that violates the binomial assumptions.

1.2.2 Multinomial Distribution

Some trials have more than two possible outcomes. Suppose that each of n independent, identical trials can have outcome in any of c categories. Let $y_{ij} = 1$ if trial i has outcome in category j and $y_{ij} = 0$ otherwise. Then $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ic})$ represents a multinomial trial, with $\sum_j y_{ij} = 1$; for instance, (0, 0, 1, 0) denotes outcome in category 3 of four possible categories. Note that y_{ic} is redundant, being linearly dependent on the others. Let $n_j = \sum_i y_{ij}$ denote the number of trials having outcome in category j. The counts $(n_1, n_2, \ldots, n_c)$ have the *multinomial distribution*.

Let $\pi_j = P(Y_{ij} = 1)$ denote the probability of outcome in category j for each trial. The multinomial probability mass function is

$$p(n_1, n_2, \ldots, n_{c-1}) = \left(\frac{n!}{n_1!n_2!\cdots n_c!}\right)\pi_1^{n_1}\pi_2^{n_2}\cdots\pi_c^{n_c}. \tag{1.2}$$

Since $\sum_j n_j = n$, this is $(c-1)$-dimensional, with $n_c = n - (n_1 + \cdots + n_{c-1})$. The binomial distribution is the special case with $c = 2$.

For the multinomial distribution,

$$E(n_j) = n\pi_j, \quad \text{var}(n_j) = n\pi_j(1-\pi_j), \quad \text{cov}(n_j, n_k) = -n\pi_j\pi_k. \tag{1.3}$$

We derive the covariance in Section 16.1.4. The marginal distribution of each n_j is binomial.

1.2.3 Poisson Distribution

Sometimes, count data do not result from a fixed number of trials. For instance, if $Y =$ number of automobile accidents today on motorways in Italy, there is no fixed upper bound n for Y (as you are aware if you have driven in Italy!). Since Y must take a nonnegative integer value, its distribution should place its mass on that range. The simplest such distribution is the *Poisson*. Its probabilities depend on a single parameter, the mean μ. The Poisson probability mass function (Poisson 1837, p. 206) is

$$p(y) = \frac{e^{-\mu}\mu^y}{y!}, \quad y = 0, 1, 2, \ldots. \tag{1.4}$$

It satisfies $E(Y) = \text{var}(Y) = \mu$. It is unimodal with mode equal to the integer part of μ. Its skewness is described by $E(Y-\mu)^3/\sigma^3 = 1/\sqrt{\mu}$. The Poisson distribution approaches normality as μ increases, the normal approximation being quite good when μ is at least about 10.

The Poisson distribution is used for counts of events that occur randomly over time or space, when outcomes in disjoint periods or regions are independent. It also applies as an approximation for the binomial when n is large and π is small, with $\mu = n\pi$. For example, suppose $Y =$ *number of deaths* today in auto accidents in Italy (rather than the *number of accidents*). Then, Y has an upper bound. If each of the 50 million people driving in Italy is an independent trial with probability 0.0000003 of dying today in an auto accident, the

number of deaths Y is a bin(50000000, 0.0000003) variate. This is approximately Poisson with $\mu = n\pi = 50000000(0.0000003) = 15$.

A key feature of the Poisson distribution is that its variance equals its mean. Sample counts vary more when their mean is higher. When the mean number of daily fatal accidents equals 15, greater variability occurs from day to day than when the mean equals 2.

1.2.4 Overdispersion

In practice, count observations often exhibit variability exceeding that predicted by the binomial or Poisson. This phenomenon is called *overdispersion*. We assumed above that each person has the same probability each day of dying in a fatal auto accident. More realistically, these probabilities vary from day to day according to the amount of road traffic and weather conditions and vary from person to person according to factors such as the amount of time spent in autos, whether the person wears a seat belt, how much of the driving is at high speeds, gender, and age. Such variation causes fatality counts to display more variation than predicted by the Poisson model.

Suppose that Y is a random variable with variance $\text{var}(Y|\mu)$ for given μ, but μ itself varies because of unmeasured factors such as those just described. Let $\theta = E(\mu)$. Then unconditionally,

$$E(Y) = E[E(Y|\mu)], \quad \text{var}(Y) = E[\text{var}(Y|\mu)] + \text{var}[E(Y|\mu)].$$

When Y is conditionally Poisson (given μ), then $E(Y) = E(\mu) = \theta$ and $\text{var}(Y) = E(\mu) + \text{var}(\mu) = \theta + \text{var}(\mu) > \theta$.

Assuming a Poisson distribution for a count variable is often too simplistic, because of factors that cause overdispersion. The *negative binomial* is a related distribution for count data that has a second parameter and permits the variance to exceed the mean. We introduce it in Section 4.3.4.

Analyses assuming binomial (or multinomial) distributions are also sometimes invalid because of overdispersion. This might happen because the true distribution is a mixture of different binomial distributions, with the parameter varying because of unmeasured variables. To illustrate, suppose that an experiment exposes pregnant mice to a toxin and then after a week observes the number of fetuses in each mouse's litter that show signs of malformation. Let n_i denote the number of fetuses in the litter for mouse i. The pregnant mice also vary according to other factors, such as their weight, overall health, and genetic makeup. Extra variation then occurs because of the variability from litter to litter in the probability π of malformation. The distribution of the number of fetuses per litter showing malformations might cluster near 0 and near n_i, showing more dispersion than expected for binomial sampling with a single value of π. Overdispersion could also occur when π varies among fetuses in a litter according to some distribution (Exercise 1.17). In Chapters 4, 13, and 14 we introduce methods for data that are overdispersed relative to binomial and Poisson assumptions.

1.2.5 Connection Between Poisson and Multinomial Distributions

For adult residents of Britain who visit France this year, let $Y_1 =$ number who fly there, $Y_2 =$ number who travel there by train without a car (Eurostar), $Y_3 =$ number who travel there by ferry without a car, and $Y_4 =$ number who take a car (by Eurotunnel Shuttle or

a ferry). A Poisson model for (Y_1, Y_2, Y_3, Y_4) treats these as independent Poisson random variables, with parameters $(\mu_1, \mu_2, \mu_3, \mu_4)$. The joint probability mass function for $\{Y_i\}$ is the product of the four mass functions of form (1.4). The total $n = \sum_i Y_i$ also has a Poisson distribution, with parameter $\sum_i \mu_i$.

With Poisson sampling the total count n is random rather than fixed. If we assume a Poisson model but condition on n, $\{Y_i\}$ no longer have Poisson distributions, since each Y_i cannot exceed n. Given n, $\{Y_i\}$ are also no longer independent, since the value of one affects the possible range for the others.

For c independent Poisson variates, with $E(Y_i) = \mu_i$, the conditional probability of a set of counts $\{n_i\}$ satisfying $\sum_i Y_i = n$ is

$$\begin{aligned} P\Big[(Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c) \Big| \sum_j Y_j = n\Big] \\ = \frac{P(Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c)}{P\left(\sum_j Y_j = n\right)} \\ = \frac{\prod_i [\exp(-\mu_i)\mu_i^{n_i}/n_i!]}{\exp\left(-\sum_j \mu_j\right)\left(\sum_j \mu_j\right)^n/n!} = \frac{n!}{\prod_i n_i!} \prod_i \pi_i^{n_i}, \end{aligned} \tag{1.5}$$

where $\left\{\pi_i = \mu_i / \left(\sum_j \mu_j\right)\right\}$. This is the multinomial $(n, \{\pi_i\})$ distribution, characterized by the sample size n and the probabilities $\{\pi_i\}$.

Many categorical data analyses assume a multinomial distribution. Such analyses usually have the same inferential results as those of analyses assuming a Poisson distribution, because of the similarity in the likelihood functions.

1.2.6 The Chi-Squared Distribution

Another distribution of fundamental importance for categorical data is the *chi-squared*, not as a distribution for the data but rather as a sampling distribution for many statistics. Because of its importance, we summarize here a few of its properties.

The chi-squared distribution with degrees of freedom denoted by df has mean df, variance 2(df), and skewness $\sqrt{8/\text{df}}$. It converges (slowly) to normality as df increases, the approximation being reasonably good when df is at least about 50.

Let Z denote a standard normal random variable (mean 0, variance 1). Then Z^2 has a chi-squared distribution with df = 1. A chi-squared random variable with df = ν has representation $Z_1^2 + \cdots + Z_\nu^2$, where $Z_1, \ldots, Z_\nu$ are independent standard normal variables. Thus, a chi-squared statistic having df = ν has partitionings into independent chi-squared components—for example, into ν components each having df = 1. Conversely, the *reproductive property* states that if X_1^2 and X_2^2 are independent chi-squared random variables having degrees of freedom ν_1 and ν_2, then $X^2 = X_1^2 + X_2^2$ has a chi-squared distribution with df = $\nu_1 + \nu_2$.

1.3 STATISTICAL INFERENCE FOR CATEGORICAL DATA

In practice, the probability distribution assumed for the response variable has unknown parameter values. In this section we review methods of using sample data to make

inferences about the parameters. Sections 1.4 and 1.5 illustrate these methods for binomial and multinomial parameters.

1.3.1 Likelihood Functions and Maximum Likelihood Estimation

In this book we use *maximum likelihood* for parameter estimation. Maximum likelihood estimators have desirable properties: They have large-sample normal distributions; they are asymptotically consistent, converging to the parameter as n increases; and they are asymptotically efficient, producing large-sample standard errors no greater than those from other estimation methods. These results hold under weak regularity conditions, mainly that the number of parameters remains constant as n increases and that the true values of those parameters fall in the interior (rather than on the boundary) of the parameter space.

Given the data, for a chosen probability distribution the *likelihood function* is the probability of those data, treated as a function of the unknown parameter. The maximum likelihood (ML) estimate is the parameter value that maximizes this function. This is the parameter value under which the data observed have the highest probability of occurrence. We denote a parameter for a generic problem by β and its ML estimate by $\hat{\beta}$. We denote the likelihood function by $\ell(\beta)$. The β value that maximizes $\ell(\beta)$ also maximizes $L(\beta) = \log[\ell(\beta)]$. It is simpler to maximize $L(\beta)$ since it is a sum rather than a product of terms. For many models, $L(\beta)$ has concave shape and $\hat{\beta}$ is the point at which the derivative equals 0. The ML estimate is then the solution of the likelihood equation, $\partial L(\beta)/\partial\beta = 0$. Often, β is multidimensional, denoted by $\boldsymbol{\beta}$, and $\hat{\boldsymbol{\beta}}$ is the solution of a set of likelihood equations.

Let $\text{cov}(\hat{\boldsymbol{\beta}})$ denote the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$. Under regularity conditions (Rao 1973, p. 364), $\text{cov}(\hat{\boldsymbol{\beta}})$ is the inverse of the *information matrix*. The (j, k) element of the information matrix is

$$-E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k}\right). \tag{1.6}$$

The standard errors are the square roots of the diagonal elements for the inverse of the information matrix. The greater the curvature of the log likelihood function, the smaller the standard errors. This is reasonable, since large curvature implies that the log likelihood drops quickly as $\boldsymbol{\beta}$ moves away from $\hat{\boldsymbol{\beta}}$; hence, the data would have been much more likely to occur if $\boldsymbol{\beta}$ took a value near $\hat{\boldsymbol{\beta}}$ rather than a value far from $\hat{\boldsymbol{\beta}}$.

1.3.2 Likelihood Function and ML Estimate for Binomial Parameter

The part of a likelihood function involving the parameters is called the *kernel*. Since the maximization of the likelihood is done with respect to the parameters, the rest is irrelevant.

To illustrate, consider the binomial distribution (1.1). The binomial coefficient $n!/[y!(n-y)!]$ has no influence on where the maximum occurs with respect to π. Thus, we ignore it and treat the kernel as the likelihood function. The binomial log likelihood function is then

$$L(\pi) = \log[\pi^y(1-\pi)^{n-y}] = y\log(\pi) + (n-y)\log(1-\pi). \tag{1.7}$$

Differentiating with respect to π yields

$$\partial L(\pi)/\partial\pi = y/\pi - (n-y)/(1-\pi) = (y - n\pi)/\pi(1-\pi). \tag{1.8}$$

Equating this to 0 gives the likelihood equation, which has solution $\hat{\pi} = y/n$, the sample proportion of successes for the n trials.

Calculating $\partial^2 L(\pi)/\partial\pi^2$, taking the expectation, and combining terms, we get

$$-E[\partial^2 L(\pi)/\partial\pi^2] = E[y/\pi^2 + (n-y)/(1-\pi)^2] = n/[\pi(1-\pi)]. \tag{1.9}$$

Thus, the asymptotic variance of $\hat{\pi}$ is $\pi(1-\pi)/n$. This is no surprise. Since $E(Y) = n\pi$ and $\text{var}(Y) = n\pi(1-\pi)$, the distribution of $\hat{\pi} = Y/n$ has mean and standard deviation

$$E(\hat{\pi}) = \pi, \quad \sigma(\hat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

1.3.3 Wald–Likelihood Ratio–Score Test Triad

There are three standard ways to use the likelihood function to perform large-sample inference. We introduce these for a significance test of a null hypothesis H_0: $\beta = \beta_0$ and then discuss their relation to interval estimation. They all exploit the large-sample normality of ML estimators.

Standard errors obtained from the inverse of the information matrix depend on the unknown parameter values. When we substitute the unrestricted ML estimates (i.e., not assuming the null hypothesis) we obtain an estimated standard error of $\hat{\beta}$, which we denote by SE. Denote $-E[\partial^2 L(\beta)/\partial\beta^2]$ (i.e., the information) evaluated at $\hat{\beta}$ by $\iota(\hat{\beta})$. The first large-sample inference method has test statistic using this estimated standard error,

$$z = (\hat{\beta} - \beta_0)/SE, \quad \text{where} \quad SE = 1/\sqrt{\iota(\hat{\beta})}.$$

This statistic has an approximate standard normal distribution when $\beta = \beta_0$. We refer z to the standard normal table to obtain one- or two-sided P-values. Equivalently, for the two-sided alternative, z^2 has an approximate chi-squared null distribution with df = 1; the P-value is then the right-tailed chi-squared probability above the observed value. This type of statistic, using the nonnull estimated standard error, is called a *Wald statistic* (Wald 1943).

The multivariate extension[2] for the Wald test of H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ has test statistic

$$W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T [\text{cov}(\hat{\boldsymbol{\beta}})]^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

The nonnull covariance is based on the curvature (1.6) of the log-likelihood function at $\hat{\boldsymbol{\beta}}$ and typically itself requires estimation. The asymptotic multivariate normal distribution for $\hat{\boldsymbol{\beta}}$ implies an asymptotic chi-squared distribution for W. The df equal the rank of $\text{cov}(\hat{\boldsymbol{\beta}})$, which is the number of nonredundant parameters in $\boldsymbol{\beta}$.

[2]The ^{T} superscript on a vector or matrix denotes the transpose.

A second general-purpose method uses the likelihood function through the ratio of two maximizations: (1) the maximum over the possible parameter values under H_0, and (2) the maximum over the larger set of parameter values permitting H_0 or an alternative H_a to be true. Let ℓ_0 denote the maximized value of the likelihood function under H_0, and let ℓ_1 denote the maximized value generally (i.e., under $H_0 \cup H_a$). For instance, for parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$ and H_0: $\boldsymbol{\beta}_0 = \mathbf{0}$, ℓ_1 is the likelihood function calculated at the $\boldsymbol{\beta}$ value for which the data would have been most likely; ℓ_0 is the likelihood function calculated at the $\boldsymbol{\beta}_1$ value for which the data would have been most likely, when $\boldsymbol{\beta}_0 = \mathbf{0}$. Then ℓ_1 is always at least as large as ℓ_0, since ℓ_0 results from maximizing over a restricted set of the parameter values.

The ratio $\Lambda = \ell_0/\ell_1$ of the maximized likelihoods cannot exceed 1. Wilks (1935, 1938) showed that $-2 \log \Lambda$ has a limiting null chi-squared distribution, as $n \to \infty$. The df equal the difference in the dimensions of the parameter spaces under $H_0 \cup H_a$ and under H_0. The *likelihood-ratio test statistic* equals

$$-2 \log \Lambda = -2 \log(\ell_0/\ell_1) = -2(L_0 - L_1),$$

where L_0 and L_1 denote the maximized log-likelihood functions. [In this book, we use the natural logarithm throughout, for which its inverse is the exponential function; so, if $a = \log(b)$, then $b = \exp(a) = e^a$.]

The third method uses the *score statistic*, due to R. A. Fisher and C. R. Rao. The score test, referred to in some literature as the *Lagrange multiplier test*, is based on the slope and expected curvature of the log-likelihood function $L(\beta)$ at the null value β_0. It utilizes the size of the *score function*

$$u(\beta) = \partial L(\beta)/\partial \beta,$$

evaluated at β_0. The value $u(\beta_0)$ tends to be larger in absolute value when $\hat{\beta}$ is farther from β_0. Denote $-E[\partial^2 L(\beta)/\partial \beta^2]$ evaluated at β_0 by $\iota(\beta_0)$. The score statistic is the ratio of $u(\beta_0)$ to its null *SE*, which is $[\iota(\beta_0)]^{1/2}$. This has an approximate standard normal null distribution. The chi-squared form of the score statistic is

$$\frac{[u(\beta_0)]^2}{\iota(\beta_0)} = \frac{[\partial L(\beta)/\partial \beta_0]^2}{-E[\partial^2 L(\beta)/\partial \beta_0^2]},$$

where the notation reflects derivatives with respect to β that are evaluated at β_0. In the multiparameter case, the score statistic is a quadratic form based on the vector of partial derivatives of the log likelihood with respect to $\boldsymbol{\beta}$ and the inverse information matrix, both evaluated at the H_0 estimates (i.e., assuming that $\boldsymbol{\beta} = \boldsymbol{\beta}_0$).

Figure 1.1 shows a plot of a generic log-likelihood function $L(\beta)$ for the univariate case. It illustrates the three tests of H_0: $\beta = 0$. The Wald test uses the behavior of $L(\beta)$ at the ML estimate $\hat{\beta}$, having chi-squared form $(\hat{\beta}/SE)^2$. The *SE* of $\hat{\beta}$ depends on the curvature of $L(\beta)$ at $\hat{\beta}$. The score test is based on the slope and curvature of $L(\beta)$ at $\beta = 0$. The likelihood-ratio test combines information about $L(\beta)$ at both $\hat{\beta}$ and $\beta_0 = 0$. It compares the log-likelihood values L_1 at $\hat{\beta}$ and L_0 at $\beta_0 = 0$ using the chi-squared statistic $-2(L_0 - L_1)$. In Figure 1.1, this statistic is twice the vertical distance between values of $L(\beta)$ at $\hat{\beta}$ and at 0.

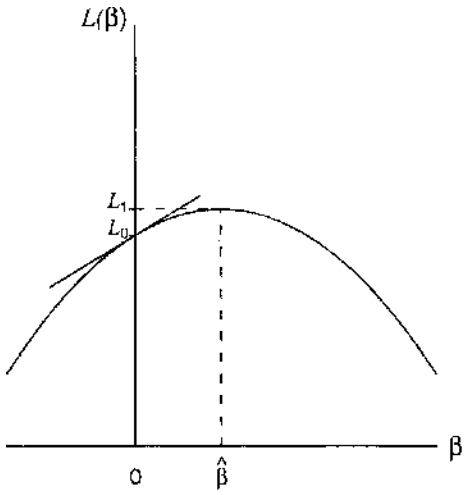

Figure 1.1 Log-likelihood function and information used in three tests of H_0: $\beta = 0$.

Section 1.4.1 illustrates the Wald, likelihood-ratio, and score tests for inference about a binomial parameter. As $n \to \infty$, the three tests have certain asymptotic equivalences (Cox and Hinkley 1974, Sec. 9.3). For small to moderate sample sizes, the likelihood-ratio and score tests are usually more reliable than the Wald test, having actual error rates closer to the nominal level.

1.3.4 Constructing Confidence Intervals by Inverting Tests

In practice, it is more informative to construct confidence intervals for parameters than to test hypotheses about their values. For any of the three test methods, we can construct a confidence interval by inverting the test. For instance, a 95% confidence interval for β is the set of β_0 for which the test of H_0: $\beta = \beta_0$ has P-value exceeding 0.05.

Let z_a denote the z-score from the standard normal distribution having right-tailed probability a; this is the $100(1-a)$ percentile of that distribution. A $100(1-\alpha)\%$ confidence interval based on asymptotic normality uses $z_{\alpha/2}$, for instance $z_{0.025} = 1.96$ for 95% confidence. The Wald confidence interval is the set of β_0 for which $|\hat{\beta} - \beta_0|/SE < z_{\alpha/2}$. This gives the interval $\hat{\beta} \pm z_{\alpha/2}(SE)$. Let $\chi^2_{df}(a)$ denote the $100(1-a)$ percentile of the chi-squared distribution with degrees of freedom df. The likelihood-ratio-based confidence interval is the set of β_0 for which $-2[L(\beta_0) - L(\hat{\beta})] < \chi^2_1(\alpha)$. [Note that $\chi^2_1(\alpha) = z^2_{\alpha/2}$.]

When $\hat{\beta}$ has a normal distribution, the log-likelihood function has a parabolic shape. For small samples with categorical data, $\hat{\beta}$ may be far from normality and the log-likelihood function can be far from a symmetric, parabolic-shaped curve. This can also happen with moderate to large samples when β falls near the boundary of the parameter space, such as a population proportion that is near 0 or near 1. In such cases, inference based on asymptotic normality of $\hat{\beta}$ may have inadequate performance. A marked divergence in results of Wald and likelihood-ratio inference indicates that the distribution of $\hat{\beta}$ may not be close to normality. The example in Section 1.4.3 illustrates.

The Wald confidence interval is commonly used in practice, because it is simple to construct using ML estimates and standard errors reported by statistical software. The

likelihood-ratio-test-based interval is becoming more widely available in software and is preferable for categorical data with small to moderate n. The score-test-based interval is widely available only in certain cases, such as for proportions as outlined in Section 1.4.2. For the best known statistical model, regression for a normal response, the three types of inference provide identical results. In later chapters, we'll use versions of these intervals that apply for models with multiple parameters. Especially useful is the *profile likelihood* approach based on inverting likelihood-ratio tests (e.g., Section 3.2.6).

1.4 STATISTICAL INFERENCE FOR BINOMIAL PARAMETERS

In this section we illustrate inference methods for categorical data by presenting tests and confidence intervals for the binomial parameter π. With y successes in n independent trials, recall that the ML estimator of π is $\hat{\pi} = y/n$, for which $E(\hat{\pi}) = \pi$ and $\text{var}(\hat{\pi}) = \pi(1-\pi)/n$.

1.4.1 Tests About a Binomial Parameter

Consider H_0: $\pi = \pi_0$. Since H_0 has a single parameter, we use the normal rather than chi-squared forms of Wald and score test statistics. They permit tests against one-sided as well as two-sided alternatives.

The Wald statistic for testing H_0: $\pi = \pi_0$ is

$$z_W = \frac{\hat{\pi} - \pi_0}{SE} = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}}. \tag{1.10}$$

To find the score statistic, we evaluate the binomial score (1.8) and information (1.9) at π_0. This yields

$$u(\pi_0) = \frac{y}{\pi_0} - \frac{n-y}{1-\pi_0}, \quad \iota(\pi_0) = \frac{n}{\pi_0(1-\pi_0)}.$$

The normal form of the score statistic simplifies to

$$z_S = \frac{u(\pi_0)}{[\iota(\pi_0)]^{1/2}} = \frac{y - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}. \tag{1.11}$$

Whereas the Wald statistic z_W uses the standard error evaluated at $\hat{\pi}$, the score statistic z_S uses it evaluated at π_0. The score statistic is preferable, as it uses the actual null *SE* rather than an estimate. Its null sampling distribution is closer to standard normal than that of the Wald statistic.

The binomial log-likelihood function (1.7) equals $L_0 = y \log \pi_0 + (n-y)\log(1-\pi_0)$ under H_0 and $L_1 = y \log \hat{\pi} + (n-y)\log(1-\hat{\pi})$ more generally. The likelihood-ratio test statistic simplifies to

$$-2(L_0 - L_1) = 2\left[y \log \frac{\hat{\pi}}{\pi_0} + (n-y) \log \frac{1-\hat{\pi}}{1-\pi_0} \right].$$

Expressed as

$$-2(L_0 - L_1) = 2\left[y \log \frac{y}{n\pi_0} + (n-y)\log \frac{n-y}{n-n\pi_0}\right],$$

it compares observed success and failure counts with fitted counts under H_0 by

$$2\sum \text{observed}\left[\log\left(\frac{\text{observed}}{\text{fitted}}\right)\right]. \tag{1.12}$$

We'll see that this formula also holds for tests about Poisson and multinomial parameters. Since no unknown parameters occur under H_0 and one occurs under H_a, the asymptotic chi-squared distribution for (1.12) has df $= 1 - 0 = 1$.

1.4.2 Confidence Intervals for a Binomial Parameter

Inverting the Wald test statistic gives the interval of π_0 values for which $|z_W| < z_{\alpha/2}$, or

$$\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}. \tag{1.13}$$

Historically, this was one of the first confidence intervals used for any parameter (Laplace 1812, p. 283). Unfortunately, it performs poorly unless n is very large (e.g., Brown et al. 2001), in the sense that the actual probability that the interval contains π usually falls below the nominal confidence coefficient, much below when π is near 0 or 1.

The likelihood-ratio-based confidence interval is more complex computationally, but simple in principle. It is the set of π_0 for which the likelihood-ratio test has a P-value exceeding α. Equivalently, it is the set of π_0 for which double the log likelihood drops by less than $\chi^2_1(\alpha)$ from its value at the ML estimate $\hat{\pi} = y/n$. For example, the endpoints of the 95% confidence interval can be found using numerical methods to iteratively solve for the values of π_0 that satisfy

$$2\left[y \log \frac{\hat{\pi}}{\pi_0} + (n-y)\log \frac{1-\hat{\pi}}{1-\pi_0}\right] = \chi^2_1(0.05) = 3.84.$$

The score confidence interval contains π_0 values for which $|z_S| < z_{\alpha/2}$. Its endpoints are the π_0 solutions to the equations

$$(\hat{\pi} - \pi_0)/\sqrt{\pi_0(1-\pi_0)/n} = \pm z_{\alpha/2}.$$

These are quadratic in π_0. First discussed by Wilson (1927), this interval is

$$\left[\hat{\pi}\left(\frac{n}{n+z^2_{\alpha/2}}\right) + \frac{1}{2}\left(\frac{z^2_{\alpha/2}}{n+z^2_{\alpha/2}}\right)\right]$$
$$\pm z_{\alpha/2}\sqrt{\frac{1}{n+z^2_{\alpha/2}}\left[\hat{\pi}(1-\hat{\pi})\left(\frac{n}{n+z^2_{\alpha/2}}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{z^2_{\alpha/2}}{n+z^2_{\alpha/2}}\right)\right]}. \tag{1.14}$$

The midpoint is a weighted average of $\hat{\pi}$ and $\frac{1}{2}$, where the weight $n/(n + z^2_{\alpha/2})$ given $\hat{\pi}$ increases as n increases. Combining terms, this midpoint equals $\tilde{\pi} = (y + z^2_{\alpha/2}/2)/(n + z^2_{\alpha/2})$. This is the sample proportion for an adjusted sample that adds $z^2_{\alpha/2}$ observations, half of each type, for example, $z^2_{0.025}/2 = 1.96^2/2 \approx 2$ of each type for 95% intervals. The square of the coefficient of $z_{\alpha/2}$ in (1.14) is a weighted average of the variance of a sample proportion when $\pi = \hat{\pi}$ and the variance of a sample proportion when $\pi = \frac{1}{2}$, using the adjusted sample size $n + z^2_{\alpha/2}$ in place of n.

For 95% confidence, the score interval can be approximated by a simple adjustment of the Wald interval (see Exercise 1.25) that adds 2 observations of each type to the sample before using the Wald formula (1.13). This interval and the ordinary score interval tend to have actual coverage probability much closer to the nominal level than the Wald interval (Agresti and Coull 1998, Agresti and Caffo 2000).

1.4.3 Example: Estimating the Proportion of Vegetarians

To collect data to illustrate concepts in introductory statistics courses, often I have given the students a questionnaire. One year I asked each student in an honors class at the University of Florida whether he or she was a vegetarian. Of $n = 25$ students, $y = 0$ answered "yes." They were not a random sample of a particular population, but we use these data to illustrate 95% confidence intervals for a binomial parameter π.

Since $y = 0$, the ML estimate $\hat{\pi} = 0/25 = 0$. With the Wald method, the 95% confidence interval for π is

$$\hat{\pi} \pm 1.96\sqrt{\hat{\pi}(1-\hat{\pi})/n}, \quad \text{which is} \quad 0 \pm 1.96\sqrt{(0.0 \times 1.0)/25}, \quad \text{or} \quad (0, 0).$$

When a parameter falls near the boundary of the sample space, often sample estimates of standard errors are poor and the Wald method does not provide a sensible answer.

By contrast, the 95% score interval equals (0.0, 0.133). That is, when $\hat{\pi} = 0.0$ and $n = 25$, the two roots for π_0 that satisfy the equation

$$|\hat{\pi} - \pi_0| = 1.96\sqrt{\pi_0(1-\pi_0)/n}$$

are $\pi_0 = 0.0$ and $\pi_0 = 0.133$. This interval provides a more believable inference. It contains the values not rejected in corresponding score tests with size (probability of type I error) 0.05. For H_0: $\pi = 0.20$, for instance, the score test statistic is $z_S = (0 - 0.20)/\sqrt{(0.20 \times 0.80)/25} = -2.50$, which has two-sided P-value $0.012 < 0.05$, so 0.20 does not fall in the interval. By contrast, for H_0: $\pi = 0.10$, $z_S = (0 - 0.10)/\sqrt{(0.10 \times 0.90)/25} = -1.67$, which has P-value $0.096 > 0.05$, so 0.10 falls in the interval.

When $y = 0$ and $n = 25$, the kernel of the likelihood function is $\ell(\pi) = \pi^0(1-\pi)^{25} = (1-\pi)^{25}$. The log-likelihood function (1.7) is $L(\pi) = 25\log(1-\pi)$. Note that $L(\hat{\pi}) = L(0) = 0$. The 95% likelihood-ratio confidence interval is the set of π_0 for which the likelihood-ratio statistic

$$\begin{aligned} -2(L_0 - L_1) &= -2[L(\pi_0) - L(\hat{\pi})] \\ &= -50\log(1-\pi_0) < \chi^2_1(0.05) = 3.84. \end{aligned}$$

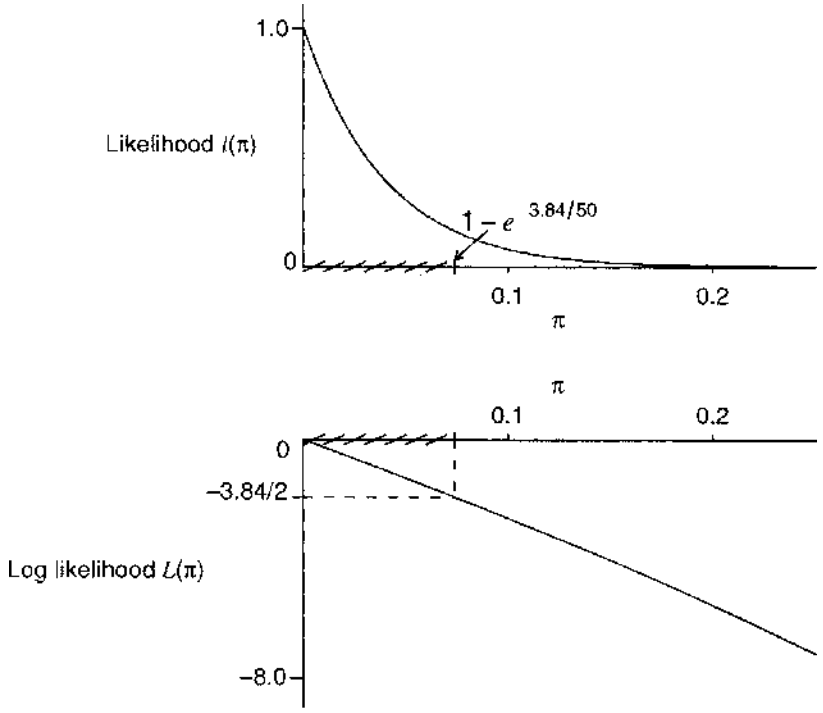

Figure 1.2 Binomial likelihood and log likelihood when $y = 0$ in $n = 25$ trials, and likelihood-ratio test-based confidence interval for π.

The upper bound is $1 - \exp(-3.84/50) = 0.074$, and the confidence interval equals (0.0, 0.074). Figure 1.2 shows the likelihood and log-likelihood functions and the corresponding confidence region for π.

The three large-sample methods yield quite different results. When π is near 0, the sampling distribution of $\hat{\pi}$ is highly skewed to the right for small n. From numerical evaluations, we prefer the interval based on inverting the score test.

1.4.4 Exact Small-Sample Inference and the Mid *P*-Value

With modern computational power, it is not necessary to rely on large-sample approximations for the distribution of estimators such as $\hat{\pi}$. Tests and confidence intervals can directly use the binomial distribution rather than its normal approximation. Such inferences occur naturally for small samples, but apply for any n.

We illustrate by testing H_0: $\pi = 0.50$ against H_a: $\pi \neq 0.50$ for the survey results on vegetarianism just discussed, namely, $y = 0$ with $n = 25$. We noted that the score statistic equals $z = -5.0$. The exact P-value for this statistic, based on the null bin(25, 0.50) distribution, is

$$P(|z| \geq 5.0) = P(Y = 0 \text{ or } Y = 25) = 0.50^{25} + 0.50^{25} = 0.00000006.$$

Because of discreteness, in testing H_0: $\pi = \pi_0$, it is not usually possible to achieve a particular fixed size such as 0.05. With a finite number of possible samples, there is a finite

number of possible P-values, of which 0.05 may not be one. When $n = 25$ and $\pi_0 = 0.50$, for example, the two-sided P-value using the binomial probabilities is 0.043 if $y = 7$ or if $y = 18$ and it is 0.108 if $y = 8$ or if $y = 17$. Thus, if we reject H_0 when $y \leq 7$ or $y \geq 18$, the test is *conservative*, in the sense that the actual size (i.e., 0.043) is less than the nominal size (0.05).

To adjust somewhat for discreteness in small-sample distributions, we can base inference on the *mid P-value* (Lancaster 1949b, 1961). For a test statistic T with observed value t_o and one-sided H_a such that large T contradicts H_0,

$$\text{mid } P\text{-value} = \tfrac{1}{2}P(T = t_o) + P(T > t_o),$$

with probabilities calculated from the null distribution. Thus, the mid P-value is less than the ordinary P-value by half the probability of the observed result. Although discrete, compared with the ordinary P-value, the mid P-value behaves more like the P-value for a test statistic having a continuous distribution: The sum of its two one-sided P-values equals 1.0. Under H_0, it has a null expected value of 0.50 (like the uniform distribution that occurs in the continuous case), whereas this expected value exceeds 0.50 for the ordinary P-value for a discrete test statistic.

Unlike an exact test with ordinary P-value, a test using the mid P-value does not guarantee that the size of the test is no greater than a nominal value (Exercise 1.12). However, it usually performs well. It is less conservative than the ordinary exact test. Inference based on the mid P-value compromises between the conservativeness of exact methods and the uncertain adequacy of large-sample methods.

Similarly, we can use small-sample distributions to construct confidence intervals for parameters. Some subtle issues arise such that the choice of such an interval is not straightforward, and we defer this topic to a special section (16.6) in Chapter 16 about small-sample intervals for categorical data.

1.5 STATISTICAL INFERENCE FOR MULTINOMIAL PARAMETERS

Next we consider inference for multinomial parameters $\{\pi_j\}$. Of n observations in c categories, n_j occur in category j, $j = 1, \ldots, c$.

1.5.1 Estimation of Multinomial Parameters

First, we obtain ML estimates of $\{\pi_j\}$. As a function of $\{\pi_j\}$, the multinomial probability mass function (1.2) is proportional to the kernel

$$\prod_j \pi_j^{n_j}, \quad \text{where all} \quad \pi_j \geq 0 \quad \text{and} \quad \sum_j \pi_j = 1. \tag{1.15}$$

The ML estimates are the $\{\pi_j\}$ that maximize (1.15).

The multinomial log-likelihood function is

$$L(\boldsymbol{\pi}) = \sum_j n_j \log \pi_j.$$

To eliminate redundancies, we treat L as a function of $(\pi_1, \ldots, \pi_{c-1})$, since $\pi_c = 1 - (\pi_1 + \cdots + \pi_{c-1})$. Thus, $\partial\pi_c/\partial\pi_j = -1$, $j = 1, \ldots, c-1$. Since

$$\frac{\partial \log \pi_c}{\partial \pi_j} = \frac{1}{\pi_c}\frac{\partial \pi_c}{\partial \pi_j} = -\frac{1}{\pi_c},$$

differentiating $L(\boldsymbol{\pi})$ with respect to π_j gives the likelihood equation

$$\frac{\partial L(\boldsymbol{\pi})}{\partial \pi_j} = \frac{n_j}{\pi_j} - \frac{n_c}{\pi_c} = 0.$$

The ML solution satisfies $\hat{\pi}_j/\hat{\pi}_c = n_j/n_c$. Now

$$\sum_j \hat{\pi}_j = 1 = \frac{\hat{\pi}_c\left(\sum_j n_j\right)}{n_c} = \frac{\hat{\pi}_c n}{n_c},$$

so $\hat{\pi}_c = n_c/n$ and then $\hat{\pi}_j = n_j/n$. From general results presented later in the book (Section 9.6), this solution does maximize the likelihood. Thus, the ML estimates of $\{\pi_j\}$ are the sample proportions.

1.5.2 Pearson Chi-Squared Test of a Specified Multinomial

In 1900 the eminent British statistician Karl Pearson introduced a hypothesis test that was one of the first inferential methods. It had a revolutionary impact on categorical data analysis. Pearson's test evaluates whether multinomial parameters equal certain values. His original motivation in developing this test was to analyze whether possible outcomes on a particular Monte Carlo roulette wheel were equally likely (Stigler 1986).

Consider H_0: $\pi_j = \pi_{j0}$, $j = 1, \ldots, c$, where $\sum_j \pi_{j0} = 1$. When H_0 is true, the expected values of $\{n_j\}$, called *expected frequencies*, are $\mu_j = n\pi_{j0}$, $j = 1, \ldots, c$. Pearson proposed the test statistic

$$X^2 = \sum_j \frac{(n_j - \mu_j)^2}{\mu_j}. \tag{1.16}$$

Greater differences $|n_j - \mu_j|$ produce greater X^2 values, for fixed $\{\pi_{j0}\}$ and n. Let X_o^2 denote the observed value of X^2. The P-value is the null value of $P(X^2 \geq X_o^2)$. This equals the sum of the null multinomial probabilities of all count arrays (having a sum of n) with $X^2 \geq X_o^2$.

For large samples, X^2 has approximately a chi-squared distribution with $\text{df} = c - 1$. The P-value is approximated by $P(\chi^2_{c-1} \geq X_o^2)$, where χ^2_{c-1} denotes a chi-squared random variable with $\text{df} = c - 1$. Statistic (1.16) is called the *Pearson chi-squared statistic*.

1.5.3 Likelihood-Ratio Chi-Squared Test of a Specified Multinomial

An alternative test for multinomial parameters uses the likelihood-ratio test. The kernel of the multinomial likelihood is (1.15). Under H_0 the likelihood is maximized when $\hat{\pi}_j = \pi_{j0}$.

In the general case, it is maximized when $\hat{\pi}_j = n_j/n$. The ratio of the likelihoods equals

$$\Lambda = \frac{\prod_j (\pi_{j0})^{n_j}}{\prod_j (n_j/n)^{n_j}}.$$

Thus, the likelihood-ratio statistic, denoted by G^2, is

$$G^2 = -2\log\Lambda = 2\sum_j n_j \log(n_j/n\pi_{j0}). \tag{1.17}$$

This statistic, which has form (1.12), is called the *likelihood-ratio chi-squared statistic*. The larger the value of G^2, the greater the evidence against H_0.

In the general case, the parameter space consists of $\{\pi_j\}$ subject to $\sum_j \pi_j = 1$, so the dimensionality is $c - 1$. Under H_0, the $\{\pi_j\}$ are specified completely, so the dimension is 0. The difference in these dimensions equals $(c - 1)$. For large n, G^2 has a chi-squared null distribution with $\text{df} = c - 1$.

When H_0 holds, the Pearson X^2 and the likelihood ratio G^2 both have large-sample chi-squared distributions with $\text{df} = c - 1$. In fact, they are asymptotically equivalent in that case; specifically, $X^2 - G^2$ converges in probability to zero. [This means that for any $\epsilon > 0$, $P(|X^2 - G^2| < \epsilon) \to 1$ as $n \to \infty$; See Section 16.3.4.] When H_0 is false, X^2 and G^2 grow in expectation proportionally to n; they need not take similar values, however, even for very large n.

For fixed c, as n increases the distribution of X^2 usually converges to chi-squared more quickly than that of G^2. The chi-squared approximation is often poor for G^2 when $n/c < 5$. When c is large, it can be decent for X^2 for n/c as small as 1 if the table does not contain both very small and moderately large expected frequencies.

Alternatively, the multinomial probabilities induce exact distributions of these test statistics. When it is not feasible to quickly enumerate all the possible samples, it is simple to simulate the exact distributions by randomly generating a very large number of multinomial samples of size n with the null probabilities, and calculating X^2 and or G^2 for each sample (Hirji 2005, Chap. 13). The simulated P-value is the proportion of test statistic values that are at least as large as the observed value.

1.5.4 Example: Testing Mendel's Theories

Among its many applications, Pearson's test was used in genetics to test Mendel's theories of natural inheritance. Mendel crossed pea plants of pure yellow strain with plants of pure green strain. He predicted that second-generation hybrid seeds would be 75% yellow and 25% green, yellow being the dominant strain. One experiment produced $n = 8023$ seeds, of which $n_1 = 6022$ were yellow and $n_2 = 2001$ were green. The expected frequencies for H_0: $\pi_{10} = 0.75$, $\pi_{20} = 0.25$ are $\mu_1 = 8023(0.75) = 6017.25$ and $\mu_2 = 2005.75$. The Pearson statistic $X^2 = 0.015$ and the likelihood-ratio statistic $G^2 = 0.015$ (df = 1) have P-values of $P = 0.90$. They do not contradict Mendel's hypothesis.

When $c = 2$, Pearson's X^2 simplifies to the square of the normal score statistic (1.11). For Mendel's data, $\hat{\pi}_1 = 6022/8023$, $\pi_{10} = 0.75$, $n = 8023$, and $z_S = 0.123$, for which $X^2 = (0.123)^2 = 0.015$. In fact, for general c the Pearson test is the score test about specified values for multinomial parameters.

Mendel performed several experiments of this type. In 1936, R. A. Fisher summarized Mendel's results. He used the reproductive property of chi-squared: If $X_1^2, \ldots, X_k^2$ are independent chi-squared statistics with degrees of freedom $\nu_1, \ldots, \nu_k$, then $\sum_{i=1}^k X_i^2$ has a chi-squared distribution with df $= \sum_{i=1}^k \nu_i$. Fisher obtained a summary chi-squared statistic equal to 42, with df $= 84$. A chi-squared distribution with df $= 84$ has mean 84 and standard deviation $(2 \times 84)^{1/2} = 13.0$, and the right-tailed probability above 42 is $P = 0.99996$. In other words, the chi-squared statistic was so small that the fit seemed *too* good.

Fisher commented: "The general level of agreement between Mendel's expectations and his reported results shows that it is closer than would be expected in the best of several thousand repetitions . . .. I have no doubt that Mendel was deceived by a gardening assistant, who knew only too well what his principal expected from each trial made." In a letter written at the time, he stated: "Now, when data have been faked, I know very well how generally people underestimate the frequency of wide chance deviations, so that the tendency is always to make them agree too well with expectations" (Box 1978, p. 297). In summary, goodness-of-fit tests can reveal not only when a fit is inadequate, but also when it is better than random fluctuations would have us expect. [Fisher's daughter, Joan Fisher Box (1978, pp. 295–300), discussed Fisher's analysis of Mendel's data and the accompanying controversy. See also Pires and Branco (2010). Despite possible difficulties with Mendel's data, subsequent work led to general acceptance of his theories.]

1.5.5 Testing with Estimated Expected Frequencies

The chi-squared statistics (1.16) and (1.17) compare a sample distribution to a hypothetical one $\{\pi_{j0}\}$. In some applications, $\{\pi_{j0} = \pi_{j0}(\boldsymbol{\theta})\}$ are functions of a smaller set of unknown parameters $\boldsymbol{\theta}$. ML estimates $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ determine ML estimates $\{\pi_{j0}(\hat{\boldsymbol{\theta}})\}$ of $\{\pi_{j0}\}$ and hence ML estimates $\{\hat{\mu}_j = n\pi_{j0}(\hat{\boldsymbol{\theta}})\}$ of expected frequencies.

Replacing $\{\mu_j\}$ by estimates $\{\hat{\mu}_j\}$ affects the distribution of X^2 and G^2. When $\dim(\boldsymbol{\theta}) = p$, the true df $= (c-1) - p$ (Section 16.3.3). Pearson (1917) realized this but did not always take it into account (Section 17.2).

1.5.6 Example: Pneumonia Infections in Calves

We now show a goodness-to-fit test with estimated expected frequencies. A sample of 156 dairy calves born in Okeechobee County, Florida, were classified according to whether they caught pneumonia within 60 days of birth. Calves that got a pneumonia infection were also classified according to whether they got a secondary infection within 2 weeks after the first infection cleared up. Table 1.1 shows the data. Calves that did not get a primary infection

Table 1.1 Primary and Secondary Pneumonia Infections in Calves

	Secondary Infection[a]	
Primary Infection	Yes	No
Yes	30 (38.1)	63 (39.0)
No	0 (—)	63 (78.9)

[a] Values in parentheses are estimated expected frequencies.
Source: Data courtesy of Thang Tran and G. A. Donovan, College of Veterinary Medicine, University of Florida.

Table 1.2 Probability Structure for Hypothesis

	Secondary Infection		
Primary Infection	Yes	No	Total
Yes	π^2	$\pi(1-\pi)$	π
No	—	$1-\pi$	$1-\pi$

could not get a secondary infection, so no observations can fall in the category for "no" primary infection and "yes" secondary infection. That combination is called a *structural zero*.

A goal of this study was to test whether the probability of primary infection was the same as the conditional probability of secondary infection, given that the calf got the primary infection. In other words, if π_{ab} denotes the probability that a calf is classified in row a and column b of this table, the null hypothesis is

$$H_0: \pi_{11} + \pi_{12} = \pi_{11}/(\pi_{11} + \pi_{12})$$

or $\pi_{11} = (\pi_{11} + \pi_{12})^2$. Let $\pi = \pi_{11} + \pi_{12}$ denote the probability of primary infection. The null hypothesis states that the probabilities satisfy the structure that Table 1.2 shows; that is, probabilities in a trinomial for the categories (yes–yes, yes–no, no–no) for primary–secondary infection equal $[\pi^2, \pi(1-\pi), 1-\pi]$.

Let n_{ab} denote the number of observations in row a and column b of Table 1.1. The ML estimate of π is the value maximizing the kernel of the multinomial likelihood

$$(\pi^2)^{n_{11}}(\pi - \pi^2)^{n_{12}}(1-\pi)^{n_{22}}.$$

The log likelihood is

$$L(\pi) = n_{11}\log\pi^2 + n_{12}\log(\pi - \pi^2) + n_{22}\log(1-\pi).$$

Differentiation with respect to π gives the likelihood equation

$$\frac{2n_{11}}{\pi} + \frac{n_{12}}{\pi} - \frac{n_{12}}{1-\pi} - \frac{n_{22}}{1-\pi} = 0.$$

The solution is

$$\hat{\pi} = (2n_{11} + n_{12})/(2n_{11} + 2n_{12} + n_{22}).$$

For Table 1.1, $\hat{\pi} = 0.494$. Since $n = 156$, the estimated expected frequencies are $\hat{\mu}_{11} = n\hat{\pi}^2 = 38.1$, $\hat{\mu}_{12} = n(\hat{\pi} - \hat{\pi}^2) = 39.0$, and $\hat{\mu}_{22} = n(1-\hat{\pi}) = 78.9$. Table 1.1 shows them. Pearson's statistic is $X^2 = 19.7$. Since the $c = 3$ possible responses have $p = 1$ parameter (π) determining the expected frequencies, df $= (3-1) - 1 = 1$. There is strong evidence against H_0 ($P = 0.00001$). Inspection of Table 1.1 reveals that many more calves got a primary infection but not a secondary infection than H_0 predicts. The researchers concluded that the primary infection had an immunizing effect that reduced the likelihood of a secondary infection.

1.5.7 Chi-Squared Theoretical Justification

We now outline why Pearson's statistic for a specified multinomial has a limiting chi-squared distribution. Derivations for the likelihood-ratio statistic and cases with estimated expected frequencies are given in Section 16.3.

For a multinomial sample $(n_1, \ldots, n_c)$ of size n, the marginal distribution of n_j is the bin(n, π_j) distribution. For large n, by the normal approximation to the binomial, n_j (and $\hat{\pi}_j = n_j/n$) have approximate normal distributions. More generally, by the central limit theorem, the sample proportions $\hat{\boldsymbol{\pi}} = (n_1/n, \ldots, n_{c-1}/n)^T$ have an approximate multivariate normal distribution (Section 16.1.4). Let $\boldsymbol{\Sigma}_0$ denote the null covariance matrix of $\sqrt{n}\,\hat{\boldsymbol{\pi}}$, and let $\boldsymbol{\pi}_0 = (\pi_{10}, \ldots, \pi_{c-1,0})^T$. Under H_0, since $\sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)$ converges to a $N(\mathbf{0}, \boldsymbol{\Sigma}_0)$ distribution, the quadratic form

$$n(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)^T \boldsymbol{\Sigma}_0^{-1} (\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) \tag{1.18}$$

has distribution converging to chi-squared with df $= c - 1$.

In Section 16.1.4 we show that the covariance matrix of $\sqrt{n}\hat{\boldsymbol{\pi}}$ has elements

$$\sigma_{jk} = \begin{cases} -\pi_j \pi_k & \text{if } j \neq k \\ \pi_j(1 - \pi_j) & \text{if } j = k \end{cases}.$$

The matrix $\boldsymbol{\Sigma}_0^{-1}$ has (j, k)th element $1/\pi_{c0}$ when $j \neq k$ and $(1/\pi_{j0} + 1/\pi_{c0})$ when $j = k$. (You can verify this by showing that $\boldsymbol{\Sigma}_0\ \boldsymbol{\Sigma}_0^{-1}$ equals the identity matrix.) With this substitution, direct calculation with appropriate combining of terms yields that (1.18) simplifies to X^2. In Section 16.3 we provide a formal proof in a more general setting.

This argument is similar to Pearson's in 1900. R. A. Fisher (1922) gave a simpler justification, the gist of which follows: Suppose that $(n_1, \ldots, n_c)$ are independent Poisson random variables with means $(\mu_1, \ldots, \mu_c)$. For large $\{\mu_j\}$, the standardized values $\{z_j = (n_j - \mu_j)/\sqrt{\mu_j}\}$ have approximate standard normal distributions. Thus, $\sum_j z_j^2 = X^2$ has an approximate chi-squared distribution with c degrees of freedom. Adding the single linear constraint $\sum_j (n_j - \mu_j) = 0$, thus converting the Poisson distributions to a multinomial, we lose a degree of freedom.

1.6 BAYESIAN INFERENCE FOR BINOMIAL AND MULTINOMIAL PARAMETERS

This book mainly uses the traditional, so-called *frequentist*, approach to statistical inference. We regard parameter values as fixed and apply probability statements to possible values for the data, given the parameter values. Recent years have seen increasing popularity of the *Bayesian* approach, which has probability distributions for parameters as well as for data. This yields inferences in the form of probability statements about possible values for the parameters, given the data.

1.6.1 The Bayesian Approach to Statistical Inference

The Bayesian approach assumes a *prior distribution* for the parameters. This probability distribution may reflect subjective prior beliefs. Or, it may reflect information about the

parameter values from other studies. Or, it may be relatively uninformative, so that inferential results are based almost entirely on the current data. The prior distribution combines with the information that the data provide to generate a *posterior distribution* for the parameters. Different choices for the prior distribution can result in quite different posterior inferences, especially for small sample sizes, so the choice should be given careful thought.

By Bayes' theorem, the posterior probability density function h of a parameter $\boldsymbol{\theta}$, given the data $\boldsymbol{y}$, relates to the probability mass function f for $\boldsymbol{y}$, given $\boldsymbol{\theta}$, and the prior density function g for $\boldsymbol{\theta}$, by

$$h(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta})g(\boldsymbol{\theta})}{f(\boldsymbol{y})}.$$

The denominator $f(\boldsymbol{y})$ on the right-hand side is the marginal probability mass function of the data, that is, $\int_{\boldsymbol{\Theta}} f(\boldsymbol{y} \mid \boldsymbol{\theta})g(\boldsymbol{\theta})d\boldsymbol{\theta}$. This is a constant with respect to $\boldsymbol{\theta}$, so irrelevant for inference about $\boldsymbol{\theta}$. When we plug in the observed data, $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ is the likelihood function when viewed as a function of $\boldsymbol{\theta}$. So, the prior density function for $\boldsymbol{\theta}$ multiplied by the likelihood function determines the posterior density for $\boldsymbol{\theta}$.

Except in specialized cases such as presented in Sections 1.6.2 and 1.6.3, there is not a closed-form expression for the posterior distribution. The difficulty is in finding the denominator integral that determines $f(\boldsymbol{y})$. The key part of the Bayes equation is the numerator, because of the proportionality in terms of $\boldsymbol{\theta}$,

$$h(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto f(\boldsymbol{y} \mid \boldsymbol{\theta})g(\boldsymbol{\theta}).$$

Simulation methods are used to approximate the posterior distribution. The primary method for doing this is Markov chain Monte Carlo (MCMC). It is beyond our scope to discuss the technical details of how an MCMC algorithm works. In a nutshell, a stochastic process of Markov chain form is designed so that its long-run stationary distribution is the posterior distribution. One or more such Markov chains provide a very large number of simulated values from the posterior distribution, and the distribution of the simulated values approximates the posterior distribution. Enough observations are taken after a burn-in period so that the Monte Carlo error is small in approximating the posterior distribution and summary measures of interest for that distribution, such as the mean and standard deviation, certain percentiles, and intervals formed using those percentiles.

For an arbitrary parameter β, such as a coefficient in a regression-type model, Bayesian methods of inference using the posterior distribution parallel those for frequentist inference. For example, in lieu of P-values, posterior tail probabilities are useful. Information about the direction of an effect is contained in the posterior probabilities $P(\beta > 0 \mid \boldsymbol{y})$ and $P(\beta < 0 \mid \boldsymbol{y})$. With a flat prior distribution, $P(\beta < 0 \mid \boldsymbol{y})$ corresponds to the frequentist P-value for the one-sided test with H_a: $\beta > 0$.

Analogous to the frequentist confidence interval is an interval that contains most of the posterior distribution. Such an interval is referred to as a *posterior interval* or *credible interval*. A common approach for constructing a posterior interval uses percentiles of the posterior distribution, with equal probabilities in the two tails. For example, the 95% equal-tail posterior interval for β is the region between the 2.5 and 97.5 percentiles of the posterior distribution for β. For unimodal posteriors, an alternative Bayesian *highest posterior density* (HPD) interval has higher posterior density for every value inside the interval than for every value outside it, subject to the posterior probability over the interval

equaling the desired confidence level. This method produces the shortest possible interval with the given level.

We next summarize the Bayesian approach for binomial and multinomial parameters. Then, in the rest of the book, we'll occasionally present Bayesian alternatives to frequentist model-based inference.

1.6.2 Binomial Estimation: Beta and Logit-Normal Prior Distributions

The simplest Bayesian inference for a binomial parameter π uses a member of the *beta distribution* as the prior distribution. The beta(α_1, α_2) probability density function for π is proportional to

$$\pi^{\alpha_1 - 1}(1 - \pi)^{\alpha_2 - 1}.$$

The parameters $\alpha_1 > 0$ and $\alpha_2 > 0$ of the prior are often referred to as *hyperparameters*, to distinguish them from the parameter that is the object of inference (in this case, π). The beta distribution has

$$E(\pi) = \alpha_1/(\alpha_1 + \alpha_2) \quad \text{and} \quad \text{var}(\pi) = \alpha_1\alpha_2/[(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)].$$

The family of beta probability density functions has a wide variety of shapes over the interval (0, 1), including uniform when $\alpha_1 = \alpha_2 = 1$, unimodal symmetric ($\alpha_1 = \alpha_2 > 1$), unimodal skewed left ($\alpha_1 > \alpha_2 > 1$), unimodal skewed right ($\alpha_2 > \alpha_1 > 1$), and bimodal U-shaped ($\alpha_1 < 1, \alpha_2 < 1$).

Often prior knowledge about π can be expressed in terms of a mean and standard deviation for a prior for π. Then, the one-to-one correspondence between those moments and (α_1, α_2) based on the above moment expressions determines a beta prior. By contrast, lack of prior knowledge about π might suggest using a uniform prior distribution. The posterior distribution then has the same shape as the binomial likelihood function. Alternatively, a popular prior distribution with Bayesians is the *Jeffreys prior*, which is proportional to the square root of the determinant of the Fisher information matrix for the parameters of interest. With this approach, prior distributions for different scales of measurement for the parameters (e.g., for π or for $\phi = \log[\pi/(1 - \pi)]$) are equivalent. For a binomial parameter, the Jeffreys prior is the beta distribution with $\alpha_1 = \alpha_2 = 0.5$.

The beta distribution is the *conjugate prior distribution* for inference about a binomial parameter. This means that it is the family of probability distributions such that, when combined with the likelihood function, the posterior distribution falls in the same family. When we combine a beta(α_1, α_2) prior distribution with a binomial likelihood function, the posterior distribution is a beta$(y + \alpha_1, n - y + \alpha_2)$ distribution, for which the mean is

$$\frac{y + \alpha_1}{n + \alpha_1 + \alpha_2} = \left(\frac{n}{n + \alpha_1 + \alpha_2}\right)\hat{\pi} + \left(\frac{\alpha_1 + \alpha_2}{n + \alpha_1 + \alpha_2}\right)\frac{\alpha_1}{\alpha_1 + \alpha_2}.$$

This is a weighted average of the sample proportion $\hat{\pi} = y/n$ and the prior mean, with more weight given the sample proportion as n increases. Conjugate priors were the primary method of conducting Bayesian analysis before the development of computationally intensive methods, such as Markov chain Monte Carlo, for evaluating the integral that determines the posterior distribution.

An alternative prior distribution assumes a normal distribution for the *logit* parameter, $\log[\pi/(1-\pi)]$. This parameter, which is relevant for many analyses presented in this book, takes values over the entire real line. With a $N(0, \sigma^2)$ prior distribution for $\log[\pi/(1-\pi)]$, on the π scale the shape of this *logit-normal* (also called *logistic-normal*) density is symmetric[3], being unimodal when $\sigma^2 \leq 2$ and bimodal when $\sigma^2 > 2$, but always tapering off toward 0 as π approaches 0 or 1. Specifically, it is mound-shaped for small σ, roughly uniform except near the boundaries when $\sigma \approx 1.5$, and with more pronounced peaks for the modes when σ is about 2 or larger. The peaks for the modes get closer to 0 and 1 as σ increases further, and the curve has appearance that is essentially U-shaped when $\sigma = 3$ and similar to that of a beta(0.5, 0.5) prior. For $\sigma = (1, 2, 3)$, the standard deviations on the π scale of these priors are (0.21, 0.31, 0.37), similar to the values (0.22, 0.29, 0.35) for the beta priors with $\alpha_1 = \alpha_2 = (2.0, 1.0, 0.5)$. The logit-normal prior with $\sigma = 2.67$ matches the Jeffreys prior in the first two moments (on the probability scale), and the logit-normal prior with $\sigma = 1.69$ matches the uniform prior in the first two moments. With a $N(\mu, \sigma^2)$ prior distribution for the logit, the density for π is skewed left when $\mu > 0$ and skewed right when $\mu < 0$.

Yet another possibility, hierarchical in nature, uses beta or logit-normal priors but assumes a distribution for their hyperparameters instead of assigning fixing values. That second-stage distribution may have its own hyperparameters. See Section 3.6.7, Albert (2010), Good (1965), and Leonard (1972).

1.6.3 Multinomial Estimation: Dirichlet Prior Distributions

For $c > 2$ categories, the beta distribution generalizes to the *Dirichlet distribution*. It is defined over the simplex of nonnegative values $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_c)$ that sum to 1. Expressed in terms of gamma functions and c hyperparameters $\{\alpha_i > 0\}$, the Dirichlet probability density function is

$$g(\boldsymbol{\pi}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\left[\prod_i \Gamma(\alpha_i)\right]} \prod_{i=1}^{c} \pi_i^{\alpha_i - 1} \quad \text{for } 0 < \pi_i < 1 \text{ all } i, \quad \sum_i \pi_i = 1.$$

The case $\{\alpha_i = 1\}$ is the uniform density over the possible probability values. The case $\left\{\alpha_i = \frac{1}{2}\right\}$ is the *Jeffreys prior* for multinomial parameters. Let $K = \sum_i \alpha_i$. The Dirichlet distribution has $E(\pi_i) = \alpha_i/K$ and $\text{var}(\pi_i) = \alpha_i(K - \alpha_i)/[K^2(K+1)]$. For particular relative sizes of $\{\alpha_i\}$, such as identical values, the distribution is more tightly concentrated around the means as K increases.

Let $\boldsymbol{n} = (n_1, \ldots, n_c)$ denote cell counts from $n = \sum_i n_i$ independent observations with cell probabilities $\boldsymbol{\pi}$. Formula (1.2) showed the multinomial probability mass function for $\boldsymbol{n}$. Multiplying this by the Dirichlet prior density function $g(\boldsymbol{\pi})$ contributes to a posterior density function $h(\boldsymbol{\pi} \mid \boldsymbol{n})$ for $\boldsymbol{\pi}$ that is also Dirichlet, but with the hyperparameters $\{\alpha_i\}$ replaced by $\{\alpha_i' = n_i + \alpha_i\}$. The mean of the posterior distribution of π_i is

$$E(\pi_i \mid n_1, \ldots, n_c) = (n_i + \alpha_i)/(n + K).$$

[3]See `logitnorm.r-forge.r-project.org` and the "Logit-normal distribution" entry in `wikipedia.org` for figures illustrating the shapes described below.

Let $\gamma_i = E(\pi_i) = \alpha_i/K$. This Bayesian estimator equals the weighted average

$$\left(\frac{n}{n+K}\right)p_i + \left(\frac{K}{n+K}\right)\gamma_i \tag{1.19}$$

of the sample proportion $p_i = n_i/n$ and the mean γ_i of the prior distribution for π_i. This posterior mean takes the form of a sample proportion when the prior information corresponds to K additional observations of which α_i were outcomes of type i. (We'll consider a formal way of setting such *data augmentation priors* in Section 7.2.4.) With identical $\{\alpha_i\}$, the Bayes estimate shrinks each sample proportion toward the equi-probability value $\gamma_i = 1/c$. Greater shrinkage occurs as K increases, for fixed n.

Bayesian estimators of multinomial parameters, unlike the sample proportions, are slightly biased for finite n. Usually, though, they have smaller total mean squared error (MSE) than the sample proportions. They are not uniformly better for all possible parameter values, however. For instance, if a particular $\pi_i = 0$, then $p_i = 0$ with probability one, so the sample proportion is then better than any other estimator. We do not expect $\pi_i = 0$ in practice, and the parameter space is often defined under the restriction that all $\pi_i > 0$, but this limiting behavior explains why the ML estimator can have smaller MSE than the Bayes estimator when π_i is very near 0.

1.6.4 Example: Estimating Vegetarianism Revisited

In Section 1.4.3 we estimated the population proportion of vegetarians with a sample of size $n = 25$ for which $y = 0$. The ML estimate of π is $\hat{\pi} = 0.0$, and the 95% score confidence interval is (0.0, 0.133). How does this compare to Bayesian point and interval estimates?

First, we use a uniform prior distribution for π, reflecting prior ignorance. For this beta(1, 1) prior with $y = 0$ and $n = 25$, the posterior distribution is beta(1, 26). The posterior mean is 1/27 = 0.037. The posterior 95% equal-tail interval is (0.001, 0.132), the endpoints being the 2.5 and 97.5 percentiles of the beta posterior density. This interval is similar to the frequentist 95% score interval, but the prior information has the impact of moving the left boundary slightly away from 0.0. By contrast, since the posterior density is proportional to $(1-\pi)^{25}$ and hence monotone decreasing, the 95% highest posterior density (HPD) interval has lower limit of 0 and upper limit that is the 95th percentile of the beta(1, 26) density, which is 0.109.

For contrast, let's use a much more informative beta prior. Suppose we used a subjective approach and were quite sure *a priori* that π falls between about 0 and 0.16. We might summarize this by a prior mean of 0.08 and standard deviation of 0.04. These moments correspond to beta hyperparameters of $\alpha_1 = 3.6$ and $\alpha_2 = 41.4$, for which 0.16 is the 96th percentile. Then, the posterior is the beta(3.6, 66.4), which has mean = 0.051 and 95% posterior equal-tail interval of (0.013, 0.114) and HPD interval of (0.008, 0.103). Stronger prior beliefs result in greater shrinkage of the Bayes estimate toward the prior mean and a narrower posterior equal-tail interval.

1.6.5 Binomial and Multinomial Estimation: Improper Priors

For multinomial data, the sample proportion p_i is the ML estimate of π_i. It results as the special case of the Bayesian estimate (1.19) when each $\alpha_i = 0$. But when any $\alpha_i = 0$, the

Dirichlet formula is not a legitimate probability density function, as it integrates to ∞ instead of 1. It is then an example of an *improper prior distribution*. Bayesian inference sometimes uses such improper prior distributions, as long as the posterior distribution is proper (e.g., Lindley 1964). The Dirichlet posterior is proper as long as $n_i > 0$ for each i having $\alpha_i = 0$.

For parameters that can take value over the entire real line, a common improper distribution is uniform over all real numbers. For a binomial parameter π, the improper beta(0,0) prior for π corresponds to an improper uniform distribution for logit(π). Haldane (1948) suggested that this prior is often sensible in genetics applications, such as for mutation rates for which $\log(\pi)$ might be approximately uniform for π close to 0.

NOTES

Section 1.1: Categorical Response Data

1.1 **Measurement scales**: Stevens (1951) defined (nominal, ordinal, interval) scales of measurement. Other scales result from mixtures of these types. For instance, *partially ordered* scales occur when subjects respond to questions having categories that are ordered except for don't know or undecided categories.

Section 1.3: Statistical Inference for Categorical Data

1.2 **Chi-squared**: Greenwood and Nikulin (1996), Kendall and Stuart (1979), and Lancaster (1969) presented in-depth overviews of the chi-squared distribution. Cochran (1952) presented a historical survey of chi-squared tests of fit. See also Cressie and Read (1989), Koch and Bhapkar (1982), Koehler (2005), Moore (1986b), Read and Cressie (1988), and Watson (1959).

1.3 **Wald/LR/score**: Disadvantages of the Wald method compared with the score and likelihood-ratio methods is that it does not apply when $\hat{\beta}$ is on the boundary of the parameter space (such as a sample proportion $\hat{\pi} = 0$) and its results depend on the parameterization; inference based on $\hat{\beta}$ and its *SE* is not equivalent to inference based on a nonlinear function of it, such as $\log(\hat{\beta})$ and its *SE*. See Section 5.2.6. "Higher-order asymptotics" improve on simple normal and chi-squared approximations for distributions of these statistics (Brazzale et al. 2007, Davison et al. 2006).

Section 1.4: Statistical Inference for Binomial Parameters

1.4 **Score CI**: The superiority of the score interval to the Wald interval for π was shown by, among others, Agresti and Coull (1998), Blyth and Still (1983), Brown et al. (2001), Ghosh (1979), Newcombe (1998a), and Schader and Schmid (1990).

1.5 **Continuity correction**: Using continuity corrections with large-sample methods provides approximations to exact small-sample methods. We do not present them, since if you prefer an exact method, with modern computational power you can usually implement it directly rather than approximate it. However, we'll see in Sections 3.5.5, 3.5.7, 7.3.7, 16.6.1, and 16.6.4 that exact methods have the disadvantage that they behave conservatively.

1.6 **Discreteness**: Suppose a statistic T has discrete distribution with cdf $F(t)$. Then, $F(T)$ is *stochastically larger* than uniform over [0, 1], its cdf being everywhere no greater than that of the uniform (Casella and Berger 2001, pp. 77, 434). Likewise, a P-value based on T has null distribution stochastically larger than uniform. In theory, we can eliminate issues with discreteness in tests by performing a supplementary randomization on the boundary of a

critical region (see Exercise 1.12). In rejecting H_0 at the boundary with a certain probability, we can obtain type I error probability = α even when α is not an achievable P-value. For such randomization, the P-value is

$$\text{randomized } P\text{-value} = U \times P(T = t_o) + P(T > t_o),$$

where U denotes a uniform (0, 1) random variable (Stevens 1950). In practice, this is not done, as it is absurd to let a random number determine a decision. The mid P-value replaces the arbitrary uniform multiple $U \times P(T = t_o)$ by its expected value $0.50 \times P(T = t_o)$.

Section 1.5: Statistical Inference for Multinomial Parameters

1.7 **Multinomials**: Other references on testing a specified multinomial include Good et al. (1970) and Baglivo et al. (1992). For simultaneous confidence intervals for multinomial parameters and their differences, see Exercise 1.36, Chafaï (2009), Fitzpatrick and Scott (1987), Goodman (1965), and Sison and Glaz (1995).

Section 1.6: Bayesian Inference for Binomial and Multinomial Parameters

1.8 **Beta/Dirichlet priors**: Agresti and Hitchcock (2005) surveyed Bayesian methods for categorical data. Lindley (1964) and Good (1965) were influential early articles about Bayesian estimation of multinomial parameters using a Dirichlet prior. Brown et al. (2001) showed that the Jeffreys beta prior yields posterior intervals for the binomial parameter that perform well, having actual coverage probability close to the nominal level. Good (1967) gave a Bayesian goodness-of-fit test that multinomial probabilities are identical, using a hierarchical approach with a symmetric Dirichlet prior that has a log Cauchy distribution for its hyperparameter.

1.9 **Loss functions**: In decision-theoretic terms, the Bayes estimator minimizes the posterior expected value of a loss function that measures the distance between an estimator $T(\boldsymbol{y})$ and a parameter θ. It is the posterior mean for squared error loss and posterior median for absolute error loss. For loss function $w(\theta)(T - \theta)^2$, it is $E[\theta w(\theta)|\boldsymbol{y}]/E[w(\theta)|\boldsymbol{y}]$. With loss function $(T - \pi)^2/[\pi(1 - \pi)]$ and uniform prior, the Bayes estimator of π is the ML estimator $p = y/n$. Its risk function (the expected loss, treated as a function of π) is constant. Bayes estimators with constant risk are *minimax*, the maximum risk being no greater than the maximum risk for any other estimator. Johnson (1971) showed that p is an admissible estimator, for standard loss functions. For other cases, see DasGupta and Zhang (2004). Blyth (1980) noted that for large n, $E|\hat{\pi} - \pi| \approx \sqrt{2\pi(1 - \pi)/\pi_c n}$, where $\pi_c = 3.14\ldots$ is the mathematical constant.

EXERCISES

Applications

1.1 Identify each variable as nominal, ordinal, or interval.

a. UK political party preference (Labour, Liberal Democrat, Conservative)

b. Anxiety rating (none, mild, moderate, severe, very severe)

c. Patient survival (in number of months)

d. Clinic location (London, Boston, Madison, Rochester, Montreal)

e. Response of tumor to chemotherapy (complete elimination, partial reduction, stable, growth progression)

f. Favorite grocery store for UK residents (Sainsbury, Tesco, Waitrose, other)

1.2 Each of 100 multiple-choice questions on an exam has four possible answers, one of which is correct. For each question, a student guesses by selecting an answer randomly.

a. Specify the distribution of the number of correct answers.

b. Find the mean and standard deviation of that distribution. Would it be surprising if the student made at least 50 correct responses? Why?

c. Specify the distribution of (n_1, n_2, n_3, n_4), where n_j is the number of times the student picked choice j.

d. Find $E(n_j)$ and $\text{var}(n_j)$. Show that $\text{cov}(n_j, n_k) = -6.25$ and $\text{corr}(n_j, n_k) = -0.333$.

1.3 An experiment studies the number of insects that survive a certain dose of an insecticide, using several batches of insects of size n each. The insects are sensitive to factors that vary among batches during the experiment but were not measured, such as temperature level. Explain why the distribution of the number of insects per batch surviving the experiment might show overdispersion relative to a $\text{bin}(n, \pi)$ distribution.

1.4 In his autobiography *A Sort of Life*, British author Graham Greene described a period of severe mental depression during which he played Russian roulette. This "game" consists of putting a bullet in one of the six chambers of a pistol, spinning the chambers to select one at random, and then firing the pistol once at one's head.

a. Greene played this game six times and was lucky that none of them resulted in a bullet firing. Find the probability of this outcome.

b. Suppose that he had kept playing this game until the bullet fired. Let Y denote the number of the game on which it fires. Explain why the probability mass function for Y is the *geometric*, $p(y) = (5/6)^{y-1}(1/6)$, $y = 1, 2, 3, \ldots$.

1.5 When the 2010 General Social Survey asked, "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if she is married and does not want any more children," 587 replied "yes" and 636 replied "no." Let π denote the population proportion who would reply "yes." Find the P-value for testing H_0: $\pi = 0.50$ using the score test, and construct a 95% confidence interval for π. Interpret the results.

1.6 Refer to the vegetarianism example in Section 1.4.3. For testing H_0: $\pi = 0.50$ against H_a: $\pi \neq 0.50$, show that:

a. The likelihood-ratio statistic equals $2[25 \log(25/12.5)] = 34.7$.

b. The chi-squared form of the score statistic equals 25.0.

c. The Wald z or chi-squared statistic is infinite.

1.7 In a crossover trial comparing a new drug to a standard, π denotes the probability that the new one is judged better. It is desired to estimate π and test H_0: $\pi = 0.50$ against H_a: $\pi \neq 0.50$. In 20 independent observations, the new drug is better each time.

a. Find and sketch the likelihood function. Is it close to the quadratic shape that large-sample normal approximations utilize?

b. Give the ML estimate of π. Conduct a Wald test and construct a 95% Wald confidence interval for π. Are these sensible?

c. Conduct a score test, reporting the P-value. Construct a 95% score confidence interval. Interpret.

d. Conduct a likelihood-ratio test and construct a likelihood-based 95% confidence interval. Interpret.

e. Construct an exact binomial test. Interpret.

1.8 Refer to the previous exercise. Suppose you wanted a large enough sample to estimate the probability of preferring the new drug to within 0.05, with confidence 0.95. If the true probability is 0.80, about how large a sample is needed?

1.9 In an experiment on chlorophyll inheritance in maize, for 1103 seedlings of self-fertilized heterozygous green plants, 854 seedlings were green and 249 were yellow. Theory predicts the ratio of green to yellow is 3:1. Test the hypothesis that 3:1 is the true ratio. Report the P-value, and interpret.

1.10 Table 1.3 contains Ladislaus von Bortkiewicz's data on deaths of soldiers in the Prussian army from kicks by army mules (Fisher 1934, Quine and Seneta 1987). The data refer to 10 army corps, each observed for 20 years. In 109 corps-years of exposure, there were no deaths, in 65 corps-years there was one death, and so on. Estimate the mean and test whether probabilities of occurrences in these five categories follow a Poisson distribution (truncated for 4 and above).

1.11 A binomial experiment tests H_0: $\pi = 0.50$ against H_a: $\pi \neq 0.50$ using significance level 0.05. Only $n = 5$ observations are available. Show that the true null probability of rejecting H_0 is 0.00 for an exact binomial test and $\frac{1}{16}$ using the large-sample score test.

1.12 A researcher routinely tests using a nominal $P(\text{type I error}) = 0.05$, rejecting H_0 if the P-value ≤ 0.05. An exact test using test statistic T has null distribution $P(T = 0) = 0.30$, $P(T = 1) = 0.62$, and $P(T = 2) = 0.08$, where a higher T provides more evidence against the null.

Table 1.3 Data on Deaths by Mule Kicks, for Exercise 1.10

Number of Deaths	Number of Corps-Years
0	109
1	65
2	22
3	3
4	1
≥ 5	0

a. With the usual P-value, show that the actual $P(\text{type I error}) = 0$.

b. With the mid P-value, show that the actual $P(\text{type I error}) = 0.08$.

c. Find $P(\text{type I error})$ in parts (a) and (b) when $P(T = 0) = 0.30$, $P(T = 1) = 0.66$, $P(T = 2) = 0.04$. Note that the test with mid P-value can be conservative or liberal. The exact test with ordinary P-value cannot be liberal.

d. In part (a), a randomized-decision test generates a uniform random variable U from [0, 1] and rejects H_0 if both $T = 2$ and $U \leq \frac{5}{8}$. Show the actual $P(\text{type I error}) = 0.05$. Is this a sensible test?

1.13 The 2006 General Social Survey asked respondents how much government should spend on culture and the arts, with categories (much more, more, the same, less, much less). For 18–21 year-old females, the counts in these categories were (0, 8, 10, 9, 1). Find the Bayes estimates of the population proportions based on a Dirichlet prior distribution with $\{\alpha_i = K/5\}$ for values of $K = 1, 2.5, 5$. For each case, compare the estimate for the "much more" category to the ML estimate.

1.14 Refer to Example 1.6.4 on estimating the proportion of vegetarians. For the Jeffreys prior, find the posterior mean, the posterior 95% equal-tail interval, and the 95% highest posterior density interval.

1.15 You plan to use Bayesian methods to estimate binomial parameters in two cases, using n observations. In case (1) you want to estimate the probability that a new treatment for skin cancer is effective. In case (2) you want to estimate the probability of a head when you repeatedly flip a particular coin. Select prior distributions that you think would be sensible for each case. If they differ, explain why.

Theory and Methods

1.16 It is easier to get a precise estimate of the binomial parameter when π is near 0 or 1 than when it is near $\frac{1}{2}$. Explain why.

1.17 Suppose that $P(Y_i = 1) = 1 - P(Y_i = 0) = \pi, i = 1, \ldots, n$, where $\{Y_i\}$ are independent. Let $Y = \sum_i Y_i$.

a. What is the distribution of Y? What are $E(Y)$ and $\text{var}(Y)$?

b. When $\{Y_i\}$ instead have pairwise correlation $\rho > 0$, show that $\text{var}(Y) > n\pi(1 - \pi)$, overdispersion relative to the binomial. [Altham (1978) and Ochi and Prentice (1984) discussed generalizations of the binomial that allow correlated trials.]

c. Suppose that heterogeneity exists: $P(Y_i = 1|\pi) = \pi$ for all i, but π is a random variable with density function $g(\cdot)$ on [0, 1] having mean ρ and positive variance. Show that $\text{var}(Y) > n\,\rho(1 - \rho)$. (When π has a beta distribution, Y has the *beta-binomial distribution* of Section 14.3.)

1.18 For a sequence of independent Bernoulli trials, let Y be the number of successes before the kth failure. Explain why its probability mass function is the *negative*

binomial,

$$p(y) = \frac{(y+k-1)!}{y!(k-1)!}\pi^y(1-\pi)^k, \quad y = 0, 1, 2, \ldots .$$

[For it, $E(Y) = k\pi/(1-\pi)$ and $\text{var}(Y) = k\pi/(1-\pi)^2$, so $\text{var}(Y) > E(Y)$; the Poisson is the limit as $k \to \infty$ and $\pi \to 0$ with $k\pi = \mu$ fixed.]

1.19 For the multinomial distribution, show that

$$\text{corr}(n_j, n_k) = -\pi_j\pi_k/\sqrt{\pi_j(1-\pi_j)\pi_k(1-\pi_k)}.$$

When $c = 2$, show that this simplifies to $\text{corr}(n_1, n_2) = -1$, and explain why this makes intuitive sense.

1.20 Show that the moment generating function (mgf) is **(a)** $m(t) = (1-\pi+\pi e^t)^n$ for the binomial distribution, **(b)** $m(t) = \exp\{\mu[\exp(t)-1]\}$ for the Poisson distribution. For each distribution, use them to obtain the first two moments and to show a reproductive property.

1.21 A likelihood-ratio statistic equals t_o. At the ML estimates, show that the data are $\exp(t_o/2)$ times more likely under H_a than under H_0.

1.22 Suppose that $y_1, y_2, \ldots, y_n$ are independent from a Poisson distribution.

a. Obtain the likelihood function. Show that the ML estimator $\hat{\mu} = \bar{y}$.

b. Construct a large-sample test statistic for H_0: $\mu = \mu_0$ using (i) the Wald method, (ii) the score method, and (iii) the likelihood-ratio method.

c. Explain how to construct a large-sample confidence interval for μ using (i) the Wald method, (ii) the score method, and (iii) the likelihood-ratio method.

1.23 Inference for Poisson parameters can often be based on connections with binomial and multinomial distributions. Show how to test H_0: $\mu_1 = \mu_2$ for two populations based on independent Poisson counts (y_1, y_2), using a corresponding binomial test. [*Hint:* Condition on $n = y_1 + y_2$ and identify $\pi = \mu_1/(\mu_1+\mu_2)$.] How can you construct a confidence interval for μ_1/μ_2 based on one for π?

1.24 Since the Wald confidence interval for a binomial parameter π is degenerate when $\hat{\pi} = 0$ or 1, argue that the probability that the interval covers π cannot exceed $[1-\pi^n-(1-\pi)^n]$; hence, the infimum of the coverage probability over $0 < \pi < 1$ equals 0, regardless of n.

1.25 We noted in Section 1.4.2 that the midpoint $\tilde{\pi}$ of the score confidence interval (1.14) for π is the sample proportion after adding $z^2_{\alpha/2}$ observations to the sample, half of each type. This motivates a simple confidence interval,

$$\tilde{\pi} \pm z_{\alpha/2}\sqrt{\tilde{\pi}(1-\tilde{\pi})/n^*}, \quad \text{where } n^* = n + z^2_{\alpha/2}.$$

Show that the variance $\tilde{\pi}(1-\tilde{\pi})/n^*$ at the weighted average is at least as large as the weighted average of the variances that appears under the square root sign in the score interval. [*Hint:* Use Jensen's inequality.] Thus, this interval, which is sometimes referred to as the *Agresti–Coull confidence interval*, contains the score interval. [Agresti and Coull (1998) and Brown et al. (2001) showed that it performs much better than the Wald interval. It does not have the score interval's disadvantage (Exercise 16.32) of poor coverage near 0 and 1. With 95% confidence, this motivates a simple method that uses the Wald method after adding 2 observations of each type (Agresti and Coull 1998, Agresti and Caffo 2000); this is sometimes called the *plus four confidence interval*.]

1.26 A binomial sample of size n has $y = 0$ successes.

a. Show that the confidence interval for π based on the likelihood function is $[0.0, 1 - \exp(-z^2_{\alpha/2}/2n)]$. For $\alpha = 0.05$, use the expansion of an exponential function to show that this is approximately $[0, 1.92/n]$.

b. For the score method, show that the confidence interval is $[0, z^2_{\alpha/2}/(n + z^2_{\alpha/2})]$, or $[0, 3.84/(n + 3.84)]$ when $\alpha = 0.05$. (See Exercise 16.30 for small-sample intervals when $y = 0$.)

1.27 Suppose that $P(T = t_j) = \pi_j,\ j = 1, \ldots$. Show that $E(\text{mid } P\text{-value}) = 0.50$. [*Hint:* Show that $\sum_j \pi_j(\pi_j/2 + \pi_{j+1} + \cdots) = \left(\sum_j \pi_j\right)^2/2$.]

1.28 For a statistic T with cdf $F(t)$ and $p(t) = P(T = t)$, the *mid distribution function* is $F_{\text{mid}}(t) = F(t) - 0.50p(t)$ (Parzen 1997). Given $T = t_o$, show that the mid P-value equals $1 - F(t_o)$. (It also satisfies $E[F_{\text{mid}}(T)] = 0.50$ and $\text{var}[F_{\text{mid}}(T)] = (1/12)\{1 - E[p^2(T)]\}$.)

1.29 Genotypes AA, Aa, and aa occur with probabilities $[\theta^2, 2\theta(1-\theta), (1-\theta)^2]$. A multinomial sample of size n has frequencies (n_1, n_2, n_3) of these three genotypes.

a. Form the log likelihood. Show that $\hat{\theta} = (2n_1 + n_2)/(2n_1 + 2n_2 + 2n_3)$.

b. Show that $-\partial^2 L(\theta)/\partial\theta^2 = [(2n_1 + n_2)/\theta^2] + [(n_2 + 2n_3)/(1-\theta)^2]$ and that its expectation is $2n/\theta(1-\theta)$. Use this to obtain an asymptotic standard error of $\hat{\theta}$.

c. Explain how to test whether the probabilities truly have this pattern.

1.30 Refer to Section 1.5.6 and the model for pneumonia infections in calves. Using the likelihood function to obtain the information, show that the approximate standard error of $\hat{\pi}$ is $\sqrt{\pi(1-\pi)/n(1+\pi)}$.

1.31 Refer to Section 1.5.6. Let a denote the number of calves that got a primary, secondary, and tertiary infection, b the number that received a primary and secondary but not a tertiary infection, c the number that received a primary but not a secondary infection, and d the number that did not receive a primary infection. Let π be the probability of a primary infection. Consider the hypothesis that the probability of infection at time t, given infection at times $1, \ldots, t-1$, is also π, for $t = 2, 3$. Show that $\hat{\pi} = (3a + 2b + c)/(3a + 3b + 2c + d)$.

1.32 Refer to quadratic form (1.18) that leads to the Pearson chi-squared.

a. Verify that the matrix quoted in the text for $\mathbf{\Sigma}_0^{-1}$ is the inverse of $\mathbf{\Sigma}_0$.

b. Show that (1.18) simplifies to Pearson's statistic (1.16).

c. For the z_S statistic (1.11), show that $z_S^2 = X^2$ for $c = 2$.

1.33 For testing H_0: $\pi_j = \pi_{j0}$, $j = 1, \ldots, c$, using sample multinomial proportions $\{\hat{\pi}_j\}$, the likelihood-ratio statistic (1.17) is

$$G^2 = -2n \sum_j \hat{\pi}_j \log(\pi_{j0}/\hat{\pi}_j).$$

Show that $G^2 \geq 0$, with equality if and only if $\hat{\pi}_j = \pi_{j0}$ for all j. [*Hint:* Apply Jensen's inequality to $E(-2n \log X)$, where X equals $\pi_{j0}/\hat{\pi}_j$ with probability $\hat{\pi}_j$.]

1.34 For counts $\{n_i\}$, the *power divergence statistic* for testing goodness of fit (Cressie and Read 1984, Read and Cressie 1988) is

$$\frac{2}{\lambda(\lambda+1)} \sum n_i[(n_i/\hat{\mu}_i)^\lambda - 1] \quad \text{for } -\infty < \lambda < \infty.$$

a. For $\lambda = 1$, show that this equals X^2.

b. As $\lambda \to 0$, show that it converges to G^2. [*Hint:* $\log t = \lim_{h\to 0}(t^h - 1)/h$.]

c. As $\lambda \to -1$, show that it converges to $2\sum \hat{\mu}_i \log(\hat{\mu}_i/n_i)$, the *minimum discrimination information* statistic (Gokhale and Kullback 1978).

d. For $\lambda = -2$, show that it equals $\sum(n_i - \hat{\mu}_i)^2/n_i$, the *Neyman modified chi-squared* statistic (Neyman 1949).

e. For $\lambda = -\frac{1}{2}$, show that it equals $4\sum(\sqrt{n_i} - \sqrt{\hat{\mu}_i})^2$, the *Freeman–Tukey* statistic (Freeman and Tukey 1950).

[Under regularity conditions, their asymptotic distributions are identical (Drost et al. 1989). The chi-squared null approximation works best for λ near $\frac{2}{3}$.]

1.35 The chi-squared mgf with df $= \nu$ is $m(t) = (1 - 2t)^{-\nu/2}$, for $|t| < \frac{1}{2}$. Use it to prove the reproductive property of the chi-squared distribution.

1.36 For the multinomial $(n, \{\pi_j\})$ distribution with $c > 2$, a possible set of score-type simultaneous confidence limits for π_j are the solutions of

$$(\hat{\pi}_j - \pi_j)^2/[\pi_j(1 - \pi_j)/n] = (z_{\alpha/2c})^2, \quad j = 1, \ldots, c.$$

a. Using the Bonferroni inequality, argue that for large n these c intervals simultaneously contain all $\{\pi_j\}$ with probability at least $1 - \alpha$.

b. Show that the standard deviation of $\hat{\pi}_j - \hat{\pi}_k$ is $[\pi_j + \pi_k - (\pi_j - \pi_k)^2]/n$. Let $a = c(c-1)/2$. For large n, explain why the probability is at least $1 - \alpha$ that the

Wald confidence intervals

$$(\hat{\pi}_j - \hat{\pi}_k) \pm z_{\alpha/2a}\{[\hat{\pi}_j + \hat{\pi}_k - (\hat{\pi}_j - \hat{\pi}_k)^2]/n\}^{1/2}$$

simultaneously contain the a differences $\{\pi_j - \pi_k\}$ (Goodman 1965).

1.37 Consider the Bayesian equal-tail posterior interval for a binomial parameter π, using a beta or logit-normal prior. When $y = 0$, explain why the lower limit for π can never be 0, unlike the frequentist approach based on inverting a score or likelihood-ratio test.

1.38 Consider estimating the ratio π_i/π_j of two multinomial parameters. Should the estimate depend at all on the counts in other categories?

a. With a frequentist approach, explain why the ML estimate of π_i/π_j is n_i/n_j.

b. For a Dirichlet prior, show that using the Bayes estimates of π_i and π_j to estimate π_i/π_j uses also the counts in other categories. (However, the posterior distribution of $\gamma = \pi_i/(\pi_i + \pi_j)$ is the same as its posterior distribution ignoring the other counts and treating y_i as binomial with sample size $(y_i + y_j)$ and parameter γ.)

1.39 Given π, Y has a bin(n, π) distribution, and π has a uniform prior distribution. Show that the marginal distribution of Y is uniform over $0, 1, \ldots, n$.

1.40 Consider the Bayes estimator of the binomial parameter π using a beta prior distribution.

a. Show that the ML estimator is a limit of Bayes estimators, for a certain sequence of beta prior parameter values.

b. Find an improper prior density such that the Bayes estimator coincides with the ML estimator. (In this sense, the ML estimator is a *generalized Bayes estimator*.)

1.41 For the Dirichlet prior for multinomial probabilities, show the posterior expected value of π_i is formula (1.19). Derive the expression for this Bayes estimator as a weighted average of p_i and $E(\pi_i)$.