CHAPTER 1

INTRODUCTION

From helping to assess the value of new medical treatments to evaluating the factors that affect our opinions on controversial issues, scientists today are finding myriad uses for categorical data analyses. It is primarily for these scientists and their collaborating statisticians – as well as those training to perform these roles – that this book was written.

This first chapter reviews the most important probability distributions for categorical data: the *binomial* and *multinomial* distributions. It also introduces *maximum likelihood*, the most popular method for using data to estimate parameters. We use this type of estimate and a related *likelihood function* to conduct statistical inference. We also introduce the *Bayesian* approach to statistical inference, which utilizes probability distributions for the parameters as well as for the data. We begin by describing the major types of categorical data.

1.1 CATEGORICAL RESPONSE DATA

A *categorical* variable has a measurement scale consisting of a set of categories. For example, political ideology might be measured as liberal, moderate, or conservative; choice of accommodation might use categories house, condominium, and apartment; a diagnostic test to detect e-mail spam might classify an incoming e-mail message as spam or legitimate. Categorical variables are often referred to as *qualitative*, to distinguish them from *quantitative* variables, which take numerical values, such as age, income, and number of children in a family.

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e

Categorical variables are pervasive in the social sciences for measuring attitudes and opinions, with categories such as (agree, disagree), (yes, no), and (favor, oppose, undecided). They also occur frequently in the health sciences, for measuring responses such as whether a medical treatment is successful (yes, no), mammogram-based breast diagnosis (normal, benign, probably benign, suspicious, malignant with cancer), and stage of a disease (initial, intermediate, advanced). Categorical variables are common for service-quality ratings of any company or organization that has customers (e.g., with categories excellent, good, fair, poor). In fact, categorical variables occur frequently in most disciplines. Other examples include the behavioral sciences (e.g., diagnosis of type of mental illness, with categories schizophrenia, depression, neurosis), ecology (e.g., primary land use in satellite image, with categories woodland, swamp, grassland, agriculture, urban), education (e.g., student responses to an exam question, with categories correct, incorrect), and marketing (e.g., consumer cell-phone preference, with categories Samsung, Apple, Nokia, LG, Other). They even occur in highly quantitative fields such as the engineering sciences and industrial quality control, when items are classified according to whether or not they conform to certain standards.

1.1.1 Response Variable and Explanatory Variables

Most statistical analyses distinguish between a *response* variable and *explanatory* variables. For instance, ordinary regression models describe how the mean of a quantitative response variable, such as annual income, changes according to levels of explanatory variables, such as number of years of education and number of years of job experience. The response variable is sometimes called the *dependent variable* and the explanatory variable is sometimes called the *independent variable*. When we want to emphasize that the response variable is a random variable, such as in a probability statement, we use upper-case notation for it (e.g., Y). We use lower-case notation to refer to a particular value (e.g., $y = 0$).

This text presents statistical models that relate a categorical response variable to explanatory variables that can be categorical or quantitative. For example, a study might analyze how opinion about whether same-sex marriage should be legal (yes or no) is associated with explanatory variables such as number of years of education, annual income, political party affiliation, religious affiliation, age, gender, and race.

1.1.2 Binary–Nominal–Ordinal Scale Distinction

Many categorical variables have only two categories, such as (yes, no) for possessing health insurance or (favor, oppose) for legalization of marijuana. Such variables are called *binary variables*.

When a categorical variable has more then two categories, we distinguish between two types of categorical scales. Categorical variables having *unordered* scales are called *nominal* variables. Examples are religious affiliation (categories Christian, Jewish, Muslim, Buddhist, Hindu, none, other), primary mode of transportation to work (automobile, bicycle, bus, subway, walk), and favorite type of music (classical, country, folk, jazz, pop, rock). Variables having naturally *ordered* categories are called *ordinal* variables. Examples are perceived happiness (not too happy, pretty happy, very happy), frequency of feeling anxiety (never, occasionally, often, always), and headache pain (none, slight, moderate, severe).

A variable's measurement scale determines which statistical methods are appropriate. For nominal variables, the order of listing the categories is arbitrary, so methods designed for them give the same results no matter what order is used. Methods designed for ordinal variables utilize the category ordering.

1.1.3 Organization of this Book

Chapters 1 and 2 describe basic non model-based methods of categorical data analysis. These include analyses of proportions and of association between categorical variables.

Chapters 3 to 7 introduce models for categorical response variables. These models resemble regression models for quantitative response variables. In fact, Chapter 3 shows they are special cases of a class of *generalized linear models* that also contains the ordinary normal-distribution-based regression models. *Logistic regression* models, which apply to binary response variables, are the focus of Chapters 4 and 5. Chapter 6 extends logistic regression to multicategory responses, both nominal and ordinal. Chapter 7 introduces *loglinear* models, which analyze associations among multiple categorical response variables.

The methods in Chapters 1 to 7 assume that observations are independent. Chapters 8 to 10 introduce logistic regression models for observations that are correlated, such as for matched pairs or for repeated measurement of individuals in longitudinal studies. Chapter 11 introduces some advanced methods, including ways of classifying and clustering observations into categories and ways of dealing with data sets having huge numbers of variables. The book concludes (Chapter 12) with a historical overview of the development of categorical data methods.

Statistical software packages can implement methods for categorical data analysis. We illustrate throughout the text for the free software `R`. The Appendix discusses the use of SAS, Stata, and SPSS. A companion website for the book, `www.stat.ufl.edu/~aa/cat`, has additional information, including complete data sets for the examples. The data files are also available at `https://github.com/alanagresti/categorical-data`.

1.2 PROBABILITY DISTRIBUTIONS FOR CATEGORICAL DATA

Parametric inferential statistical analyses require an assumption about the probability distribution of the response variable. For regression models for quantitative variables, the normal distribution plays a central role. This section presents the key probability distributions for categorical variables: the *binomial* and *multinomial* distributions.

1.2.1 Binomial Distribution

When the response variable is binary, we refer to the two outcome categories as *success* and *failure*. These labels are generic and the *success* outcome need not be a preferred result.

Many applications refer to a fixed number n of independent and identical trials with two possible outcomes for each. *Identical trials* means that the probability of success is the same for each trial. *Independent trials* means the response outcomes are independent random variables. In particular, the outcome of one trial does not affect the outcome of another. These are often called *Bernoulli trials*. Let π denote the probability of success for

each trial. Let Y denote the number of successes out of the n trials. Under the assumption of n independent, identical trials, Y has the *binomial distribution* with index n and parameter π. The probability of a particular outcome y for Y equals

$$P(y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y}, \quad y = 0, 1, 2, \ldots, n. \tag{1.1}$$

To illustrate, suppose a quiz has ten multiple-choice questions, with five possible answers for each. A student who is completely unprepared randomly guesses the answer for each question. Let Y denote the number of correct responses. For each question, the probability of a correct response is 0.20, so $\pi = 0.20$ with $n = 10$. The probability of $y = 0$ correct responses, and hence $n - y = 10$ incorrect ones, equals

$$P(0) = \frac{10!}{0!10!}(0.20)^0(0.80)^{10} = (0.80)^{10} = 0.107.$$

The probability of 1 correct response equals

$$P(1) = \frac{10!}{1!9!}(0.20)^1(0.80)^9 = 10(0.20)(0.80)^9 = 0.268.$$

Table 1.1 shows the binomial distribution for all the possible values, $y = 0, 1, 2, \ldots, 10$. For contrast, it also shows the binomial distributions when $\pi = 0.50$ and when $\pi = 0.80$.

Table 1.1 Binomial distributions with $n = 10$ and $\pi = 0.20$, 0.50, and 0.80. The binomial distribution is symmetric when $\pi = 0.50$.

y	$P(y)$ when $\pi = 0.20$ ($\mu = 2.0$, $\sigma = 1.26$)	$P(y)$ when $\pi = 0.50$ ($\mu = 5.0$, $\sigma = 1.58$)	$P(y)$ when $\pi = 0.80$ ($\mu = 8.0$, $\sigma = 1.26$)
0	0.107	0.001	0.000
1	0.268	0.010	0.000
2	0.302	0.044	0.000
3	0.201	0.117	0.001
4	0.088	0.205	0.005
5	0.027	0.246	0.027
6	0.005	0.205	0.088
7	0.001	0.117	0.201
8	0.000	0.044	0.302
9	0.000	0.010	0.268
10	0.000	0.001	0.107

The binomial distribution for n trials with parameter π has mean and standard deviation

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

The binomial distribution with $\pi = 0.20$ in Table 1.1 has $\mu = 10(0.20) = 2.0$. The standard deviation is $\sigma = \sqrt{10(0.20)(0.80)} = 1.26$, which σ also equals when $\pi = 0.80$.

The binomial distribution is symmetric when $\pi = 0.50$. For fixed n, it becomes more bell-shaped as π gets closer to 0.50. For fixed π, it becomes more bell-shaped as n increases.

When n is large, it can be approximated by a normal distribution with $\mu = n\pi$ and $\sigma = \sqrt{n\pi(1-\pi)}$. A guideline[1] is that the expected number of outcomes of the two types, $n\pi$ and $n(1-\pi)$, should both be at least about 5. For $\pi = 0.50$ this requires only $n \geq 10$, whereas $\pi = 0.10$ (or $\pi = 0.90$) requires $n \geq 50$. When π gets nearer to 0 or 1, larger samples are needed before a symmetric, bell shape occurs.

1.2.2 Multinomial Distribution

Nominal and ordinal response variables have more than two possible outcomes. When the observations are independent with the same category probabilities for each, the probability distribution of counts in the outcome categories is the *multinomial*.

Let c denote the number of outcome categories. We denote their probabilities by $(\pi_1, \pi_2, \ldots, \pi_c)$, where $\sum_j \pi_j = 1$. For n independent observations, the multinomial probability that y_1 fall in category 1, y_2 fall in category 2, ... , y_c fall in category c, where $\sum_j y_j = n$, equals

$$P(y_1, y_2, ..., y_c) = \left(\frac{n!}{y_1!y_2!\cdots y_c!}\right)\pi_1^{y_1}\pi_2^{y_2}\cdots\pi_c^{y_c}.$$

The binomial distribution is the special case with $c = 2$ categories. We will not need to use this formula, because our focus is on inference methods that use *sampling distributions* of statistics computed from the multinomial counts, and those sampling distributions are approximately *normal* or *chi-squared*.

1.3 STATISTICAL INFERENCE FOR A PROPORTION

In practice, the parameter values for binomial and multinomial distributions are unknown. Using sample data, we estimate the parameters. This section introduces the *maximum likelihood* estimation method and illustrates it for the binomial parameter.

1.3.1 Likelihood Function and Maximum Likelihood Estimation

The parametric approach to statistical modeling assumes a family of probability distributions for the response variable, indexed by an unknown parameter. For a particular family, we can substitute the observed data into the formula for the probability function and then view how that probability depends on the unknown parameter value. For example, in $n = 10$ trials, suppose a binomial count equals $y = 0$. From the binomial formula (1.1) with parameter π, the probability of this outcome equals

$$P(0) = \frac{10!}{0!10!}\pi^0(1-\pi)^{10} = (1-\pi)^{10}.$$

This probability is defined for all the potential values of π between 0 and 1.

[1] You can explore this with the binomial distribution applet at `www.artofstat.com/webapps.html`.

The probability of the observed data, expressed as a function of the parameter, is called the *likelihood function*. With $y = 0$ successes in $n = 10$ trials, the binomial likelihood function is $\ell(\pi) = (1 - \pi)^{10}$, for $0 \leq \pi \leq 1$. If $\pi = 0.40$, for example, the probability that $y = 0$ is $\ell(0.40) = (1 - 0.40)^{10} = 0.006$. Likewise, if $\pi = 0.20$ then $\ell(0.20) = (1 - 0.20)^{10} = 0.107$, and if $\pi = 0.0$ then $\ell(0.0) = (1 - 0.0)^{10} = 1.0$. Figure 1.1 plots this likelihood function for all π values between 0 and 1.

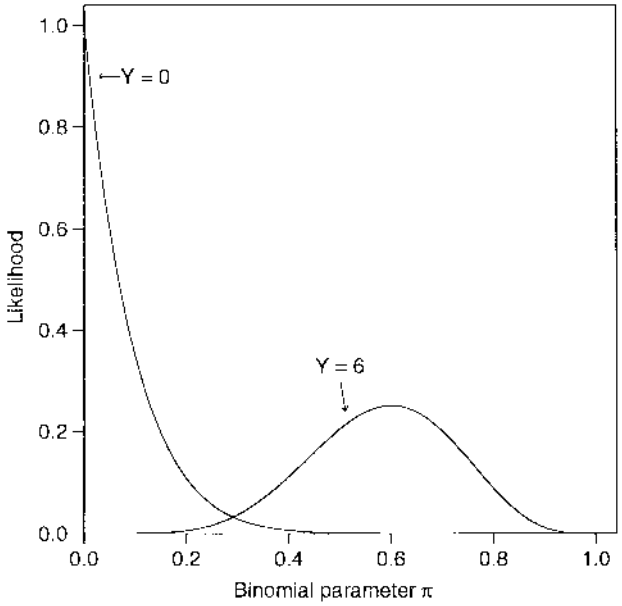

Figure 1.1 Binomial likelihood functions for $y = 0$ successes and for $y = 6$ successes in $n = 10$ trials.

The *maximum likelihood estimate* of a parameter is the parameter value at which the likelihood function takes its maximum. That is, it is the parameter value for which the probability of the observed data takes its greatest value. Figure 1.1 shows that the likelihood function $\ell(\pi) = (1 - \pi)^{10}$ has its maximum at $\pi = 0.0$. Therefore, when $n = 10$ trials have $y = 0$ successes, the maximum likelihood estimate of π equals 0.0. This means that the result $y = 0$ in $n = 10$ trials is more likely to occur when $\pi = 0.00$ than when π equals any other value.

We use the abbreviation *ML* to symbolize *maximum likelihood*. The ML estimate is often denoted by the parameter symbol with a ^ (a *hat*) over it. We denote the ML estimate of the binomial parameter π by $\hat{\pi}$, called *pi-hat*. In general, for the binomial outcome of y successes in n trials, the maximum likelihood estimate of π is $\hat{\pi} = y/n$. This is the sample proportion of successes for the n trials. If we observe $y = 6$ successes in $n = 10$ trials, then the maximum likelihood estimate of π is $\hat{\pi} = 6/10 = 0.60$. Figure 1.1 also plots the likelihood function when $n = 10$ with $y = 6$, which from formula (1.1) equals $\ell(\pi) = [10!/(6!4!)]\pi^6(1 - \pi)^4$. The maximum value occurs at $\hat{\pi} = 0.60$. The result $y = 6$ in $n = 10$ trials is more likely to occur when $\pi = 0.60$ than when π equals any other value.

If we denote each success by a 1 and each failure by a 0, then the sample proportion equals the sample mean of the data. For instance, for 4 failures followed by 6 successes in 10 trials, the data are $(0, 0, 0, 0, 1, 1, 1, 1, 1, 1)$ and the sample mean is

$$\hat{\pi} = (0 + 0 + 0 + 0 + 1 + 1 + 1 + 1 + 1 + 1)/10 = 0.60.$$

Thus, results that apply to sample means with random sampling apply also to sample proportions. These include the *Central Limit Theorem*, which states that the sampling distribution of the sample proportion $\hat{\pi}$ is approximately normal for large n, and the *Law of Large Numbers*, which states that $\hat{\pi}$ converges to the population proportion π as n increases.

Before we observe the data, the value of the ML estimate is unknown. The estimate is then a random variable having some sampling distribution. We refer to it as an *estimator* and its value for observed data as an *estimate*. Estimators based on the method of maximum likelihood are popular because they have good large-sample behavior. Sampling distributions of ML estimators are typically approximately normal and no other "good" estimator has a smaller standard error.

1.3.2 Significance Test About a Binomial Parameter

For the binomial distribution, we now use the ML estimator in statistical inference for the parameter π. The ML estimator $\hat{\pi}$ is the sample proportion. Its sampling distribution has mean and standard error

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

Consider the null hypothesis H_0: $\pi = \pi_0$ that the parameter equals some fixed value, π_0, such as 0.50. When H_0 is true, the standard error of $\hat{\pi}$ is $SE_0 = \sqrt{\pi_0(1-\pi_0)/n}$, which we refer to as the *null standard error*. The test statistic

$$z = \frac{\hat{\pi} - \pi_0}{SE_0} = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \tag{1.2}$$

divides the difference between the sample proportion $\hat{\pi}$ and the null hypothesis value π_0 by the null standard error. The z test statistic measures the number of standard errors that $\hat{\pi}$ falls from the H_0 value. For large samples, the null sampling distribution of z is the standard normal, which has mean $= 0$ and standard deviation $= 1$.

1.3.3 Example: Surveyed Opinions About Legalized Abortion

Do a majority, or minority, of adults in the United States believe that a pregnant woman should be able to obtain an abortion? Let π denote the proportion of the American adult population that responds *yes* when asked, "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if she wants it for any reason." We test H_0: $\pi = 0.50$ against the two-sided alternative hypothesis, H_a: $\pi \neq 0.50$.

This item was one of many about legalized abortion included in the 2016 General Social Survey (GSS). This survey, conducted every other year by the National Opinion Research Center (NORC) at the University of Chicago, asks a sample of adult Americans their opinions about a wide variety of issues.[2] The GSS is a multi-stage sample, but it has characteristics similar to a simple random sample. Of 1810 respondents to this item in 2016, 837 replied *yes* and 973 replied *no*. The sample proportion of *yes* responses was $\hat{\pi} = 837/1810 = 0.4624$.

[2] You can view responses to surveys since 1972 at `sda.berkeley.edu/archive.htm`.

The test statistic for H_0: $\pi = 0.50$ is

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.4624 - 0.50}{\sqrt{\frac{0.50(0.50)}{1810}}} = -3.20.$$

The two-sided P-value is the probability that the absolute value of a standard normal variate exceeds 3.20, which is $P = 0.0014$. The evidence is very strong that, in 2016, $\pi < 0.50$, that is, that fewer than half of Americans favored unrestricted legal abortion. In some other situations, such as when the mother's health was endangered, an overwhelming majority favored legalized abortion. Responses depended strongly on the question wording.

1.3.4 Confidence Intervals for a Binomial Parameter

A significance test merely indicates whether a particular value for a parameter (such as 0.50) is plausible. We learn more by constructing a confidence interval to determine the range of plausible values. Let $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$ denote the estimated standard error of $\hat{\pi}$. This formula obtains SE by substituting the ML estimate $\hat{\pi}$ for the unknown parameter π in $\sigma(\hat{\pi}) = \sqrt{\pi(1-\pi)/n}$. One way to form a $100(1-\alpha)$% confidence interval for π uses the formula

$$\hat{\pi} \pm z_{\alpha/2}(SE), \text{ with } SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n}, \tag{1.3}$$

where $z_{\alpha/2}$ denotes the standard normal percentile having right-tail probability equal to $\alpha/2$; for example, for 95% confidence, $\alpha = 0.05$, $z_{\alpha/2} = z_{0.025} = 1.96$.

For the opinion about legalized abortion example just discussed, $\hat{\pi} = 0.462$ for $n = 1810$ observations. The 95% confidence interval equals

$$0.462 \pm 1.96\sqrt{0.462(0.538)/1810}, \text{ which is } 0.462 \pm 0.023, \text{ or } (0.439, 0.485).$$

We can be 95% confident that the population proportion of Americans in 2016 who favored unrestricted legalized abortion is between 0.439 and 0.485.

The significance test and confidence interval for π, as well as other confidence intervals presented next, are readily available in software and at web sites.[3]

1.3.5 Better Confidence Intervals for a Binomial Proportion *

Formula (1.3) is simple. When π is near 0 or near 1, however, it performs poorly unless n is very large. Its *actual* coverage probability, that is, the probability that the method produces an interval that captures the true parameter value, may be much less than the nominal value (such as 0.95).

A better way to construct confidence intervals uses a duality with significance tests. The confidence interval consists of all H_0 values π_0 that are judged plausible in the z test of Section 1.3.2. A 95% confidence interval contains all values π_0 for which the two-sided P-value exceeds 0.05. That is, it contains all values that are *not rejected* at the 0.05

[3] For instance, see `https://istats.shinyapps.io/Inference_prop`. The confidence interval (1.3) is the *Wald* type listed in the menu.

significance level. These are the H_0 values for π_0 that have test statistic z less than 1.96 in absolute value. This alternative method, called the *score confidence interval*, has the advantage that we do not need to estimate π in the standard error, because the standard error $SE_0 = \sqrt{\pi_0(1-\pi_0)/n}$ in the test statistic uses the null value π_0.

To illustrate, suppose that a clinical trial to evaluate a new treatment has 9 successes in the first 10 trials. For a sample proportion of $\hat{\pi} = 0.90$ based on $n = 10$, the value $\pi_0 = 0.596$ for the H_0 parameter value yields the test statistic value

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.90 - 0.596}{\sqrt{\frac{0.596(0.404)}{10}}} = 1.96$$

and a two-sided P-value of $P = 0.05$. The value $\pi_0 = 0.982$ yields

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.90 - 0.982}{\sqrt{\frac{0.982(0.018)}{10}}} = -1.96$$

and also a two-sided P-value of $P = 0.05$. All π_0 values between 0.596 and 0.982 have $|z| < 1.96$ and P-value > 0.05. Therefore, the 95% score confidence interval for π is (0.596, 0.982). For particular values of $\hat{\pi}$ and n, the π_0 values that have test statistic value $z = \pm 1.96$ are the solutions to the equation

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

for π_0. We will not deal here with how to solve this equation, as this confidence interval is readily available in software and at web sites.[4]

The simple formula (1.3) using estimated standard error fails spectacularly when $\hat{\pi} = 0$ or when $\hat{\pi} = 1$, regardless of how large n is. To illustrate, suppose the clinical trial had 10 successes in the 10 trials. Then, $\hat{\pi} = 10/10 = 1.0$ and $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n} = \sqrt{1.0(0.0)/10} = 0$, so the 95% confidence interval $1.0 \pm 1.96(SE)$ is 1.0 ± 0.0. This interval (1.0, 1.0) is completely unrealistic. When a sample estimate is at or near the boundary of the parameter space, having that estimate in the middle of the confidence interval results in poor performance of the method. By contrast, the 95% score confidence interval based on the corresponding significance test with null standard error SE_0 is (0.72, 1.0).

The score confidence interval itself has actual coverage probability a bit too small when π is very close to 0 or 1. A simple alternative confidence interval approximates the score interval but is a bit wider and has better coverage probability when π is near 0 or 1. It uses the simple formula (1.3) with the estimated standard error after adding 2 to the number of successes and 2 to the number of failures (and thus 4 to n). With 10 successes in 10 trials, you apply formula (1.3) to 12 successes in 14 trials and get (0.68, 1.0). This simple method,[5] called the *Agresti–Coull confidence interval*, has adequate coverage probability for small n even when π is very close to 0 or 1.

[4] Such as the "Wilson score" option at `https://istats.shinyapps.io/Inference_prop`.

[5] More precisely, software and the website `https://istats.shinyapps.io/Inference_prop` adds $(z^2_{\alpha/2})/2$ to each count (e.g., $(1.96)^2/2 = 1.92$ for 95% confidence); the CI then performs well because it has the same midpoint as the score CI but is a bit wider.

1.4 STATISTICAL INFERENCE FOR DISCRETE DATA

In summary, two methods we have presented for constructing a confidence interval for a proportion (1) use $\hat{\pi} \pm z_{\alpha/2}(SE)$ with the estimated standard error, $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$, or (2) invert results of a significance test using test statistic $z = (\hat{\pi} - \pi_0)/SE_0$ with the null standard error, $SE_0 = \sqrt{\pi_0(1-\pi_0)/n}$. These methods apply two of the three standard ways of conducting statistical inference (confidence intervals and significance tests) about parameters. We present the methods in a more general context in this section and also introduce a third standard inference method that uses the likelihood function.

1.4.1 Wald, Likelihood-Ratio, and Score Tests

Let β denote an arbitrary parameter, such as a linear effect of an explanatory variable in a model. Consider a significance test of H_0: $\beta = \beta_0$, such as H_0: $\beta = 0$ for which $\beta_0 = 0$. The simplest test statistic exploits the large-sample normality of the ML estimator $\hat{\beta}$. Let SE denote the unrestricted standard error of $\hat{\beta}$, evaluated by substituting the ML estimate for the unknown parameter in the expression for the true standard error. (For example, for the binomial parameter π, $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$.) When H_0 is true, the test statistic

$$z = (\hat{\beta} - \beta_0)/SE$$

has approximately a standard normal distribution. Equivalently, z^2 has approximately a chi-squared distribution with $df = 1$. This type of statistic, which uses the standard error evaluated at the ML estimate, is called a *Wald statistic*. The z test using this test statistic, or the corresponding chi-squared test that uses z^2, is called a *Wald test*.[6]

We can refer z to the standard normal distribution to get one-sided or two-sided P-values. For the two-sided alternative H_a: $\beta \neq \beta_0$, the P-value is also the right-tail chi-squared probability with $df = 1$ above the observed value of z^2. That is, the two-tail probability beyond $\pm z$ for the standard normal distribution equals the right-tail probability above z^2 for the chi-squared distribution with $df = 1$. For example, the two-tail standard normal probability of 0.05 that falls below -1.96 and above 1.96 equals the right-tail chi-squared probability above $(1.96)^2 = 3.84$ when $df = 1$. With the software `R` and its functions `pnorm` and `pchisq` for *cumulative probabilities* (i.e., probabilities *below* fixed values) for normal and chi-squared distributions, we find (with comments added following the # symbol):

```
------------------------------------------------------------------------------
> 2*pnorm(-1.96) # 2(standard normal cumulative probability below -1.96)
[1] 0.0499958    # essentially equals 0.05
> pchisq(1.96^2, 1) # pchisq gives chi-squared cumulative probability
[1] 0.9500042       # here, cumul. prob. at (1.96)(1.96) = 3.84 when df=1
> 1 - pchisq(1.96^2, 1) # right-tail prob. above (1.96)(1.96) when df=1
[1] 0.0499958    # same as normal two-tail probability
> # can also get this by  pchisq(1.96^2, 1, lower.tail=FALSE)
------------------------------------------------------------------------------
```

You can also find chi-squared and normal tail probabilities with applets on the Internet.[7]

[6] Proposed by the statistician Abraham Wald in 1943.

[7] See, for example, the applets at `www.artofstat.com/webapps.html#Distributions`.

A second possible test is called the *score test*.[8] This test uses standard errors that are valid when H_0 is true, rather than estimated more generally. For example, the z test (1.2) for a binomial parameter that uses the null standard error $SE_0 = \sqrt{\pi_0(1-\pi_0)/n}$ of $\hat{\pi}$ is a score test. The z test that uses $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$ instead of SE_0 is the Wald test.

A third possible test of H_0: $\beta = \beta_0$ uses the likelihood function through the ratio of two of its values. For a single parameter β, these are (1) the value ℓ_0 when H_0 is true (so $\beta = \beta_0$), (2) the maximum ℓ_1 over all possible parameter values, which is the likelihood function calculated at the ML estimate $\hat{\beta}$. Then ℓ_1 is always at least as large as ℓ_0, because ℓ_1 refers to maximizing over the entire parameter space rather than just at β_0. The *likelihood-ratio* test statistic[9] equals

$$2\log(\ell_1/\ell_0).$$

The reason for taking the log transform and doubling is that it yields an approximate chi-squared sampling distribution. Under H_0: $\beta = \beta_0$, this test statistic has a large-sample chi-squared distribution with $df = 1$. The test statistic $2\log(\ell_1/\ell_0)$ is nonnegative, and the P-value is the chi-squared right-tail probability. Larger values of (ℓ_1/ℓ_0) yield larger values of $2\log(\ell_1/\ell_0)$ and smaller P-values and stronger evidence against H_0.

For ordinary regression models that assume a normal distribution for Y, the Wald, score, and likelihood-ratio tests provide identical test statistics and P-values. For parameters in other statistical models, they have similar behavior when the sample size n is large and H_0 is true. When n is small to moderate, the Wald test is the least reliable of the three tests. The likelihood-ratio inference and score-test based inference are better in terms of actual inferential error probabilities, coming closer to matching nominal levels.

For any of the three tests, the P-value that software reports is an approximation for the true P-value. This is because the normal (or chi-squared) sampling distribution used is a large-sample approximation for the actual sampling distribution. Thus, when you report a P-value, it is overly optimistic to use many decimal places. If you are lucky, the P-value approximation is good to the second decimal place. Therefore, for a P-value that software reports as 0.028374, it makes more sense to report it as 0.03 (or, at best, 0.028) rather than 0.028374. An exception is when the P-value is zero to many decimal places, in which case it is sensible to report it as $P < 0.001$ or $P < 0.0001$. A P-value merely summarizes the strength of evidence against H_0, and accuracy to two or three decimal places is sufficient for this purpose.

Each significance test method has a corresponding confidence interval. The 95% confidence interval for β is the set of β_0 values for the test of H_0: $\beta = \beta_0$ such that the P-value is larger than 0.05. For example, the 95% *Wald confidence interval* is the set of β_0 values for which $z = (\hat{\beta} - \beta_0)/SE$ has $|z| < 1.96$. It is $\hat{\beta} \pm 1.96(SE)$.

1.4.2 Example: Wald, Score, and Likelihood-Ratio Binomial Tests

We illustrate the Wald, score, and likelihood-ratio tests by testing H_0: $\pi = 0.50$ against H_a: $\pi \neq 0.50$ for the toy example mentioned on page 16 of a clinical trial to evaluate a new treatment that has 9 successes in $n = 10$ trials. The sample proportion is $\hat{\pi} = 0.90$.

[8] Proposed by the statistician Calyampudi Radhakrishna Rao in 1948.

[9] Proposed by the statistician Sam Wilks in 1938; in this text, we use the *natural log*, which has $e = 2.718...$ as the base. It is often denoted on calculators by LN.

For the Wald test, the estimated standard error is $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n} = \sqrt{0.90(0.10)/10} = 0.095$. The z test statistic is

$$z = (\hat{\pi} - \pi_0)/SE = (0.90 - 0.50)/0.095 = 4.22.$$

The corresponding chi-squared statistic is $(4.22)^2 = 17.78$ ($df = 1$). The P-value < 0.001.

For the score test, the null standard error is $SE_0 = \sqrt{\pi_0(1-\pi_0)/n} = \sqrt{0.50(0.50)/10} = 0.158$. The z test statistic is

$$z = (\hat{\pi} - \pi_0)/SE_0 = (0.90 - 0.50)/0.158 = 2.53.$$

The corresponding chi-squared statistic is $(2.53)^2 = 6.40$ ($df = 1$). The P-value $= 0.011$.

The likelihood function is the binomial probability of the observed result of 9 successes in 10 trials, viewed as a function of the parameter,

$$\ell(\pi) = \frac{10!}{9!1!}\pi^9(1-\pi)^1 = 10\pi^9(1-\pi).$$

The likelihood-ratio test compares this when H_0: $\pi = 0.50$ is true, for which $\ell_0 = 10(0.50)^9(0.50) = 0.00977$, to the value at the ML estimate of $\hat{\pi} = 0.90$, for which $\ell_1 = 10(0.90)^9(0.10) = 0.3874$. The likelihood-ratio test statistic equals

$$2\log(\ell_1/\ell_0) = 2[\log(0.3874/0.00977)] = 7.36.$$

From the chi-squared distribution with $df = 1$, this statistic has P-value $= 0.007$.

A marked divergence in the values of the three statistics, such as often happens when n is small and the ML estimate is near the boundary of the parameter space, indicates that the sampling distribution of the ML estimator may be far from normality and an estimate of the standard error may be poor. In that case, special small-sample methods are more reliable.

1.4.3 Small-Sample Binomial Inference and the Mid P-Value *

For statistical inference about a binomial parameter, the large-sample likelihood-ratio and two-sided score tests and the confidence intervals based on those tests perform reasonably well when $n\pi \geq 5$ and $n(1-\pi) \geq 5$. Otherwise, it is better to use the binomial distribution directly. With modern software, we can use this direct approach with any n.

To illustrate using the binomial directly, consider testing H_0: $\pi = 0.50$ against H_a: $\pi > 0.50$ for the toy example of a clinical trial, with $y = 9$ successes in $n = 10$ trials. The exact P-value, based on the right tail of the null binomial distribution with $\pi = 0.50$, is the binomial probability

$$P(Y \geq 9) = P(9) + P(10) = \frac{10!}{9!1!}(0.50)^9(0.50)^1 + \frac{10!}{10!0!}(0.50)^{10}(0.50)^0 = 0.011.$$

For the two-sided alternative H_a: $\pi \neq 0.50$, the P-value is

$$P(Y \geq 9 \text{ or } Y \leq 1) = 2[P(Y \geq 9)] = 0.021.$$

With discrete probability distributions, small-sample inference using the ordinary P-value is *conservative*. This means that when H_0 is true, the P-value is ≤ 0.05 (thus leading to rejection of H_0 at the 0.05 significance level) not *exactly* 5% of the time, but *no more* than 5% of the time. Then, the actual P(Type I error) is not exactly 0.05, but may be much less than 0.05. For example, for testing H_0: $\pi = 0.50$ against H_a: $\pi > 0.50$ with $y = 9$ successes in $n = 10$ trials, from the binomial probabilities with $\pi = 0.50$ in Table 1.1 in Section 1.2.1, the right-tail P-value is ≤ 0.05 only when $y = 9$ or 10. This happens with probability $0.010 + 0.001 = 0.011$. Thus, the probability of rejecting H_0 (i.e., getting a P-value ≤ 0.05) is only 0.011. That is, the actual P(Type I error) $= 0.011$, much smaller than the intended significance level of 0.05.

This illustrates an awkward aspect of small-sample significance testing when the test statistic has a discrete distribution. Imagine how a P-value, regarded as a random variable, may vary from study to study. For test statistics having a *continuous* distribution, the P-value has a *uniform* null distribution over the interval [0, 1]. That is, when H_0 is true, the P-value is equally likely to fall anywhere between 0 and 1. Then, the probability that the P-value falls below 0.05 equals exactly 0.05. The expected value of the P-value, that is, its long-run average value, is exactly 0.50. By contrast, for a test statistic having a *discrete* distribution, the null distribution of the P-value is discrete and has an expected value greater than 0.50 (e.g., it can equal 1.00 but never exactly 0.00). In this average sense, ordinary P-values for discrete distributions tend to be too large.

To address the conservatism difficulty, with discrete data we recommend using a different type of P-value. Called the *mid P-value*, it adds only *half* the probability of the observed result to the probability of the more extreme results. To illustrate, with $y = 9$ successes in $n = 10$ trials, the ordinary P-value for H_a: $\pi > 0.50$ is $P(9) + P(10) = 0.010 + 0.001 = 0.011$. The mid P-value is $[P(9)/2] + P(10) = (0.010/2) + 0.001 = 0.006$. The two-sided mid P-value for H_a: $\pi \neq 0.50$ is 0.012. The mid P-value has a null expected value of 0.50, the same as the regular P-value for test statistics that have a continuous distribution. Also, the two separate one-sided mid P-values sum to 1.0. By contrast, the observed result has probability counted in each tail for the ordinary one-sided P-values, so the two one-sided P values have a sum exceeding 1.

Inference based on the mid P-value compromises between the conservativeness of small-sample methods and the potential inadequacy of large-sample methods. It is also possible to construct a confidence interval for π from the set of π_0 values not rejected in the corresponding binomial test using the mid P-value. We shall do this with software in Section 1.6. In that section, we will see that it is straightforward to use software to obtain all the results for the examples in this chapter.

1.5 BAYESIAN INFERENCE FOR PROPORTIONS *

This book mainly uses the traditional, so-called *frequentist*, approach to statistical inference. This approach treats parameter values as fixed and data as realizations of random variables that have some assumed probability distribution. That is, probability statements refer to possible values for the data, given the parameter values. Recent years have seen increasing popularity of the *Bayesian* approach, which also treats parameters as random variables and therefore has probability distributions for them as well as for the data. This yields inferences

in the form of probability statements about possible values for the parameters, given the observed data.

1.5.1 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 non-informative so that inferential results are more objective, based almost entirely on the data. The prior distribution combines with the information that the data provide through the likelihood function to generate a *posterior distribution* for the parameters. The posterior distribution reflects the information about the parameters based both on the prior distribution and the data observed in the study.

For a parameter β and data denoted by y, let $f(\beta)$ denote the probability function[10] for the prior distribution of β. For example, when β is the binomial parameter π, this is a probability distribution over the interval [0, 1] of possible values for the probability π. Also, let $p(y \mid \beta)$ denote the probability function for the data, given the parameter value. (The vertical slash $\mid$ symbolizes "given" or "conditional on.") An example is the binomial formula (1.1), treating it as a function of y for fixed π. Finally, let $g(\beta \mid y)$ denote the probability function for the posterior distribution of β after we observe the data. In these symbols, from *Bayes' Theorem*,

$$g(\beta \mid y) \text{ is proportional to } p(y \mid \beta) f(\beta).$$

Now, after we observe the data, $p(y \mid \beta)$ is the likelihood function $\ell(\beta)$ when we view it as a function of the parameter. Therefore, the posterior distribution of the parameter is determined by the product of the likelihood function with the probability function for the prior distribution. When the prior distribution is relatively flat, as data analysts often choose in practice, the posterior distribution for the parameter has a similar shape to the likelihood function.

Except in a few simple cases, such as presented next for the binomial parameter, the posterior distribution cannot be easily calculated and software uses simulation methods to approximate it. The primary method for doing this is called *Markov chain Monte Carlo* (MCMC). It is beyond our scope to discuss the technical details of how an MCMC algorithm works. In a nutshell, software generates a very long sequence of values taken from an approximation for the posterior distribution. The data analyst takes the sequence to be long enough so that the Monte Carlo error is small in approximating the posterior distribution and summary measures of it, such as the mean.

For a particular parameter, Bayesian inference methods using the posterior distribution parallel those for frequentist inference. For example, analogous to the frequentist 95% confidence interval, we can construct an interval that contains 95% of the posterior distribution. Such an interval is referred to as a *posterior interval* or *credible interval*. A simple posterior interval uses percentiles of the posterior distribution, with equal probabilities in the two tails. For example, the 95% equal-tail posterior interval for a parameter is the region between the 2.5 and 97.5 percentiles of the posterior distribution. The mean of the posterior

[10] For a continuous distribution such as the normal, this is called the *probability density function*.

distribution is a Bayesian point estimator of the parameter. In lieu of P-values, posterior tail probabilities are useful, such as the posterior probability that an effect parameter in a model is positive.

1.5.2 Bayesian Binomial Inference: Beta Prior Distributions

Bayesian inference for a binomial parameter π can use a *beta distribution* as the prior distribution. The beta probability density function for π is proportional to

$$f(\pi) \ \propto \ \pi^{\alpha-1}(1-\pi)^{\beta-1}, \quad 0 \le \pi \le 1.$$

The distribution depends on two indices $\alpha > 0$ and $\beta > 0$, which are often referred to as *hyperparameters* to distinguish them from the parameter π that is the object of the inference. The mean of the beta distribution is

$$E(\pi) = \alpha/(\alpha+\beta).$$

The family of beta probability density functions has a wide variety of shapes.[11] When $\alpha = \beta$, it is symmetric around 0.50. The *uniform distribution*, $f(\pi) = 1$ over [0, 1], spreads the mass uniformly over the interval. It is the special case of a beta distribution with $\alpha = \beta = 1$. The beta density has a bimodal U-shape when $\alpha = \beta < 1$ and a bell shape when $\alpha = \beta > 1$. The variability decreases as $\alpha = \beta$ increases.

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 so-called *Jeffreys prior*, for which prior distributions for different scales of measurement for the parameter (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 $\alpha = \beta = 0.5$, which has a symmetric U-shape. Although it is not flat, this prior distribution is relatively noninformative, in the sense that it has greater variability than the uniform distribution and yields inferential results similar to those of the best frequentist methods. For example, its posterior intervals have actual coverage probability close to the nominal level.[12] Unless you have reason to use something else, we recommend using it or the uniform prior distribution.

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, β) prior distribution with a binomial likelihood function, the posterior distribution is a beta distribution indexed by $\alpha^* = y + \alpha$ and $\beta^* = n - y + \beta$. The Bayesian point estimate of π is the mean of this posterior distribution,

$$\frac{\alpha^*}{\alpha^*+\beta^*} = \frac{y+\alpha}{n+\alpha+\beta} = \left(\frac{n}{n+\alpha+\beta}\right)\frac{y}{n} + \left(\frac{\alpha+\beta}{n+\alpha+\beta}\right)\frac{\alpha}{\alpha+\beta}.$$

This estimate is a weighted average of the sample proportion $\hat{\pi} = y/n$ and the mean of the prior distribution, $\alpha/(\alpha+\beta)$. The weight $n/(n+\alpha+\beta)$ given to the sample proportion

[11] For graphs of such shapes, see `https://en.wikipedia.org/wiki/Beta_distribution`.
[12] For example, Chapter 6 of *Statistical Intervals* by W. Meeker, G. Hahn, and L. Escobar (Wiley, 2017).

increases toward 1 as n increases. With $\alpha = \beta$, the estimate shrinks the sample proportion toward 0.50. To construct the equal-tail 95% posterior interval, software uses the 2.5 and 97.5 percentiles of this posterior beta distribution.

1.5.3 Example: Opinions about Legalized Abortion, Revisited

In Section 1.3.3 we estimated the population proportion of Americans who support unrestricted legalized abortion. For a sample of $n = 1810$ people, $y = 837$ were in support and $n - y = 973$ were not. The ML estimate of π is $\hat{\pi} = 0.462$ and the 95% score confidence interval is (0.440, 0.485). How does this compare to Bayesian point and interval estimates?

For the Jeffreys beta(0.5, 0.5) prior distribution with $y = 837$ and $n - y = 973$, the posterior distribution is beta(α^*, β^*) with $\alpha^* = y + \alpha = 837.5$ and $\beta^* = n - y + \beta = 973.5$. The posterior mean estimate of π is $\alpha^*/(\alpha^* + \beta^*) = 837.5/(837.5 + 973.5) = 0.462$. Software (e.g., as shown in Section 1.6.2) reports the posterior 95% equal-tail interval of (0.440, 0.485), the endpoints being the 2.5 and 97.5 percentiles of the beta posterior density.

The Bayesian point estimate and posterior interval are the same, to three decimal places, as the ML estimate and frequentist 95% score interval. Frequentist and Bayesian inferences tend to be very similar when n is large and the prior distribution is highly disperse. However, the interpretations are quite different. With the frequentist approach, the actual parameter value π either *is* or *is not* in the confidence interval of (0.440, 0.485). Our 95% confidence means that if we used this method over and over with separate, independent samples, in the long run 95% of the confidence intervals would contain π. That is, the probability applies to possible data in future samples, not to the parameter. By contrast, with the Bayesian approach, after observing the data, we can say that the probability is 0.95 that π falls between 0.440 and 0.485.

The ordinary frequentist P-value for the score test (Section 1.3.2) of H_0: $\pi = 0.50$ against H_a: $\pi < 0.50$ is 0.000695. For such a one-sided test, the implicit null hypothesis is H_0: $\pi \geq 0.50$, and we use the boundary value to form the test statistic. A corresponding Bayesian posterior probability of interest is $P(\pi \geq 0.50)$, which equals 0.000692. The frequentist interpretation is that if H_0 were true (i.e., if $\pi = 0.50$), the probability of getting a test statistic like the observed one or even more extreme in the direction of H_a is 0.000695. This is a probability about potential data, given a parameter value. By contrast, the Bayesian interpretation of the posterior probability is that, after observing the data, the probability that $\pi \geq 0.50$ is 0.000692. With highly disperse prior distributions, such a one-tail probability is approximately equal to the frequentist one-sided P-value.

1.5.4 Other Prior Distributions

Bayesian methods for binomial parameters can use prior distributions other than the beta distribution. One possibility, hierarchical in nature, also assumes prior distributions for the beta hyperparameters instead of assigning fixed values.

For a prior distribution for $c > 2$ multinomial parameters, the beta distribution generalizes to the *Dirichlet distribution*. It is defined over the simplex of nonnegative values $(\pi_1, \ldots, \pi_c)$ that sum to 1. The posterior distribution is then also Dirichlet.

Models presented in this book have effect parameters that can take any real-number value. Bayesian methods for such parameters typically use normal prior distributions.

1.6 USING R SOFTWARE FOR STATISTICAL INFERENCE ABOUT PROPORTIONS *

Statistical software packages can implement categorical data analyses. The software package R is increasingly popular, partly because it is available to download for free at www.r-project.org and partly because users can contribute their own functions to make new methods available. Throughout the text, we show the R code and output for the examples. The Appendix shows examples for SAS, Stata, and SPSS software.

You can get help about R at many sites on the Internet, such as https://stats.idre.ucla.edu/r. Many users prefer to use RStudio, an integrated development environment (IDE) for R. It includes a code editor and tools for debugging and plotting. See www.rstudio.com and community.rstudio.com.

For a particular function or command, using R you can get help by placing a ? before its name, for example, entering

```
> ?prop.test
```

to get information about the function prop.test that can perform inference about proportions.

1.6.1 Reading Data Files and Installing Packages

In this book, we show R statements that request statistical analyses from the command line. A basic command loads a data file from the text website,[13] www.stat.ufl.edu/~aa/cat. For example, the data file called Clinical.dat at the text website has the observations for the clinical trials toy example in Section 1.3.5, with 9 successes in 10 observations. Here is how to create an R data file with the name Clinical from that data file at the text website and then view the file:

```
------------------------------------------------------------------------------
> Clinical <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Clinical.dat",
+                        header=TRUE)
> Clinical
   subject response
1        1        1
2        2        1
3        3        1
4        4        1
5        5        1
6        6        1
7        7        1
8        8        1
9        9        1
10      10        0
------------------------------------------------------------------------------
```

The *header=TRUE* part of the read.table command tells R that the first row of the data file contains the variable names. The standard form for a data file has a separate row for each subject (e.g., person) in the sample and a separate column for each variable. Here, the

[13] You can also copy the data files from https://github.com/alanagresti/categorical-data.

column labelled *response* shows the 0 and 1 values for failure and success indications of a binary outcome.

Many users of R have created packages that can perform analyses not available in basic R. For example, to install the binom package that can conduct some statistical inference for binomial parameters, use the command

```
> install.packages("binom")
```

Once it is installed, load the package to access its functions:

```
> library(binom)
```

1.6.2 Using R for Statistical Inference about Proportions

For the opinions about legalized abortion example (Section 1.3.3) with a binomial count of 837 supporting unrestricted legalization out of $n = 1810$, here is how you can conduct two-sided and one-sided significance tests and a confidence interval for the population proportion:

```
------------------------------------------------------------------------------
> prop.test(837, 1810, p=0.50, alternative="two.sided", correct=FALSE)
data:  837 out of 1810, null probability 0.5
X-squared = 10.219, df = 1, p-value = 0.00139 # chi-squared, 2-sided altern.
alternative hypothesis: true p is not equal to 0.5 # "true p" is binom. para.
95 percent confidence interval:# score CI
 0.4395653 0.4854557

> prop.test(837, 1810, p=0.50, alternative="less", correct=FALSE)
data:  837 out of 1810, null probability 0.5
X-squared = 10.219, df = 1, p-value = 0.000695 # chi-squared, 1-sided altern.
alternative hypothesis: true p is less than 0.5
------------------------------------------------------------------------------
```

The *X-squared* value in the R output is the *score* test statistic, having a chi-squared null distribution with $df = 1$. The z test statistic (1.2) that we presented is the positive or negative square root of this value. The *correct=FALSE* option stops R from using a *continuity correction*. We do not recommend using continuity corrections, because inferences then tend to be too conservative. The confidence interval displayed is the *score* confidence interval introduced in Section 1.3.5.

We can also request analyses directly from a data file. Suppose we load a data file in which one column contains 0 and 1 values for failure and success indications of a binary outcome. We can define a binomial variable that sums these indicators to obtain the number of successes and then conduct inferences using it. For example, for the Clinical data file for the 9 successes in 10 observations for the clinical trials example in Section 1.3.5, here is how we could find the score confidence interval quoted in that example:

```
------------------------------------------------------------------------------
> Clinical
   subject response
1        1        1
2        2        1
```

```
...
10      10        0
> attach(Clinical)
> y <- sum(response) # sums 0 and 1 values to get number of successes
> prop.test(y, n=10, conf.level=0.95, correct=FALSE)
95 percent confidence interval:
 0.59585  0.98212 # score CI for probability of success
-----------------------------------------------------------------------------
```

Rather than attach a data file, which can cause confusion if a variable has already been defined in the R session that has the same name as a variable in that data file, you can refer to the data file name in the command itself or you can instead imbed the data file name and desired command in a `with` function:

```
-----------------------------------------------------------------------------
> prop.test(sum(Clinical$response), 10, correct=FALSE)$conf.int
> with(Clinical, prop.test(sum(response), 10, correct=FALSE)$conf.int)
-----------------------------------------------------------------------------
```

The Wald confidence interval ("asymptotic"), score interval ("wilson," named after the statistician who first proposed this interval), and Agresti–Coull interval (near the end of Section 1.3.5) are available with the `binom` package. Here we show them for 9 successes in 10 trials:

```
-----------------------------------------------------------------------------
> library(binom)
> binom.confint(9, 10, conf.level=0.95, method="asymptotic")
      method x  n  mean    lower    upper
  asymptotic 9 10   0.9  0.71406  1.08594 # Wald confidence interval
> binom.confint(9, 10, conf.level=0.95, method="wilson")
  method x  n  mean    lower    upper
  wilson 9 10   0.9  0.59585  0.98212 # score confidence interval
> binom.confint(9, 10, conf.level=0.95, method="agresti-coull")
         method x  n  mean    lower    upper
  agresti-coull 9 10   0.9  0.57403  1.00394
-----------------------------------------------------------------------------
```

For any upper bound reported above 1.0, you should truncate it at 1.000, since π must fall between 0 and 1. Likelihood-ratio tests and corresponding confidence intervals are easy to obtain in R in the context of modeling, as we will see in future chapters.

The `binom.test` function uses the binomial distribution to obtain exact P-values. For example, with $y = 9$ and $n = 10$, we find the P-value for H_0: $\pi = 0.50$ against H_a: $\pi \neq 0.50$ and against H_a: $\pi > 0.50$:

```
-----------------------------------------------------------------------------
> binom.test(9, 10, 0.50, alternative = "two.sided")
        Exact binomial test
number of successes = 9, number of trials = 10, p-value = 0.02148
alternative hypothesis: true probability of success is not equal to 0.5
```

```
> binom.test(9, 10, 0.50, alternative = "greater")
        Exact binomial test
number of successes = 9, number of trials = 10, p-value = 0.01074
alternative hypothesis: true probability of success is greater than 0.5
------------------------------------------------------------------------
```

Binomial tests using the mid P-value are available with the `exactci` package. Binomial confidence intervals using the mid P-value are available with that package and with the `PropCIs` package. Again, here are results for $y = 9$ in $n = 10$ trials:

```
------------------------------------------------------------------------
> library(exactci)
> binom.exact(9, 10, 0.50, alternative="greater", midp=TRUE) # mid P-value
number of successes = 9, number of trials = 10, p-value = 0.005859
> library(PropCIs)
> midPci(9, 10, 0.95) # confidence interval based on test with mid P-value
 0.5966 0.9946
------------------------------------------------------------------------
```

A Bayesian posterior interval based on a beta(α, β) prior distribution and y successes and $n - y$ failures finds percentiles of the beta distribution with parameters $\alpha^* = y + \alpha$ and $\beta^* = n - y + \beta$ using the `qbeta` quantile function. We find a 95% posterior interval here using $\alpha = \beta = 0.5$ with $y = 837$ and $n - y = 973$, as in the opinion about legalized abortion example (Section 1.5.3). We can use the `pbeta` cumulative probability function to find a tail probability, such as the posterior $P(\pi \geq 0.50)$:

```
------------------------------------------------------------------------
> qbeta(c(0.025, 0.975), 837.5, 973.5)
[1] 0.43954  0.48545 # bounds of posterior interval, for Jeffreys prior

> pbeta(0.50, 837.5, 973.5) # posterior beta cumulative prob. at 0.50
[1] 0.99931
> 1 - pbeta(0.50, 837.5, 973.5) # right-tail probability above 0.50
[1] 0.00069 # can also get as pbeta(0.50, 837.5, 973.5, lower.tail=FALSE)
------------------------------------------------------------------------
```

1.6.3 Summary: Choosing an Inference Method

In summary, several methods are available for conducting inference about a binomial parameter. It can be confusing for a methodologist to decide which to use. With modern computing power, it is no longer necessary in this simple setting to rely on methods based on large-sample normal or chi-squared approximations. For a frequentist approach to significance testing or confidence intervals, we recommend exact binomial inference using the mid P-value. For the Bayesian approach, we recommend inference using the beta posterior distribution induced by a beta(0.5, 0.5) prior distribution.

EXERCISES

1.1 In the following examples, identify the natural response variable and the explanatory variables.

a. Attitude toward gun control (favor, oppose), gender (female, male), mother's education (high school, college).

b. Heart disease (yes, no), blood pressure, cholesterol level.

c. Race (white, nonwhite), religion (Catholic, Jewish, Muslim, Protestant, none), vote for president (Democrat, Republican, Green), annual income.

1.2 Which scale of measurement is most appropriate for the following variables — nominal or ordinal?

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

b. Highest educational degree obtained (none, high school, bachelor's, master's, doctorate).

c. Patient condition (good, fair, serious, critical).

d. Hospital location (London, Boston, Madison, Rochester, Toronto).

e. Favorite beverage (beer, juice, milk, soft drink, wine, other).

f. Rating of a movie with 1 to 5 stars, representing (hated it, didn't like it, liked it, really liked it, loved it)

1.3 Each of 100 multiple-choice questions on an exam has four possible answers but one correct response. For each question, a student randomly selects one response as the answer.

a. Specify the probability distribution of the student's number of correct answers on the exam.

b. Based on the mean and standard deviation of that distribution, would it be surprising if the student made at least 50 correct responses? Explain your reasoning.

1.4 In a particular city, the population proportion π supports an increase in the minimum wage. For a random sample of size 2, let $Y =$ number who support an increase.

a. Assuming $\pi = 0.50$, specify the probabilities for the possible values y for Y and find the distribution's mean and standard deviation.

b. Suppose you observe $y = 1$ and do not know π. Find and sketch the likelihood function. Using the plotted likelihood function, explain why the ML estimate $\hat{\pi} = 0.50$.

1.5 Refer to the previous exercise. Suppose $y = 0$ for $n = 2$. Find the ML estimate of π. Does this estimate seem believable? Why? Find the Bayesian estimator based on the prior belief that π is equally likely to be anywhere between 0 and 1.

1.6 Genotypes AA, Aa, and aa occur with probabilities (π_1, π_2, π_3). For $n = 3$ independent observations, the observed frequencies are (y_1, y_2, y_3).

a. Explain how you can determine y_3 from knowing y_1 and y_2. Thus, the multinomial distribution of (y_1, y_2, y_3) is actually two-dimensional.

b. Show the ten possible observations (y_1, y_2, y_3) with $n = 3$.

c. Suppose $(\pi_1, \pi_2, \pi_3) = (0.25, 0.50, 0.25)$. What probability distribution does y_1 alone have?

1.7 In his autobiography *A Sort of Life*, British author Graham Greene described a period of severe mental depression during which he played Russian Roulette — 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 he had kept playing this game until the bullet fires. Let Y denote the number of the game on which the bullet fires. Explain why the probability of the outcome y equals $(5/6)^{y-1}(1/6)$, for $y = 1, 2, 3, \ldots$. (This is called the *geometric distribution*.)

1.8 When the 2010 General Social Survey asked subjects in the US whether they would be willing to accept cuts in their standard of living to protect the environment, 486 of 1374 subjects said *yes*.

a. Estimate the population proportion who would say *yes*. Construct and interpret a 99% confidence interval for this proportion.

b. Conduct a significance test to determine whether a majority or minority of the population would say *yes*. Report and interpret the P-value.

1.9 A study of 100 women suffering from excessive menstrual bleeding considers whether a new analgesic provides greater relief than the standard analgesic. Of the women, 40 reported greater relief with the standard analgesic and 60 reported greater relief with the new one.

a. Test the hypothesis that the probability of greater relief with the standard analgesic is the same as the probability of greater relief with the new analgesic. Report and interpret the P-value for the two-sided alternative. (*Hint:* Express the hypotheses in terms of a single parameter. A test to compare matched-pairs responses in terms of which is better is called a *sign test*.)

b. Construct and interpret a 95% confidence interval for the probability of greater relief with the new analgesic.

1.10 Refer to the previous exercise. The researchers wanted a sufficiently large sample to be able to estimate the probability of preferring the new analgesic to within 0.08, with confidence 0.95. If the true probability is 0.75, how large a sample is needed to achieve this accuracy? (*Hint:* For how large an n does a 95% confidence interval have margin of error equal to about 0.08?)

1.11 When a recent General Social Survey asked 1158 American adults, "Do you believe in heaven?", the proportion who answered yes was 0.86. Treating this as a random sample, conduct statistical inference about the population proportion of American adults believing in heaven. Summarize your analysis and interpret the results in a short report.

1.12 To collect data in an introductory statistics course, I gave the students a questionnaire. One question asked whether the student was a vegetarian. Of 25 students, 0 answered

yes. They were not a random sample, but use these data to illustrate inference for a proportion. Let π denote the population proportion who would say *yes*. Consider H_0: $\pi = 0.50$ and H_a: $\pi \neq 0.50$.

a. What happens when you conduct the *Wald test*, which uses the *estimated* standard error in the z test statistic?

b. Find the 95% *Wald confidence interval* (1.3) for π. Is it believable?

c. Conduct the *score test*, which uses the *null* standard error in the z test statistic. Report and interpret the P-value.

d. Verify that the 95% score confidence interval equals (0.0, 0.133). (This is similar to the interval (0.0, 0.137) obtained with a small-sample method of Section 1.4.3, inverting the binomial test with the mid P-value.)

1.13 Refer to the previous exercise, with $y = 0$ in $n = 25$ trials for testing H_0: $\pi = 0.50$.

a. Show that ℓ_0, the maximized likelihood under H_0, equals $(1 - \pi_0)^{25} = (0.50)^{25}$. Show that ℓ_1, the maximized likelihood over all possible π values, equals 1.0. (*Hint:* This is the value at the ML estimate value of 0.0.)

b. Show that the likelihood-ratio test statistic, $2\log(\ell_1/\ell_0)$, equals 34.7. Report the P-value.

c. The 95% likelihood-ratio-test-based confidence interval for π is (0.000, 0.074). Verify that 0.074 is the correct upper bound by showing that the likelihood-ratio test of H_0: $\pi = 0.074$ against H_a: $\pi \neq 0.074$ has a chi-squared test statistic equal to 3.84 and P-value $= 0.05$.

1.14 Section 1.4.3 found binomial P-values for a clinical trial with $y = 9$ successes in 10 trials. Suppose instead $y = 8$. Using software or the binomial distribution shown in Table 1.1:

a. Find the P-value for (i) H_a: $\pi > 0.50$, (ii) H_a: $\pi < 0.50$.

b. Find the mid P-value for (i) H_a: $\pi > 0.50$, (ii) H_a: $\pi < 0.50$.

c. Why is the sum of the one-sided P-values greater than 1.0 for the ordinary P-value but equal to 1.0 for the mid P-value?

d. Using software, find the 95% confidence interval based on the binomial test with the mid P-value.

1.15 If Y is a random variable and c is a positive constant, then the standard deviation of the probability distribution of cY equals $c\sigma(Y)$. Suppose Y is a binomial variate and let $\hat{\pi} = Y/n$.

a. Based on the binomial standard deviation for Y, show that $\sigma(\hat{\pi}) = \sqrt{\pi(1-\pi)/n}$.

b. Explain why it is easier to estimate π precisely when it is near 0 or 1 than when it is near 0.50.

1.16 Using calculus, it is easier to derive the maximum of the log of the likelihood function, $L = \log \ell$, than the likelihood function ℓ itself. Both functions have a maximum at the same value, so it is sufficient to do either.

a. Calculate the log-likelihood function $L(\pi)$ for the binomial distribution (1.1).

b. One can usually determine the point at which the maximum of a log-likelihood L occurs by solving the *likelihood equation*. This is the equation resulting from

differentiating L with respect to the parameter and setting the derivative equal to zero. Find the likelihood equation for the binomial distribution and solve it to show that the ML estimate is $\hat{\pi} = y/n$.

1.17 Refer to Exercise 1.12 on estimating the population proportion π of vegetarians. For the beta(0.5, 0.5) prior, find the Bayes estimator of π, show that the posterior 95% interval is (0.00002, 0.0947), and show that the posterior $P(\pi < 0.50) = 1.000$.

1.18 For the previous exercise, explain how the Bayes estimator shrinks the sample proportion toward the prior mean.

1.19 For Exercises 1.12 and 1.17, explain the difference between the frequentist interpretation of the score confidence interval (0.0, 0.133) and the Bayesian interpretation of the posterior interval (0.00002, 0.0947).