Chapter 2

The Dichotomous Data Problem

INTRODUCTION

In this chapter the primary focus is on the dichotomous data problem. The data consists of n independent repeated Bernoulli trials having constant probability of success p. On the basis of these outcomes, we wish to make inferences about p. Section 2.1 introduces the binomial distribution and presents a binomial test for the hypothesis $p = p_0$, where p_0 is a specified success probability. Section 2.2 gives a point estimator $\widehat{p}$ for p. Section 2.3 presents confidence intervals for p. Section 2.3 also contains the generalization of the binomial distribution to the multinomial distribution, confidence intervals for multinomial probabilities and a test that the multinomial probabilities are equal to specified values. Section 2.4 presents Bayesian competitors to the frequentist estimator $\widehat{p}$ of Section 2.2. The Bayesian estimators incorporate prior information.

Data. We observe the outcomes of n independent repeated Bernoulli trials.

Assumptions

A1. The outcome of each trial can be classified as a success or a failure.

A2. The probability of a success, denoted by p, remains constant from trial to trial.

A3. The n trials are independent.

2.1 A BINOMIAL TEST

Procedure

To test

$$H_0 : p = p_0, \tag{2.1}$$

where p_0 is some specified number, $0 < p_0 < 1$, set

$$B = \text{number of successes}. \tag{2.2}$$

a. *One-Sided Upper-Tail Test.* To test

$$H_0 : p = p_0$$

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

versus

$$H_1 : p > p_0$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } B \geq b_\alpha; \text{ otherwise do not reject,} \tag{2.3}$$

where the constant b_α is chosen to make the type I error probability equal to α. The number b_α is the upper α percentile point of the binomial distribution with sample size n and success probability p_0. Due to the discreteness of the binomial distribution, not all values of α are available (unless one resorts to randomization). Comment 3 explains how to obtain the b_α values. See also Example 2.1.

b. *One-Sided Lower-Tail Test.* To test

$$H_0 : p = p_0$$

versus

$$H_2 : p < p_0$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } B \leq c_\alpha; \text{ otherwise do not reject.} \tag{2.4}$$

Values of c_α can be determined as described in Comment 3. Here, c_α is the lower α percentile point of the binomial distribution with sample size n and success probability p_0. For the special case of testing $p = \frac{1}{2}$,

$$c_\alpha = n - b_\alpha. \tag{2.5}$$

Equation 2.5 is explained in Comment 4.

c. *Two-Sided Test.* To test

$$H_0 : p = p_0$$

versus

$$H_3 : p \neq p_0$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } B \geq b_{\alpha_1} \text{ or } B \leq c_{\alpha_2}; \text{ otherwise do not reject,} \tag{2.6}$$

where b_{α_1} is the upper α_1 percentile point, c_{α_2} is the lower α_2 percentile point, and $\alpha_1 + \alpha_2 = \alpha$. See Comment 3.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic normality of B, suitably standardized. To standardize, we need to know the mean and variance of B when the null hypothesis is true. When H_0 is true, the mean and variance of B are, respectively,

$$E_{p_0}(B) = np_0, \tag{2.7}$$

$$\text{var}_{p_0}(B) = np_0(1 - p_0). \tag{2.8}$$

Comment 8 gives the derivations for (2.7) and (2.8).

The standardized version of B is

$$B^* = \frac{B - E_{p_0}(B)}{\{\text{var}_{p_0}(B)\}^{1/2}} = \frac{B - np_0}{\{np_0(1 - p_0)\}^{1/2}}. \tag{2.9}$$

When H_0 is true, B^* has, as n tends to infinity, an asymptotic $N(0,1)$ distribution. Let z_α denote the upper α percentile point of the $N(0,1)$ distribution. To find z_α, we use the `qnorm`(1-α,0,1). For example, to find $z_{.05}$, we apply qnorm(.95,0,1) and obtain $z_{.05} = 1.645$.

The normal approximation to procedure (2.3) is

$$\text{Reject } H_0 \text{ if } B^* \geq z_\alpha; \text{ otherwise do not reject.} \tag{2.10}$$

The normal approximation to procedure (2.4) is

$$\text{Reject } H_0 \text{ if } B^* \leq -z_\alpha; \text{ otherwise do not reject.} \tag{2.11}$$

The normal approximation to procedure (2.6), with $\alpha_1 = \alpha_2 = \alpha/2$, is

$$\text{Reject } H_0 \text{ if } |B^*| \geq z_{\alpha/2}; \text{ otherwise do not reject.} \tag{2.12}$$

EXAMPLE 2.1 ***Canopy Gap Closure.***

Dickinson, Putz, and Canham (1993) investigated canopy gap closure in thickets of the clonal shrub *Cornus racemosa.* Shrubs often form dense clumps where tree abundance has been kept artificially low (e.g., on power-line right of ways). These shrub clumps then retard reinvasion of the sites by trees. Individual clumps may persist for many years. Clumps outlast the lives of the individual stems of which they are formed; stems die and leave temporary holes in the canopies of the clumps. Closure of the hope (gap) left by dead stems occurs in part by the lateral growth of stems that surround the hole. Opening of the gap often occurs when individual branches of hole-edge stems die. Between sample dates, more branches in six out of seven gaps in clumps, at a site with nutrient-poor and dry soil, died than lived. Let us say we have a success if more branches die than live in the gaps in clumps. Let p denote the corresponding probability of success. We suppose that the success probability for sites that are nutrient rich with moist soil has been established by previous studies to be 15%. Do the nutrient-poor and dry soil sites have the same success probability as the nutrient-rich and moist soil sites or is it larger? This reduces to the hypothesis-testing problem

$$H_0 : p = .15$$

versus

$$H_1 : p > .15.$$

Our sample size is $n = 7$ and we observe $B = 6$ successes. From the `R` command `round(pbinom(0:7,7,.15,lower.tail=F),4)`, we obtain, rounded to four places, the probabilities $P_{.15}(B > x)$ for $x = 0, \ldots, 7$. (The notation $P_{.15}(B > x)$ is shorthand for the probability that $B > x$, computed under the assumption that the true success probability is .15.) The $P_{.15}(B > x)$ probabilities are

x	0	1	2	3	4	5	6	7
$P_{.15}(B > x)$	.6794	.2834	.0738	.0121	.0012	.0001	.0000	.0000

To find $P_{.15}(B \geq x)$ note $P_{.15}(B \geq x) = P_{.15}(B > x - 1)$. Reasonable possible choices for α are .0738, .0121, .0012, .0001. Suppose we choose to use $\alpha = .0121$. We note $P(B > 3) = P(B \geq 4) = .0121$ and thus we see $b_{.0121} = 4$. Thus the $\alpha = .0121$ test is

Reject H_0 if $B \geq 4$; otherwise do not reject.

Our observed value is $B = 6$ and thus we reject H_0 at $\alpha = .0121$. To find the P-value, which is $P_{.15}(B \geq 6)$, we can find $P_{.15}(B > 5)$ using the `R` command `pbinom(5,7,.15,lower.tail=F)`. Alternatively, we can find the P-value using the `R` command `binom.test(6,7,.15,"g")`. We find $P = .000069$, or rounded to four places, P is .0001. This is the smallest significance level at which we can reject H_0 (in favor of the alternative $p > .15$) with our observed value of B. We conclude that there is strong evidence against H_0 favoring the alternative. For more on the P-value, see Comment 9.

EXAMPLE 2.2 *Sensory Difference Tests.*

Sensory difference tests are often used in quality control and quality evaluation. The triangle test (cf. Bradley, 1963) is a sensory difference test that provides a useful application of the binomial model. In its simplest form, the triangle procedure is as follows. To each of n panelists, three test samples are presented in a randomized order. Two of the samples are known to be identical; the third is different. The panelist is then supposed to select the odd sample, perhaps on the basis of a specified sensory attribute. If the panelists are homogeneous trained judges, the experiment can be viewed as n independent repeated Bernoulli trials, where a success corresponds to a correct identification of the odd sample. (If the panelists are not homogeneous trained judges, we may question the validity of Assumption A2.) Under the hypothesis that there is no basis for discrimination, the probability p of success is $\frac{1}{3}$, whereas a basis for discrimination would correspond to values of p that exceed $\frac{1}{3}$.

Byer and Abrams (1953) considered triangular bitterness tests in which each taster received three glasses, two containing the same quinine solution and the third a different quinine solution. In their first bitterness test, the solutions contained .0075% and .0050%, respectively, of quinine sulfate. The six presentation orders, LHH, HLH, HHL, HLL, LHL, and LLH (L denotes the lower concentration, H the higher concentration), were randomly distributed among the tasters. Out of 50 trials, there were 25 correct selections and 25 incorrect selections.

We consider the binomial test of $H_0 : p = \frac{1}{3}$ versus the one-sided alternative $p > \frac{1}{3}$ and use the large-sample approximation to (2.3). We set $\alpha = .05$ for purposes of illustration. To find $z_{.05}$, the 95^{th} quantile of the $N(0, 1)$ distribution, we use the `R` command `qnorm(.95,0,1)`, and find $z_{.05} = 1.645$. Thus approximation (2.10), at the $\alpha = .05$ level, reduces to

$$\text{Reject } H_0 \text{ if } B^* \geq 1.645; \;\text{ otherwise do not reject.}$$

From the data we have $n = 50$ and B (the number of correct identifications) $= 25$. Thus from (2.9), with $p_0 = \frac{1}{3}$, we obtain

$$B^* = \frac{25 - 50\left(\frac{1}{3}\right)}{\left\{50\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right\}^{1/2}} = 2.5.$$

The large sample approximation value $B^* = 2.5 > 1.645$ and thus we reject $H_0 : p = \frac{1}{3}$ in favor of $p > \frac{1}{3}$ at the approximate $\alpha = .05$ level. Thus there is evidence of a basis for discrimination in the taste bitterness test. To find the P-value corresponding to $B^* = 2.5$, one can use the `R` command `pnorm(2.5)`. The P-value is `1-pnorm(2.5)=.0062`. Thus, the smallest significance level at which we reject H_0 in favor of $p > \frac{1}{3}$ using the large-sample approximation is .0062. (Note the exact P-value in this case is given by R as `1-pbinom(24,50,1/3)=.0108`.)

Comments

1. *Binomial Test Procedures.* Assumptions A1–A3 are the general assumptions underlying a binomial experiment. Research problems possessing these assumptional underpinnings are common, and thus the binomial test procedures find frequent use. A particularly important special case in which procedures (2.3), (2.4), and (2.6) are applicable occurs when we wish to test hypotheses about the unknown median, θ, of a population. The application of binomial theory to this problem leads to a test statistic, B, that counts the number of sample observations larger than a specified null hypothesis value of θ, say θ_0. For this particular special case, the statistic B is referred to as the *sign statistic*, and the associated test procedures are referred to as sign test procedures. See Sections 3.4 and 3.8 for a more detailed discussion of the sign test procedures corresponding to (2.3), (2.4), and (2.6).
2. *Distribution-Free Test.* The critical constant b_α of (2.3) is chosen so that the probability of rejecting H_0, when H_0 is true, is α. We can control this type I error because Assumptions A1–A3 and a specification of p (the null hypothesis specifies p to be equal to p_0) determine, without further assumptions regarding the underlying populations from which the dichotomous data emanate, the probability distribution of B. Thus, under Assumptions A1–A3, the test given by (2.3) is

said to be a distribution-free test of H_0. The same statement can be made for tests (2.4) and (2.6).

3. *Illustration of Lower-Tail and Two-Tailed Tests.* Suppose $n = 8$ and we wish to test $H_0 : p = .4$ versus $p > .4$ via procedure (2.3). Using the methods illustrated in Example 2.1 to obtain binomial tail probabilities, we can find

b	0	1	2	3	4	5	6	7	8
$P_{.4}(B \geq b)$	1	.9832	.8936	.6846	.4059	.1737	.0498	.0085	.0007

(Recall that the $P_{.4}$ notation indicates that the probabilities are computed under the assumption that $p = .4$.) Hence, we can find constants b_α that satisfy the equation $P_{.4}\{B \geq b_\alpha\} = \alpha$ only for certain values of α. For $\alpha = .0085$, $b_{.0085} = 7$. For $\alpha = .0498$, $b_{.0498} = 6$. As α increases, the critical constant b_α decreases. Thus, when we increase α, it is easier to reject H_0; hence, we increase the power or, equivalently, decrease the probability of a type II error for our test (against a particular alternative). Similarly, if we lower α, we raise the probability of a type II error. This is illustrated in Comment 9.

Again consider the case $n = 8$ and suppose we want to test $p = .4$ versus the alternative $p < .4$. We can use the lower-tail test described by (2.4). For example, suppose we want $\alpha = .1064$. Then $P_{.4}\{B \geq 2\} = .8936$ and $P_{.4}\{B \leq 1\} = 1 - .8936 = .1064$. Thus, in (2.4), $c_{.1064} = 1$ and this yields the $\alpha = .1064$ test; namely, reject H_0 if $B \leq 1$ and accept H_0 if $B > 1$.

We close this comment with an example of the two-sided test described by (2.6). For convenience, we stay with the case $n = 8$ and test $H_0 : p = .4$. Note 6 is the upper .0498 percentile point of the null distribution of B and 1 is the lower .1064 percentile point. Thus the test that rejects H_0 when $B \geq 6$ or when $B \leq 1$ and accepts H_0 when $1 < B < 6$ is an $\alpha = .0498 + .1064 = .1562$ two-tailed test.

4. *Binomial Distribution.* The statistic B has been defined as the number of successes in n independent Bernoulli trials, each trial having a success probability equal to p. The distribution of the random variable B is known as the binomial distribution with parameters n and p.

For the special case when $p = \frac{1}{2}$, it can be shown that the distribution of B is symmetric about its mean $n/2$. This implies that

$$P_{.5}\{B \geq x\} = P_{.5}\{B \leq (n - x)\} \quad \text{for} \quad x = 0, \ldots, n. \tag{2.13}$$

Equation (2.13) implies that the lower α percentile point of the binomial distribution, with $p = .5$, is equal to n minus the upper α percentile point. This result was expressed by (2.5) after we introduced the lower-tail test given by (2.4).

5. *Motivation for the Test Based on B.* The statistic B/n is an estimator (see Section 2.2) of the true unknown parameter p. Thus, if $p > p_0$, B/n will tend to be larger than p_0. This suggests rejecting $H_0 : p = p_0$ in favor of $p > p_0$ for large values of B and serves as partial motivation for (2.3).

6. *An Example of the Exact Distribution of B.* The exact distribution of B can be obtained from the equation

$$B = \sum_{i=1}^{n} \psi_i, \tag{2.14}$$

where

$$\psi_i = \begin{cases} 1, & \text{if the } i\text{th Bernoulli trial is a success,} \\ 0, & \text{if the } i\text{th Bernoulli trial is failure.} \end{cases}$$

We consider the 2^n possible outcomes of the configurations $(\psi_1, \ldots, \psi_n)$ and use the fact that under H_0, any outcome with b 1's and $(n-b)$ 0's has probability $p^b(1-p)^{n-b}$. For example, in the case $n = 2$, $p = \frac{1}{4}$, the $2^2 = 4$ possible outcomes for (ψ_1, ψ_2) and associated values of B are as follows:

(ψ_1, ψ_2)	$P_{.25}\{(\psi_1, \psi_2)\}$	$B = \psi_1 + \psi_2$
$(0, 0)$	$\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^{2-0} = \frac{9}{16}$	0
$(0, 1)$	$\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^{2-1} = \frac{3}{16}$	1
$(1, 0)$	$\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^{2-1} = \frac{3}{16}$	1
$(1, 1)$	$\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^{2-2} = \frac{1}{16}$	2

Thus, for example, $P_{.25}\{B \geq 1\} = P_{.25}\{B = 1\} + P_{.25}\{B = 2\} = \frac{6}{16} + \frac{1}{16} = \frac{7}{16}$.

7. *The Exact Distribution of B.* By methods similar to the particular case illustrated in Comment 6, it can be shown that for each of the $n + 1$ possible values of B (namely, $b = 0, \ldots, n$), we have

$$P_p\{B = b\} = \binom{n}{b} p^b (1-p)^{n-b}. \tag{2.15}$$

In (2.15), the symbol $\binom{n}{b}$ (read "binomial *n, b*") is given by

$$\binom{n}{b} = \frac{n!}{b!(n-b)!}, \tag{2.16}$$

where the symbol $m!$ (read "m factorial") is, for positive integers, defined as $m! = m(m-1)(m-2)\ldots(3)(2)(1)$, and $0!$ is defined to be equal to 1. The number $\binom{n}{b}$ is known as the number of combinations of n things taken b at a time. It is equal to the number of subsets of size b that may be formed from the members of a set of size n. The distribution given by (2.15) is known as the binomial distribution with parameters n and p.

8. *The Asymptotic Distribution of B.* Using representation (2.14), we find the mean B is

$$E_p(B) = E_p\left(\sum_{i=1}^{n} \psi_i\right) = \sum_{i=1}^{n} E_p(\psi_i) = np,$$

where we have used the calculation

$$E_p(\psi_i) = 1 \cdot P(\psi_i = 1) + 0 \cdot P(\psi_i = 0) = 1 \cdot p + 0 \cdot (1 - p) = p.$$

Then, using the fact that $\psi_1, \psi_2, \ldots, \psi_n$ are independent,

$$\mathrm{var}_p(B) = \mathrm{var}_p\left(\sum_{i=1}^{n} \psi_i\right) = \sum_{i=1}^{n} \mathrm{var}_p(\psi_i). \tag{2.17}$$

The variance of any one of the indicator random variables ψ_i is determined as follows. Note $\psi_i^2 = \psi_i$ and thus

$$E_p(\psi_i^2) = E_p(\psi_i) = p,$$

and

$$\mathrm{var}_p(\psi_i) = E_p(\psi_i^2) - \{E_p(\psi_i)\}^2 = p - p^2 = p(1 - p).$$

Hence, from (2.17),

$$\mathrm{var}_p(B) = \sum_{i=1}^{n} p(1 - p) = np(1 - p).$$

The random variable B is a sum of independent and identically distributed random variables and hence the central limit theorem (cf. Casella and Berger, 2002, p. 236) establishes that, as $n \to \infty$, $(B - np)/\sqrt{np(1 - p)}$ has a limiting $N(0, 1)$ distribution.

9. *The P-Value.* Rather than specify an α level and report whether the test rejects at that specific α level, it is more informative to state the lowest significance level at which we can reject with the observed data. This is called the *P-value*. Consider the $\alpha = .0085$ test (test T_1, say) and the $\alpha = .0498$ test (T_2) of $H_0 : p = .4$ versus $p > .4$ for the case $n = 8$. Suppose in an actual experiment that our observed value of B is 7. Then with test T_2 we reject H_0 because the critical region for test T_2 consists of the values $\{B = 6, B = 7, B = 8\}$ and our observed value 7 is in the critical region. Thus, it is correct for us to state that the value $B = 7$ is significant at the $\alpha = .0498$ level. But the value $B = 7$ is also significant at the $\alpha = .0085$ level. If we simply state that we reject H_0 at the .0498 level, we do not convey the additional information that, with the value $B = 7$, we also can reject H_0 at the .0085 level. To remedy this, the following approach is suggested.

   Suppose, as in the previous example, large values of some statistic S (say) lead to rejection of the null hypothesis. Let s denote the observed value of S. Compute $P_0\{S \geq s\}$, the probability, under the null hypothesis, that S will be greater than or equal to s. This is the lowest level at which we can reject H_0. The observation $S = s$ will be significant at all levels greater than or equal to $P_0\{S \geq s\}$ and not significant at levels less than $P_0\{S \geq s\}$.

   To further illustrate this point, consider the test of $p = \frac{1}{3}$ versus $p > \frac{1}{3}$ of Example 2.2. We apply procedure (2.10), based on the large-sample approximation to the null distribution of B. The (approximate) $\alpha = .05$ test rejects if

$B^* \geq 1.645$ and accepts otherwise. The observed value of B^* is $B^* = 2.5$ and thus we can reject $p = \frac{1}{3}$ in favor of $p > \frac{1}{3}$ at the .05 level. In Example 2.2, we found $z_{.0062} = 2.5$. Thus, the smallest significance level at which we can reject is approximately .0062, and this statement is more informative than the statement that the .05 test leads to rejection.

10. *Calculating Power.* Take $n = 8$, and consider the following two tests of $H_0 : p = .4$ versus $p > .4$, based on (2.3). Test T_1, corresponding to $\alpha = .0085$, rejects H_0 if $B \geq 7$ and accepts otherwise. Test T_2, corresponding to $\alpha = .0498$, rejects H_0 if $B \geq 6$ and accepts otherwise. Suppose, in fact, that the alternative $p = .5$ is true. Let R_1 denote the power of the test T_1 (for this alternative) and let R_2 denote the power of the test T_2. Thus, R_1 is the probability of rejecting H_0 with test T_1 and R_2 is the probability of rejecting H_0 with test T_2. These powers are to be calculated when the alternative $p = .5$ is true. Using the `R` commands `pbinom(6,8,.5,lower.tail=F)` and `pbinom(5,8,.5,lower.tail=F)`, we obtain

$$R_1 = P_{.5}\{B \geq 7\} = P_{.5}\{B > 6\} = .0352$$

$$R_2 = P_{.5}\{B \geq 6\} = P_{.5}\{B > 5\} = .1445$$

For the alternative $p = .5$, let β_1 denote the probability of a type II error using test T_1 and let β_2 denote the probability of a type II error using test T_2. We find

$$\beta_1 = 1 - R_1 = .9648, \quad \beta_2 = 1 - R_2 = .8555.$$

Test T_1 has a lower probability of a type I error than test T_2, but the probability of a type II error for test T_1 exceeds that of test T_2. Incidentally, the reader should not be shocked at the very high values of β_1 and β_2. The alternative $p = .5$ is quite close to the null hypothesis value $p = .4$ and a sample of size 8 is simply not large enough to make a better (in terms of power) distinction between the hypothesis and alternative.

11. *More Power Calculations.* We return to Example 2.2 concerning sensory difference tests. Suppose we have $n = 50$ and we decide to employ the approximate $\alpha = .05$ level test of $H_0 : p = \frac{1}{3}$ versus $H_1 : p > \frac{1}{3}$. Recall that test rejects H_0 if

$$\frac{B - n\left(\frac{1}{3}\right)}{\{n\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\}^{1/2}} > 1.645$$

and accepts H_0 otherwise. What is the power of this test if in fact $p = .6$? We approximate the power using the asymptotic normality of B, suitably standardized. If $p = .6$, then

$$\frac{B - n(.6)}{\{n(.6)(.4)\}^{1/2}}$$

has an approximate $N(0, 1)$ distribution. Using this, we find

$$\begin{aligned}
\text{Power} &= P_{.6}\left\{\frac{B - n\left(\frac{1}{3}\right)}{\left\{n\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right\}^{1/2}} > 1.645\right\} \\
&= P_{.6}\left\{B > \left[(1.645)\left\{n\left(\tfrac{1}{3}\right)\left(\tfrac{2}{3}\right)\right\}^{1/2}\right] + n\left(\tfrac{1}{3}\right)\right\} \\
&= P_{.6}\left\{\frac{B - n(.6)}{\{n(.6)(.4)\}^{1/2}} > \left[\frac{(1.645)\left\{n\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right\}^{1/2} + n\left(\frac{1}{3}\right) - n(.6)}{\{n(.6)(.4)\}^{1/2}}\right]\right\} \\
&\doteq P\{Z > -2.27\},
\end{aligned}$$

where $Z = \{B - n(.6)\}/\{n(.6)(.4)\}^{1/2}$ is approximately a $N(0, 1)$ random variable and -2.27 is the value, when $n = 50$, of the term in large square brackets. Using `1-pnorm(-2.27)`, we find power $\doteq P\{Z > -2.27\} = .9884$.

12. *Counting Failures Instead of Successes.* Define B^- to be the number of failures in the n Bernoulli trials. Note that B^- could be defined by (2.14) with ψ_i replaced by $(1 - \psi_i)$, for $i = 1, \ldots, n$. Test procedures (2.3), (2.4), and (2.6) could equivalently be based on B^-, because $B^- = n - B$.

13. *Some History.* The binomial distribution has been utilized for statistical inferences about dichotomous data for more than 300 years. Binomial probability calculations were used by the British physician Arbuthnott (1710) in the early eighteenth century as an argument for the sexual balance maintained by Divine Providence and against the practice of polygamy. Bernoulli trials are so named in honor of Jacques Bernoulli. His book "Ars Conjectandi" (1713) contains a profound study of such trials and is viewed as a milestone in the history of probability theory. (LeCam and Neyman (1965) reported that the original Latin edition was followed by several in modern languages; the last reproduction, in German, appeared in 1899 in No. 107 and No. 108 of the series *Ostwald's Klassiker der exakten Wissenschaften*, Wilhelm Engelman, Leipzig.) Today, the binomial procedures remain one of the easiest and most useful sets of procedures in the statistical catalog.

Properties

1. *Consistency.* Test procedures (2.3), (2.4), and (2.6) will be consistent against alternatives for which $p >, <$, and $\neq p_0$, respectively.

Problems

1. Stanton (1969) investigated the problem of paroling criminal offenders. He studied the behavior of all male criminals paroled from New York's correctional institutions to original parole supervision during 1958 and 1959 (exclusive of those released to other warrants or to deportation). The parolees were observed for 3 years following their releases or until they exhibited some delinquent parole behavior. In a study involving a very large number of subjects, Stanton considered criminals convicted of crimes other than first- or second-degree murder. He found that approximately 60% of these parolees did not have any delinquent behavior during the 3 years following their releases.

During the same period, Stanton found that 56 of the 65 paroled murderers (first- or second-degree murderers who were also original parolees) in the study had no delinquent parole behavior. Let a success correspond to a male murderer on original parole who does not exhibit any delinquent parole behavior in the 3-year observation period. Note that we could question Assumptions A2 in this context; parolees convicted of first-degree murder may have a different success probability than parolees convicted of second-degree murder. Even the parolees in the first-degree (or second-degree) group may have different individual success probabilities. For pedagogical purposes, we proceed as if Assumption A2 is valid and denote the common success probability by p.

It is of interest to investigate whether murderers are better risks as original parolees than are criminals convicted of lesser crimes. This suggests testing $H_0 : p = .6$ against the alternative $p > .6$. Perform this test using the large-sample approximation to procedure (2.3).

2. Describe a situation in which Assumptions A1 and A2 hold but Assumption A3 is violated.

3. Describe a situation in which Assumptions A1 and A3 hold but Assumption A2 is violated.

4. Suppose that 10 Bernoulli trials satisfying Assumptions A1–A3 result in 8 successes. Investigate the accuracy of the large-sample approximation by comparing the smallest significance level at which we would reject $H_0 : p = \frac{1}{2}$ in favor of $p > \frac{1}{2}$ when using procedure (2.3) with the corresponding smallest significance level for the large-sample approximation to procedure (2.3) given by (2.10).

5. Return to the $\alpha = .0121$ test of Example 2.1. Recall that the test of $H_0 : p = .15$ versus $H_1 : p > .15$ rejects H_0 if in $n = 7$ trials there are 4 or more successes and accepts H_0 if there are 3 or fewer successes. What is the power of that test when (a) $p = .4$, (b) $p = .6$, and (c) $p = .8$?

6. A standard surgical procedure has a success rate of .7. A surgeon claims a new technique improves the success rate. In 20 applications of the new technique, there are 18 successes. Is there evidence to support the surgeon's claim?

7. A multiple-choice quiz contains ten questions. For each question there are one correct answer and four incorrect answers. A student gets three correct answers on the quiz. Test the hypothesis that the student is guessing.

8. Return to Example 2.2 and, in the case of $n = 50$, approximate the power of the $\alpha = .05$ test when $p = .5$.

9. Forsman and Lindell (1993) studied swallowing performance of adders (snakes). Captive snakes were fed with dead field voles (rodents) of differing body masses and the number of successful swallowing attempts was recorded. Out of 67 runs resulting in swallowing attempts, 58 where successful and 9 failed. (A failure was easy to detect because the fur of a partly swallowed and regurgitated vole is slick and sticks to the anterior part of the body.) Test the hypothesis that $p = .6$ against the alternative $p > .6$.

10. Table 2.1 gives numbers of deaths in US airline accidents from 2000 to 2010. (The entry for 2001 does not include the death toll in the September 11, 2001 plane hijackings.) See the TODAY article by Levin (2011), which cites data from the National Transportation Board.

Suppose you view each trial year as a success if there are no U.S. Airline deaths and a failure otherwise. Discuss the validity of Assumptions A1 and A2. (Mann's test for trend, covered in Comment 8.14, can be used to obtain an approximate P-value for assessing the degree of trend in deaths.)

Table 2.1 Deaths in US Airlines Accidents

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
89	266	0	22	13	22	50	0	0	50	0

Source: A. Levin (2011).

2.2 AN ESTIMATOR FOR THE PROBABILITY OF SUCCESS

Procedure

The estimator of the probability of success p, associated with the statistic B, is

$$\widehat{p} = \frac{B}{n}. \tag{2.18}$$

EXAMPLE 2.3 ***Example 2.2 Continued.***

Consider the triangle test data of Example 2.2. Then $\widehat{p} = B/n = (25/50) = .5$. Thus our point estimate of p, the probability of correctly identifying the odd sample, is $\widehat{p} = .5$.

Comments

14. *Observed Relative Frequency of Success.* The statistic $\widehat{p}$ is simply the observed relative frequency of success in n Bernoulli trials satisfying Assumptions A1-A3. Thus $\widehat{p}$ qualifies as a natural estimator of p, the unknown probability of success in a single Bernoulli trial. That is, we estimate the true unknown probability of success by the observed frequency of success.
15. *Standard Deviation of* $\widehat{p}$. We have shown in Comment 8 that the variance of B is $np(1-p)$, where p is the success probability. It follows that the variance of $\widehat{p}$ is

$$\text{var}(\widehat{p}) = \frac{p(1-p)}{n}. \tag{2.19}$$

The standard deviation of $\widehat{p}$ is

$$sd(\widehat{p}) = \sqrt{\frac{p(1-p)}{n}}. \tag{2.20}$$

Note that $sd(\widehat{p})$ cannot be computed unless we know the value of p, but it can be estimated by substituting $\widehat{p}$ for p in (2.20). This quantity, which we denote as $\widehat{sd}(\widehat{p})$, is a consistent estimator of $sd(\widehat{p})$. The quantity $\widehat{sd}(\widehat{p})$ is also known as the *standard error* of $\widehat{p}$. We have

$$\widehat{sd}(\widehat{p}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}. \tag{2.21}$$

Rather than simply stating the value of $\widehat{p}$ when reporting an observed relative frequency of success, it is important to also report the value of $\widehat{sd}(\widehat{p})$, which (as does $\text{var}(\widehat{p})$) measures the variability of the estimate.

Thus, for the adder data of Problem 9, we could report

$$\widehat{p} = \frac{58}{67} = .87; \quad \widehat{sd}(\widehat{p}) = \sqrt{\frac{\left(\frac{58}{67}\right)\left(\frac{9}{67}\right)}{67}} = .04.$$

Alternatively, we could use a confidence interval for p (see Section 2.3).

16. *Sample Size Determination.* Suppose we want to choose the sample size n so that $\widehat{p}$ is within a distance D of p, with probability $1-\alpha$. That is, we want

$$P_p\left(-D < \widehat{p} - p < D\right) = 1 - \alpha.$$

This is equivalent to

$$P_p\left(\frac{-D}{\sqrt{\dfrac{p(1-p)}{n}}} < \frac{\widehat{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}} < \frac{D}{\sqrt{\dfrac{p(1-p)}{n}}}\right) \doteq 1-\alpha.$$

The variable $(\widehat{p}-p)/\sqrt{p(1-p)/n}$ has an asymptotic $N(0,1)$ distribution and thus we know

$$P\left(-z_{\alpha/2} < \frac{\widehat{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}} < z_{\alpha/2}\right) \doteq 1-\alpha.$$

From the two previous equations, we see that

$$\frac{D}{\sqrt{\dfrac{p(1-p)}{n}}} \doteq z_{\alpha/2}.$$

Solving for n yields

$$n \doteq \frac{(z_{\alpha/2})^2 p(1-p)}{D^2} \tag{2.22}$$

Expression (2.22) requires a guess or estimate for p because p is not known. The function $p(1-p)$ is maximized at $p = \frac{1}{2}$ and decreases to zero as p approaches 0 or 1. Thus we obtain the most conservative sample size by substituting $\frac{1}{2}$ for p in (2.22). This yields

$$n = \frac{(z_{\alpha/2})^2}{4D^2} \tag{2.23}$$

17. *Competing Estimators.* Suppose you observe $B = 0$ in n trials. Depending on the situation, you may have little faith in the estimate $\widehat{p} = 0$. For example, you take a random sample of 10 smokers on a college campus and find no one in the sample smokes. You do not, however, believe that the probability is 0 that a randomly selected student is a smoker. A similar dilemma occurs when $B = n$. One alternative estimator of p is $\tilde{p}$ defined by (2.24) and presented in Section 2.3 on confidence intervals for p. Other alternative estimators use the Bayes estimators presented in Section 2.4.

Properties

1. *Maximum Likelihood Estimator.* The estimator $\widehat{p}$ is the maximum likelihood estimator.
2. *Standard Deviation.* For the standard deviation of $\widehat{p}$ see Comment 15.
3. *Asymptotic Normality.* For asymptotic normality, see Casella and Berger (2002, p 236).

Problems

11. Calculate $\widehat{p}$ for the parolee data of Problem 1 and obtain an estimate of the standard deviation of $\widehat{p}$.

12. Obtain an estimate for the standard deviation of the estimate $\widehat{p}$ calculated in Example 2.1.

13. Suppose $n = 7$. What are the possible values for $\widehat{p}$? When $\alpha = .05$, what are the possible values for $\tilde{p}$ defined by (2.24)?

14. Suppose you are designing a study to estimate a success probability p. Determine the sample size n so that $\widehat{p}$ is within a distance .05 of p with probability .99.

2.3 A CONFIDENCE INTERVAL FOR THE PROBABILITY OF SUCCESS (WILSON)

Procedure

Set

$$\tilde{p} = \widehat{p}\left(\frac{n}{n + z_{\alpha/2}^2}\right) + \frac{1}{2}\left(\frac{z_{\alpha/2}^2}{n + z_{\alpha/2}^2}\right), \tag{2.24}$$

$$p_{\mathrm{L}}^{W}(\alpha) = \tilde{p} - z_{\alpha/2}V^* \tag{2.25}$$

$$p_{\mathrm{U}}^{W}(\alpha) = \tilde{p} + z_{\alpha/2}V^*, \tag{2.26}$$

where

$$V^* = \left\{\frac{1}{n + z_{\alpha/2}^2}\left[\widehat{p}(1 - \widehat{p})\left(\frac{n}{n + z_{\alpha/2}^2}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{z_{\alpha/2}^2}{n + z_{\alpha/2}^2}\right)\right]\right\}^{1/2}. \tag{2.27}$$

With $p_{\mathrm{L}}(\alpha)$ and $p_{\mathrm{U}}(\alpha)$ defined by (2.25) and (2.26),

$$P_p\{p_{\mathrm{L}}^{W}(\alpha) < p < p_{\mathrm{U}}^{W}(\alpha)\} \approx 1 - \alpha. \tag{2.28}$$

The classical large-sample confidence interval (see Comment 19) is centered at $\widehat{p}$. The Wilson confidence interval is centered at $\tilde{p}$ which is a weighted average of $\widehat{p}$ and 1/2 (see Comment 18 and (2.24)).

EXAMPLE 2.4 ***Tempting Fate.***

Risen and Gilovich (2008) conducted a number of studies designed to explore the notion that it is bad luck to tempt fate. In one study, participants were read a scenario in which a student had recently finished applying to graduate school and his top choice was Stanford University. In the scenario, the student's optimistic mother sent him a Stanford T-shirt in the mail. Risen and Gilovich asked a group of 20 participants to consider that the student could either stuff the T-shirt in a drawer while waiting for Stanford's admission decision or could wear the shirt the next day. The question asked of the 20 participants was would a person be more upset receiving a rejection from Stanford after having worn the Stanford shirt than after having stuffed the shirt in a drawer. Eighteen of the 20 participants thought the person would be more upset having worn the shirt. (The person when he wears the shirt "tempts fate" but it is more of a superstitious nature than, for example, tempting fate by walking outside in the middle of a storm replete with lightening. The latter actually increases your chance of a serious accident while wearing the shirt does not affect the chance of admission.) Let p denote the probability that a participant thought the person would be more upset having worn the shirt.

To directly find the Wilson interval for this tempting fate data, we can use the function `binom.confint` from the library `binom`. If we enter `binom.confint(x = 18, n = 20, conf.level = .95, methods = "all")` we obtain the Wilson interval along with a number of other confidence intervals including the Laplace–Wald interval of Comment 19, the Agresti–Coull interval of Comment 20, and the Clopper–Pearson interval of Comment 21. The Wilson 95% interval is (.699, .972).

The null hypothesis of no effect underlying the Risen and Glovich studies is that people are unconcerned about tempting fate, which, in terms of p, is $H_0 : p = 1/2$. With $B = 18$, $n = 20$, we find the one-sided P-value is $P_{1/2}(B \geq 18) = .0002$. Thus there is strong evidence that the participants feel people will avoid tempting fate. The P-value of .0002 can be obtained directly from the `R` function `pbinom(18,20,.5,lower.tail=F)` or equivalently from `1-pbinom(18,20,.5)`.

Comments

18. *The Wilson Confidence Interval.* In general, confidence intervals can be obtained by inverting tests. For a general parameter θ, a two-sided $100(1-\alpha)\%$ confidence interval consists of those θ_0 values for which the two-sided test of $\theta = \theta_0$ does not reject the null hypothesis $\theta = \theta_0$. The confidence interval given by (2.25) and (2.26) is due to Wilson (1927) (see also Agresti and Caffo (2000), Agresti and Coull (1998), Brown, Cai and DasGupta (2001), and Agresti (2013)). It is also called the score interval (see Agresti (2013). The interval is the set of p_0 values for which $|\widehat{p} - p_0|/\{(p_0(1-p_0)/n)\}^{1/2} < z_{\alpha/2}$. The midpoint $\tilde{p}$ of the interval is a weighted average of $\widehat{p}$ and 1/2 with the weights $n/(n + z^2_{\alpha/2})$ and $z^2_{\alpha/2}(n + z^2_{\alpha/2})$, respectively. This midpoint equals the sample proportion obtained if $z^2_{\alpha/2}/2$ pseudo observations are added to the number of successes and $z^2_{\alpha/2}/2$ pseudo observations are added to the number of failures. We can write this midpoint $\tilde{p}$ as
$$\tilde{p} = \frac{B + z^2_{\alpha/2}/2}{n + z^2_{\alpha/2}},$$
which is equivalent to (2.24).

The quantity $(V^*)^2$ (see (2.27)) is a weighted average of the variance of a sample proportion when $p = \widehat{p}$ and the variance of a sample proportion when $p = 1/2$, where $n + z^2_{\alpha/2}$ is used in place of the sample size n.

Brown, Cai, and DasGupta (2001) studied various confidence intervals for p. For $n \leq 40$, they recommended the Wilson interval and an alternative interval due to Jeffreys. For $n > 40$ they found that the Wilson interval, the Jeffreys interval, and the Agresti–Coull interval (see Comment 20) are comparable.

19. *The Laplace–Wald Confidence Interval.* This interval can be obtained by inverting large-sample Wald tests (cf. Agresti, 2013). The approximate $100(1-\alpha)\%$ interval is the set of p_0 values for which $|\widehat{p} - p_0|/\{\widehat{p}(1-\widehat{p})/n\}^{1/2} < z_{\alpha/2}$. The interval is

$$p_{\mathrm{L}}^{\mathcal{L}W}(\alpha) = \widehat{p} - z_{\alpha/2}\{(\widehat{p}(1-\widehat{p}))/n\}^{1/2}, \tag{2.29}$$

$$p_{\mathrm{U}}^{\mathcal{L}W}(\alpha) = \widehat{p} + z_{\alpha/2}\{(\widehat{p}(1-\widehat{p}))/n\}^{1/2}, \tag{2.30}$$

where $\widehat{p} = B/n$. The interval was used by Laplace (1812), and here, we denote it as the $\mathcal{L}W$ interval.

Brown, Cai, and DasGupta (2001) highlight disadvantages of the $\mathcal{L}W$ interval. There exist pairs (p, n), which they call unlucky pairs, for which the coverage probability is much smaller than the nominal coverage probability $1 - \alpha$. The phenomenon of oscillation occurs in n for fixed p and in p for fixed n. They also note that severe changes in the coverage occur in nearby p for fixed n and in nearby n for fixed p. Furthermore, even for large sample sizes, significant changes in coverage probabilities occur in nearby p and in many cases the coverage of the $\mathcal{L}W$ interval is strictly smaller than the nominal level. In particular, Brown, Cai, and DasGupta (2001) show for all $n \leq 45$, the actual coverage of the 99% $\mathcal{L}W$ interval is strictly less than the nominal level for all $0 < p < 1$. See their Examples 1–5.

The $\mathcal{L}W$ interval can be found directly using the `R` function `binom.confint`. For the tempting fate data, if we enter `binom.confint(x=18, n=20,conf. level=.95,methods="all")`, the output for the $\mathcal{L}W$ interval (labeled the "asymptotic" interval in the output) is (.769, 1.031). The parameter p must be between 0 and 1. Thus the upper value 1.031 should be changed to 1.

20. *The Agresti–Coull Confidence Interval.* The Agresti–Coull interval is also centered at $\tilde{p}$. Let $\tilde{q} = 1 - \tilde{p}$. The Agresti–Coull (1998) two-sided confidence interval for p with confidence coefficient approximately $1 - \alpha$ is

$$p_{\mathrm{L}}^{AC}(\alpha) = \tilde{p} - z_{\alpha/2}(\tilde{p}\tilde{q})^{1/2}\tilde{n}^{-1/2} \tag{2.31}$$

$$p_{\mathrm{U}}^{AC}(\alpha) = \tilde{p} + z_{\alpha/2}(\tilde{p}\tilde{q})^{1/2}\tilde{n}^{-1/2} \tag{2.32}$$

With $p_{\mathrm{L}}(\alpha)$ and $p_{\mathrm{U}}(\alpha)$ defined by (2.31) and (2.32),

$$P_p\{p_{\mathrm{L}}^{AC}(\alpha) < p < p_{\mathrm{U}}^{AC}(\alpha)\} \approx 1 - \alpha. \tag{2.33}$$

The Agresti–Coull interval is an alternative to the classical Laplace–Wald interval; one with a better centering point ($\tilde{p}$ instead of $\widehat{p}$). For the case when

$\alpha = .05$, if you substitute "2" for $z_{.025} = 1.96$, it can be thought of as the "add two successes and two failures" interval. Brown, Cai, and DasGupta (2001) recommend the Agresti–Coull interval for practical use when $n \geq 40$, although it is never shorter than the Wilson interval. Its relative simplicity and ease of description make it particularly attractive for an introductory course. See Brown, Cai, and DasGupta (2001) for comparisons of various confidence intervals for the binomial parameter. The Agresti–Coull interval can be found directly from the R function `binom.confint`. For the tempting fate data if we enter `binom.confint(x=18, n=20, conf.level=.95, methods="all")`, we find the approximate 95% interval to be (.687, .984).

21. *The Clopper–Pearson Confidence Interval.* The Clopper–Pearson (1934) confidence interval is obtained by inverting the equal-tail binomial test. That is, if $B = b$ is observed, the Clopper–Pearson interval is defined by $p_{\mathrm{L}}^{CP}(\alpha)$, $p_{\mathrm{U}}^{CP}(\alpha)$, where $p_{\mathrm{L}}^{CP}(\alpha)$ and $p_{\mathrm{U}}^{CP}(\alpha)$ are, respectively, the solutions in p to the equations
$$P_p(B \geq b) = \alpha/2, \quad P_p(B \leq b) = \alpha/2.$$

The endpoints of the $100(1-\alpha)\%$ confidence interval are defined by the following equations:

$$p_{\mathrm{L}}^{CP}(\alpha) = \frac{B}{B + (n - B + 1)f_{\alpha/2,2(n-B+1),2B}} \tag{2.34}$$

$$p_{\mathrm{U}}^{CP}(\alpha) = 1 - p_{\mathrm{L}}^{\alpha}(n, n - B), \tag{2.35}$$

where B is the number of successes in the n Bernoulli trials and f_{γ,n_1,n_2} is the upper γth percentile for the F distribution with n_1 degrees of freedom in the numerator and n_2 degrees of freedom in the denominator.

The Clopper–Pearson interval is conservative,

$$P_p\{p_{\mathrm{L}}^{CP}(\alpha) < p < p_{\mathrm{U}}^{CP}(\alpha)\} \geq 1 - \alpha. \tag{2.36}$$

The conservativeness can be extreme in that for any fixed p, the true coverage probability can be much larger than $1 - \alpha$ unless n is quite large.

The Clopper–Pearson interval can be found directly from the R function `binom.confint`. For the tempting fate data we apply `binom.confint( x=18,n=20,conf.level=.95,methods="all")` and find the CP interval is (.683, .987). In the output the CP interval is labeled as "exact".

22. *Equivariance*. Binomial confidence interval procedures that satisfy (2.35) are said to be equivariant (Casella, 1986). The motivation for the term *equivariance* is that the binomial distribution is invariant under the transformations $B \rightarrow n - B$ and $p \rightarrow 1 - p$. See Casella (1986) for further details. The Clopper–Pearson intervals are equivariant but they are not the only ones that enjoy the equivariance property. Casella (1986) gives a method for refining equivariant binomial confidence intervals to obtain new intervals with uniformly shorter lengths for the same confidence coefficient.

23. *The Multinomial Distribution.* The binomial distribution given by (2.15) can be extended to situations where an experiment has k ($k \geq 2$) possible outcomes or categories, say $A_1, A_2, \ldots, A_k$, which are mutually exclusive and exhaustive.

We let $P(A_i) = p_i, i = 1, \ldots, k$, where $\sum_{i=1}^{k} p_i = 1$. Furthermore, let X_i be the number of times A_i occurs in the n trials. The k variables $X_1, X_2, \ldots, X_k$ are said to have the multinomial distribution with parameter $n, \mathbf{p}$, where $\mathbf{p} = (p_1, p_2, \ldots, p_k)$. The distribution is given by

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \binom{n}{x_1, x_2, \ldots, x_k} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}, \tag{2.37}$$

where

$$\binom{n}{x_1, x_2, \ldots, x_k} = \frac{n!}{x_1!x_2! \ldots x_k!}. \tag{2.38}$$

The quantity $\binom{n}{x_1, x_2, \ldots, x_k}$ is known as the *multinomial coefficient*. It is equal to the number of distinguishable arrangements of x_1 A_1's, x_2 A_2's, …, x_k A_k's. The mean and variance of X_i are

$$E(X_i) = np_i, \ \text{var}(X_i) = np_i(1 - p_i), \ i = 1, \ldots k.$$

The covariance between X_i and X_j is

$$\text{cov}(X_i, X_j) = -np_i p_j.$$

24. *Some Examples Where the Multinomial Arises.*

    Example A: In Paradise Paved, Florida, 44% of the voters are Democrats, 42% are Republican, and 14% are in some other category (Independent, Tea Party, etc.). Suppose a random sample of 20 voters are polled yielding X_1 Democrats, X_2 Republicans, and X_3 in the other category. Then, (X_1, X_2, X_3) has a multinomial distribution with parameters $n = 20$ and $\mathbf{p} = (.44, .42, .14)$.

    Example B: Cohen and Bloom (2010) report on data from the National Health Interview Survey, 2008. The distribution of health insurance status for male adults aged 20–29 years was 57.5% had private insurance, 5.6% were on Medicaid, 1.6% had some other form of insurance, and 35.3% were uninsured. Suppose a random sample of 40 was obtained from this population yielding X_1 on private insurance, X_2 on Medicaid, X_3 on some other coverage, and X_4 uninsured. Then, (X_1, X_2, X_3, X_4) has a multinomial distribution with parameters $n = 40$ and $\mathbf{p} = (.575, .056, .016, .353)$.

25. *Estimation for the Multinomial Distribution.* For the multinomial distribution, if we observe the frequencies $X_1, X_2, \ldots, X_k$, where X_i is the number of times the event A_i occurs in n experiments, the standard frequentist estimators of $p_1, p_2, \ldots, p_k$ are the sample proportions

$$\widehat{p_i} = X_i/n, \quad i = 1, \ldots, k. \tag{2.39}$$

With $k \geq 2$, asymptotic $100(1 - \alpha)\%$ simultaneous confidence intervals for the $M = k(k-1)/2$ pairwise differences $\{p_i - p_j\}, 1 \leq i < j \leq k$, can be obtained using Bonferroni's inequality. They are

$$p_{\text{L}}^{(i,j)} = \widehat{p_i} - \widehat{p_j} - z_{\alpha/2M}\{[\widehat{p_i} + \widehat{p_j} - (\widehat{p_i} - \widehat{p_j})^2]/n\}^{1/2}, \tag{2.40}$$

$$p_{\text{U}}^{(i,j)} = \widehat{p_i} - \widehat{p_j} + z_{\alpha/2M}\{[\widehat{p_i} + \widehat{p_j} - (\widehat{p_i} - \widehat{p_j})^2]/n\}^{1/2}, \tag{2.41}$$

where $\widehat{p_i} = X_i/n,\ i = 1, \ldots, k$. The M intervals given by (2.40) and (2.41) satisfy, for large n,

$$P\{p_{\mathrm{L}}^{(i,j)} < p_i - p_j < p_{\mathrm{U}}^{(i,j)}, 1 \le i < j \le k\} \approx 1 - \alpha. \tag{2.42}$$

That is, the probability is approximately $1 - \alpha$ that the M intervals simultaneously contain the M differences $\{p_i - p_j, i < j\}$.

Asymptotic $100(1-\alpha)\%$ simultaneous confidence intervals for $p_i, i = 1, \ldots, k$, based on Bonferroni's inequality, are obtained by solving

$$(\widehat{p_i} - p_i)^2 = (z_{\alpha/2k})^2 p_i(1 - p_i)/n \tag{2.43}$$

for lower and upper limits $p_{\mathrm{L}}^{(i)}, p_{\mathrm{U}}^{(i)}$ (see Goodman (1965), Miller (1981a), Fitzpatrick and Scott (1987) and Agresti (2013)).

26. *Pearson's Chi-Squared Goodness-of-Fit Test for Specified Multinomial Probabilities.* Pearson's (1900) chi-squared statistic can be used to test, on the basis of n experiments with frequencies $X_1, X_2, \ldots, X_k$ corresponding to the k categories, the hypothesis that the multinomial probabilities $p_1, p_2, \ldots, p_k$ are equal to specified or known values $p_1^0, p_2^0, \ldots, p_k^0$. Pearson's chi-squared statistic is

$$\chi^2 = \sum_{i=1}^{k} \left\{ \frac{(X_i - np_i^0)^2}{np_i^0} \right\} \tag{2.44}$$

Note deviations of the X_i's from their hypothesis expected values (the np_i^0,s) are magnified by the $(X_i - np_i^0)^2$ terms leading to large values of χ^2. That is, significantly large values of χ^2 indicate a deviation from the hypothesis

$$H_0: \quad p_1 = p_1^0,\ p_2 = p_2^0,\ \ldots,\ p_k = p_k^0 \tag{2.45}$$

in favor of the alternative

$$H_1: \quad p_i \ne p_i^0 \text{ for at least on value of } i. \tag{2.46}$$

If one computes $\chi^2 = \chi^2_{obs}$ (the observed value), one can find the corresponding P-value, the probability under the null hypothesis that $\chi^2 \ge \chi^2_{obs}$, by summing the probabilities given by (2.37) over all possible multinomial outcomes yielding $\chi^2 \ge \chi^2_{obs}$. It is more convenient, however, to use a large-sample approximation.

Pearson showed that when H_0 is true, the distribution of χ^2 as $n \to \infty$, is that of a chi-squared distribution with $k - 1$ degrees of freedom. Thus, an approximate α-level test is

$$\text{Reject } H_0 \text{ if } \chi^2 \ge \chi^2_{\alpha,k-1}; \text{ otherwise do not reject.} \tag{2.47}$$

The P-value is found by referring the observed value of χ^2 to the χ^2_{k-1} distribution.

The χ^2 approximation is good when each of the np_i^0's is not too small. A general foot rule is it's good if $np_i^0 \ge 5$ for each value of i.

Table 2.2 Outcomes of Pea Plant Experiments

Trait	Dominant		Recessive		χ^2	P-value	Expected Ratio
Seed shape	Round	5474	Angular	1850	.2629	.6081	3:1
Cotyledon color	Yellow	6022	Green	2001	.015	.9025	3:1
Seed coat color	Colored	705	White	224	.3907	.5319	3:1
Pod shape	Inflated	882	Constricted	299	.0635	.801	3:1
Pod color	Green	428	Yellow	152	.4506	.5021	3:1
Flower position	Axial	651	Terminal	207	.3497	.5543	3:1
Stem length	Long	787	Short	277	.6065	.4361	3:1

Source: D.J. Fairbanks and B. Rytting (2001).

27. *Checking for Data Fudging: The Fit May be Too Good.* The chi-squared statistic rejects the goodness-of-fit null hypothesis (2.45) if χ^2 is too large. Small values in the lower tail of the null distribution of χ^2, however, can give an indication that the fit is too good and that perhaps the data have been "cooked" so that they would appear to support the hypothesized model values.

A classic example of the use of Pearson's χ^2 involves Gregor Mendel's famous genetics experiments on pea plants. Mendel, a European monk whom many biologists regard as the father of genetics, cross-pollinated purebred plants with specific traits and observed and recorded the results over many generations. Table 2.2, based on data in Fairbanks and Rytting (2001), gives the f_2 generation (the second offspring of cross-pollinated purebred plants) pertaining to seven pea characteristics: (1) seed shape (round or angular), (2) cotyledon (part of the embryo within the seed) color (yellow or green), (3) seed coat color (colored or white), (4) pod shape (inflated or constricted), (5) pod color (green or yellow), (6) flower position (axial or terminal), (7) stem length (long or short).

All of the χ^2 statistics in Table 2.2 are based on one degree of freedom (*df*). The sum of seven independent χ^2 statistics with one *df* follows a χ^2 distribution with $df = 7$. (More generally, the sum $\sum_i \chi^2$ of independent χ^2 statistics, with the i^{th} having $df = m_i$, follows a χ^2 distribution with $df = \sum m_i$.) Summing the seven χ^2 values in column 6 of Table 2.1 yields $\sum_{i=1}^{7} \chi^2 = 2.1389$ and $P(\sum_{i=1}^{7} \chi^2 \leq 2.1389) = .0482$. Thus the value 2.1389 falls in the lower tail of the distribution and arouses suspicion that the fit is too good.

Mendel did many more experiments than those represented in Table 2.2. Agresti (2013) and Fisher (1936) summarized Mendel's experiments and obtained a χ^2 value of 42 based on $df = 84$. We find $P(\chi^2_{84} \leq 42) = .000035$. This chi-square value is extremely small and it is smaller than would be expected when the model fits. Fisher suspected that an overzealous assistant might have biased the data. Other possibilities include the "several left-in-the drawer" theory in which Mendel may have only reported the "best" results and omitted the results of other experiments. Nevertheless, over time the works of Mendel and many others have led to acceptance of Mendel's genetic theories. Pires and Branco investigate a model that may alleviate the controversy. See their paper and Box (1978) for more details about the history of the Mendel-Fisher difficulty.

The χ^2 values can be readily obtained from `R` functions. For example, the $\chi^2 = .2629$ value given in the first row of Table 2.2 is obtained from `chisq.test(c(5474,1850), p=c(.75,.25))` yielding the output $\chi^2 = .2629$, $df = 1$, P-value=.6081. The lower-tail probability $P(\chi^2_{84} \leq 42) =$

.000035 corresponding to the value $\chi^2_{84} = 42$ is obtained by the R function `pchisq(42,84)`.

28. *Testing Equal Probabilities.* In the case when the multinomial probabilities are specified to be equal, that is H_0 is taken to be $p_1 = 1/k, \ldots, p_k = 1/k$, the chi-squared statistic reduces to

$$\chi^2 = \left\{(k/n)\sum_{i=1}^{k} X_i^2\right\} - n. \tag{2.48}$$

29. *The Case k = 2.* When $k = 2$, the multinomial setting reduces to the binomial setting and Pearson's χ^2 test is a test of $p = p_0$. In this case, the approximate test defined by (2.47) is equivalent to the approximate two-sided test of $p = p_0$ versus the alternative $p \neq p_0$ given by (2.12).

Properties

1. *Distribution-Freeness.* For Bernoulli trails satisfying Assumptions A1–A3, (2.36) holds. Thus, $(p_{\mathrm{L}}^{CP}(\alpha), p_{\mathrm{U}}^{CP}(\alpha))$ is a confidence interval for p with confidence coefficient at least $1 - \alpha$.

Problems

15. For the parolee data of Problem 1, obtain the Wilson, Laplace–Wald, Agresti–Coull, and Clopper–Pearson confidence intervals for p, each with an approximate confidence coefficient of .96. Compare the four intervals.

16. Shlafer and Karow (1971) considered some of the problems involved with cardiac preservation. In particular, they were interested in the morphological and physiological injury occurring in hearts that had been frozen to various temperatures without the benefit of a cryoprotectant. Hearts from adult rats were perfused with a balanced salt solution *in vitro* for 20 min, and during this time, contractions were noted. After disconnection from the perfusion apparatus, each heart, surrounded by a plastic shield, was inserted into a metal canister and chilled by an acetone bath (maintained at −20°C by addition of dry ice) until the lowest desired temperature was attained. The individual hearts were then thawed (in 1 min or less) by removing the metal canister and running 35°C tap water over the plastic shields, being careful to prevent water from flowing directly over the hearts. After thawing, the hearts were again perfused with the balanced salt solution. Hearts spontaneously resuming coordinated atrioventricular contractions within 20 min of thawing were considered to be "survivors" of the freeze–thawing process.

The authors conducted experiments where the lowest attained temperatures were −10, −12, −17, and −20°C. We focus here on the data for the −12°C investigation, in which the authors found that of six hearts frozen to −12°C, three were survivors. If we let success denote survival, then p represents the probability that a rat heart frozen to −12°C will spontaneously resume coordinated atrioventricular contractions within 20 min after thawing and perfusion with a balanced salt solution. Obtain the Wilson, Laplace–Wald, Agresti–Coull, and Clopper–Pearson confidence intervals for p, each with an approximate confidence coefficient of .90. Compare the intervals.

17. Many materials that are satisfactory for use in air will ignite and burn in pure oxygen when subjected to mechanical impact. This problem is of vital concern to the aerospace industry, which uses enormous amounts of both liquid and gaseous oxygen. In particular, there is a need for guidelines to aid the designer in selecting materials to be employed in pressurized oxygen systems. In order to provide an appropriate method for determining gaseous oxygen-material compatibility, the Kennedy Space Center developed a gaseous oxygen impact test procedure. Jamison (1971) reported on the use of this testing scheme to analyze the gaseous oxygen-material compatibility for 33 Apollo spacecraft test materials. One such material tested, silicone elastomer 342, failed the gaseous oxygen impact test (i.e., the material ignited) in 4 out of 20 trials. Let p denote the probability of ignition for silicone elastomer 342 when subjected to the conditions employed in the gaseous oxygen impact test. Obtain the Wilson, Laplace–Wald, Agresti–Coull, and Clopper–Pearson confidence intervals for p, each with an approximate confidence coefficient .95. Compare the intervals.

18. Ehlers (1995) performed a 1-year follow-up study of panic disorder. As partial motivation for the study, the author offered the following quote from Wolfe and Maser (1994, 241): "Little is known about the long-term course of disorder. The limited findings to date suggest that in most cases it is a chronic disorder that waxes and wanes in severity. However, some people may have a limited period of dysfunction that never recurs, while others tend to have a more severe and complicated course." In this study, diagnoses were made by trained interviewers (either the author, Ehlers, or trained graduate students) according to the criteria of the revised third edition of the "Diagnostic and Statistical Manual of Mental Disorders" (DSM-III-R; American Psychiatric Association, 1987). One year after initial assessments, participants were mailed a questionnaire for the purpose of assessing their current symptoms. In this problem, we discuss one small portion of the data that were obtained. Out of 46 people who were initially diagnosed as "infrequent panickers," 23 experienced panic attacks during follow-up. Obtain the approximate 95% Wilson, Laplace–Wald, Agresti–Coull, and Clopper–Pearson confidence intervals for p, the probability that an infrequent panicker will experience panic attacks during a 1-year follow-up period. Compare the intervals.

19. For the triangle bitterness tests data of Example 2.2, obtain the Wilson, Laplace–Wald, Agresti–Coull, and Clopper–Pearson confidence intervals for p, each at an approximate confidence coefficient of .90. Compare the intervals.

20. Consider the study in Problem 17 and describe the inherent sources for error when one uses a mailed questionnaire.

21. Consider the study in Problem 18 and discuss the possibility of unintentional bias entering the study, because some of the diagnoses were made by the author of the study.

22. A table of random numbers contains randomly generated digits that have been generated so that the following two properties are satisfied.

(I) *Equal Likelihood Property.* Focus on any particular spot in the table, such as the second digit in the fourth row. The probability it will be a 0 equals the probability it will be a 1 ... equals the probability it will be a 9. More succinctly, $P(0) = P(1) = \ldots = P(9) = 1/10$.

(II) *Mutual Independence Property.* Focus on any particular spot in the table, such as the third digit in the second row. Knowing some or even all of the digits in the table, except for the one you are considering, does not change the probability it will be a 0, or a 1, ..., or a 9. This probability remains a 1/10.

For the 400 randomly generated digits in Table 2.3, test the equal likelihood property.

Table 2.3 Four Hundred Random Digits

24253	39427	80642	36718	92164	77732	69754	01291	53704	33054
34302	60309	27186	22418	59962	13934	67591	17476	21559	73437
76809	84341	74012	50947	83214	19967	44219	75929	13182	34858
85183	35958	04301	49628	91493	66103	65699	04241	82441	38112
27541	79187	99777	22894	83283	56218	86183	74497	21070	78935
74188	09083	54938	79920	27158	24864	31116	33173	43032	52000
13270	57457	30968	65978	67679	91216	47969	39204	46030	93954
89150	53922	40537	23169	46948	05519	72171	85417	31580	98102

23. Devore (1991) gives an example from the paper “Linkage Studies of the Tomato” (*Tran. Royal Canadian Inst.*, 1931, pp 1-19) on phenotypes from a dihybrid cross of tall, cut-leaf tomatoes with dwarf, potato-leaf tomatoes. Dihybrid crosses arise as follows. If two different characteristics of an organism are each controlled by a single gene, and a pure strain having genotype AABB is crossed with a pure strain having genotype aabb (A,B are dominant alleles; a,b are recessive), the resulting genotype is AaBb. A dihybrid cross occurs when the first-generation organisms are crossed among themselves. The data for a dihybrid cross of tall, cut-leaf tomatoes with dwarf, potato-leaf tomatoes, based on a sample of n=1611, are as follows: Tall, cut-leaf (926); tall, potato-leaf (288); dwarf, cut-leaf (293); dwarf, potato-leaf (104) (tall is dominant for size, cut-leaf is dominant for leaf type). Mendelian inheritance theory predicts that there are four categories of probabilities occurring, 9/16, 3/16, 3/16, 1/16. Test if the data support Mendelian theory.

24. Consider the National Health Interview Survey of Example B of Comment 24. Suppose a random sample of 100 from that population yielded the results: 60 males on private insurance, 5 on Medicaid, 4 on some other form of insurance, and 31 uninsured. Is such a sample consistent with the specified population probabilities?

25. For the data of Example B of Comment 24, find approximate 90% confidence intervals for the six pairwise differences $p_1 - p_2$, $p_1 - p_3$, $p_1 - p_4$, $p_2 - p_3$, $p_2 - p_4$, $p_3 - p_4$.

26. Establish the expression for χ^2 given by (2.48), Comment 28.

27. Show the equivalence of the two tests described in Comment 29.

2.4 BAYES ESTIMATORS FOR THE PROBABILITY OF SUCCESS

The estimator $\widehat{p}$ of Section 2.2 is a frequentist estimator of p; it does not utilize prior information. In this section, we consider Bayes estimators of p that make use of prior information. The estimators are based on the Beta class $Beta(r, s)$ $(r > 0,\ s > 0)$ of prior distributions for p. In the Bayesian approach p is considered a random variable. Guidance for the choice of the prior distribution is presented in Comments 30–33.

Procedure

For squared-error loss, the Bayes estimator of p when using the prior distribution $Beta(r, s)$, is the mean of the posterior distribution of p, given $B = b$. This mean can be denoted as

$$E(p|B = b) = \frac{b + r}{r + s + n} \tag{2.49}$$

Note that the Bayes estimator can be rewritten as

$$E(p|B = b) = \left(\frac{n}{n + r + s}\right)\frac{b}{n} + \left(\frac{r + s}{n + r + s}\right)\frac{r}{r + s}. \tag{2.50}$$

In the form (2.50), we can see that the Bayes estimator is a weighted average of the observed proportion of successes b/n and the prior guess at p, namely $E(p) = r/(r + s)$. The weights are $n/(n + r + s)$ and $(r + s)/(n + r + s)$. Note that as n gets large, the second term in (2.50) tends to 0 and the first term tends to b/n. This is a reflection of the fact that as the sample size gets large, the observed data dominate the prior information.

EXAMPLE 2.5 ***Percentage of Smokers.***

In May 2010, E. Chicken polled his two statistics courses to investigate the percentage of students who smoke cigarettes. The classes were STA 4321 Introduction to Mathematical Statistics and STA 4502-5507 Applied Nonparametric Statistics. The 4321 class had 27 students comprising 5 females and 22 males. The 5507 class had 17 students comprising 5 females and 12 males. In STA 4321, there was one smoker, a female. In STA 5507, there were five smokers, three females and two males. To have a large sample size, the two classes were combined. Thus, out of a total of 44 students, 6 were smokers. The classical frequentist estimator for the true proportion p of smokers in the college population is $\widehat{p} = 6/44 = .136$ or 13.6%.

A Bayesian approach can effectively be employed because there are many studies concerning smoking rates. For example, a National Health Interview Survey (NHIS) (2008) estimated the smoking rates to be 23.1% for men and 18.3% for women. We will illustrate Bayesian approaches using a noninformative prior (see Comment 31) and an informative prior (see Comments 32 and 33). Using the noninformative Bayes–Laplace prior $Beta(1, 1)$ (see Comment 31), we find from (2.49) with $n = 44$, $b = 6$, $r = 1$, $s = 1$,

$$E(p|B = 6) = \frac{6+1}{2+44} = \frac{7}{46} = .152$$

or 15.2%.

We can also use an informative prior such as the one mentioned in Comment 32. From the NHIS (2008) results, it is reasonable to take $p^* = .20$ as a good guess at the percentage of smokers. Setting

$$p^* = .20 = r/r + s$$

as in (2.56), and taking $\sigma^* = .05$, we have from (2.57),

$$(.05)^2 = \frac{rs}{(r+s)^2(r+s+1)}$$

Solving the previous two equations for r and s yields

$$r = 12.6, \quad s = 50.4,$$

and from (2.49)

$$E(p|B = 6) = \frac{6 + 12.6}{12.6 + 50.4 + 44} = .174$$

or 17.4%.

Note that in this example, both Bayesian estimators, one based on a noninformative prior and the other on an informative prior, are closer to what might have been expected, considering the results of the NHIS survey. Yet times change, you are not reading this in 2008, and there are strong efforts in the United States to reduce the incidence of smoking. Furthermore, statisticians have played a prominent role in discovering and assessing the increased risks of various health problems (e.g., lung cancer, heart disease, and emphysema) associated with smoking, so it is not surprising that in statistics classes in a college population the incidence may be less than that in broader populations.

Comments

30. *Bayes Estimators.* The Bayesian approach incorporates prior information into the estimation procedure. One starts with a prior density for p, which is viewed as a random variable. After observing the data $B = b$, the prior and the data are used to compute the posterior density of p. The conjugate prior for p is the beta distribution, $Beta(r, s)$. A random variable Y has a beta distribution with parameters r, s ($r > 0$, $s > 0$) if Y has the density function

$$\begin{aligned} f(y) &= \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} y^{r-1}(1-y)^{s-1}, \quad && 0 < y < 1 \\ &= 0 && \text{otherwise.} \end{aligned} \tag{2.51}$$

The mean of the distribution is

$$E(y) = \frac{r}{r+s} \tag{2.52}$$

and the variance is

$$\operatorname{var}(Y) = \frac{rs}{(r+s)^2(r+s+1)}. \tag{2.53}$$

Figure 2.1 shows various beta densities.

The posterior distribution of p, given $B = b$, is readily shown to be $Beta(b + r, n - b + s)$, that is, a beta distribution with parameters $b + r$ and $n - b + s$. For squared-error loss, the Bayes estimator of p is the mean of this posterior distribution as given in (2.49), namely,

$$E(p|B = b) = \frac{b+r}{b+r+n-b+s} = \frac{b+r}{r+s+n} \tag{2.54}$$

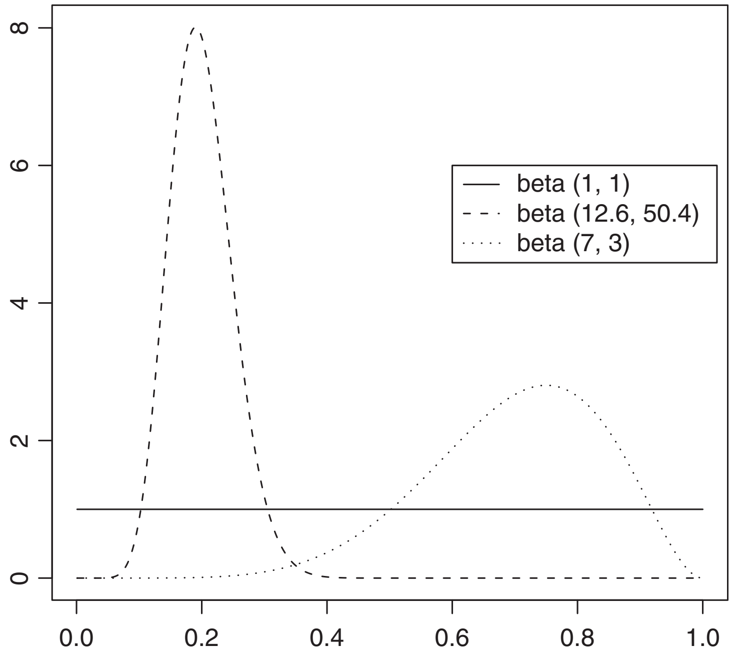

Figure 2.1 Three beta densities.

As we have noted in the Procedure subsection, the Bayes estimator can be rewritten as

$$E(p|B=b) = \left(\frac{n}{n+r+s}\right)\frac{b}{n} + \left(\frac{r+s}{n+r+s}\right)\frac{r}{r+s} \tag{2.55}$$

In the form (2.55), we see that the Bayes estimator is a weighted average of the observed proportion of successes b/n and the prior guess $E(p) = r/(r+s)$.

31. *Choice of Prior When Minimal Information is Available.* When there is little information about the parameter of interest, Bayesians, who still want to employ the Bayesian structure and machinery, favor using a noninformative prior. A noninformative prior, roughly speaking, contains little information about the parameter of interest, or favors no possible value of that parameter over other possible values (see Berger (1985, p. 82)). In the case we are discussing in this chapter, namely, where the parameter of interest is p, the probability of success, Berger (1985, p. 89) considers the following priors to be reasonable noninformative priors (i) $f_1(p) = 1$, the uniform prior corresponding to $Beta(1, 1)$, is often referred to as the Bayes–Laplace prior. (ii) $f_2(p) = p^{-1}(1-p)^{-1}$ is known as the *Haldane prior*. It can be roughly viewed as $Beta(0, 0)$ and yields as the Bayes estimator $\widehat{p} = B/n$, (iii) $f_3(p)$ proportional to $[p(1-p)]^{-1/2}$ is the $Beta(1/2, 1/2)$ prior due to Jeffreys (1961), and (iv) $f_4(p)$ proportional to $p^p(1-p)^{1-p}$, a prior not in the Beta family and one that arises from an approach due to Zellner (1977). Berger points out that f_1 is a proper density, as are f_3 and f_4 suitably normalized, whereas f_2 is improper.

32. *Choice of Prior when Prior Information is Available.* The Beta family is often used and is quite flexible for characterizing prior information. If one has a good guess, p^* say, from a prior experiment perhaps, about the location of p, it is reasonable to set p^* equal to the mean of the Beta distribution, namely,

$$p^* = \frac{r}{r+s} \tag{2.56}$$

If a reasonable choice for the standard deviation, say σ^*, is also available then set

$$(\sigma^*)^2 = \frac{rs}{(r+s)^2(r+s+1)}, \tag{2.57}$$

the Beta variance, and solve (2.56) and (2.57) for r and s, to obtain r^*, s^* say, then use $Beta(r^*, s^*)$. This is illustrated in Example 2.5.

33. *Choice of Prior in the Case of Zero Events.* Tuyl, Gerlach, and Mengerson (2008) consider the case of zero observed events, that is, $B = 0$ (or equivalently $B = n$). Their paper considers the four noninformative priors discussed by Berger (1985, p. 89) and mentioned in Comment 31. They recommend the Bayes–Laplace prior as a consensus noninformative prior. They also note that the use of a Beta prior with small Beta parameters, namely, r, s, < 1 should be avoided, both for noninformative and informative priors. One of their examples suggests that when p is known to be very small, an informative prior from $Beta(1, s)$ with $s > 1$ seems appropriate but a $Beta(r, s)$ with $r < 1$ can be too informative.

34. *The Prior Should Have Support on the Entire Parameter Space.* A prior distribution should not be so restrictive that it prevents the data from telling the true story. For example, in our problem of estimating p, if one chooses a prior that puts all of its probability on a proper subset of $[0, 1]$ and the true value of p is outside of that subset, the resulting Bayes estimator will not converge to the true value as the sample size grows. Thus, for example, if you put a uniform prior on the interval $[0, 1/2]$ and the true value of p is greater than $1/2$, the Bayes procedure will not converge to the true value.

    We want, for large samples, the data to have most of the influence. If you have a spread out prior that covers the parameter space, then for large samples the data will dominate and your posterior distribution will not be too influenced by your prior distribution.

35. *Bayesian Updating.* Suppose your prior distribution is $Beta(r, s)$, and suppose that you observe b successes in n binomial trials. Then, as noted in Comment 30, your posterior distribution is $Beta(b + r, n - b + s)$ and the Bayes estimator is $(b + r)/(r + s + n)$. Now suppose you obtain a second sample of size m and, in that second sample, you have c successes and $m - c$ failures. From your first sample, your new prior is $Beta(b + r, n - b + s)$, and after observing c successes, your new posterior is $Beta(b + r + c, s + m + n - b - c)$ and your Bayes estimator is $(b + r + c)/(b + r + c + s + m + n - b - c)$ or $(b + r + c)/(r + s + m + n)$. This agrees with what you would obtain by pooling the two samples of n and m with successes b and c, respectively. For if you start with a $Beta(r, s)$ prior, then obtain $b + c$ successes in a combined sample of $n + m$, the posterior based on the pooled sample is $Beta(r + b + c, s + m + n - b - c)$ with corresponding Bayes estimator $(r + b + c)/(r + b + c + s + m + n - b - c)$ or $(r + b + c)/(r + s + m + n)$.

36. *Bayes Estimation for the Multinomial Distribution.* For the Bayesian approach, the conjugate density is the Dirichlet distribution with parameters $\beta_1, \ldots, \beta_k$; letting $\mathbf{p} = (p_1, \ldots, p_k)$, the density is

$$f(\mathbf{p}) = \frac{\Gamma(\beta_0)}{\prod_{i=1}^{k} \Gamma(\beta_i)} \cdot \prod_{i=1}^{k} p_i^{\beta_i - 1}, \quad 0 < p_i < 1, \sum_{i=1}^{k} p_i = 1,$$

where $\beta_i > 0$ and $\beta_0 = \sum_{i=1}^{k} \beta_i$. We continue with this density in Chapter 16, where the Dirichlet distribution is generalized to the Dirichlet process.

The mean and variance of the Dirichlet are

$$E(p_i) = \frac{\beta_i}{\beta_0}, \quad \text{var}(p_i) = \frac{\beta_i(\beta_0 - \beta_i)}{\beta_0^2(\beta_0 + 1)}. \tag{2.58}$$

The posterior density is Dirichlet with parameters $X_i + \beta_i$, $i = 1, \ldots, k$ so that the posterior mean is

$$E(p_i | X_1, \ldots, X_k) = \frac{X_i + \beta_i}{n + \beta_0}. \tag{2.59}$$

The Bayes estimator, for the loss function $L(p, a) = \sum_{i=1}^{k} (p_i - a_i)^2$, is the posterior mean given by (2.59). Note that the Bayes estimator given by (2.59) can be rewritten as

$$E(p_i|X_1, \ldots, X_k) = \left(\frac{n}{n+\beta_0}\right)\frac{X_i}{n} + \left(\frac{\beta_0}{n+\beta_0}\right)\frac{\beta_i}{\beta_0}, \tag{2.60}$$

which is a weighted average of the observed sample proportion X_i/n and the prior guess at p_i, β_i/β_0. Note that as n gets large, the Bayes estimator approaches the frequentist estimator given by (2.39).

Properties

1. *Bayes Optimality of* $E(p|B = b)$. For the $Beta(r, s)$ prior and squared-error loss, $E(p|B = b)$ minimizes the Bayes risk.
2. *Bayes Optimality of* $E(p_i|X_1, \ldots, X_k)$. For the Dirichlet distribution prior and sum of squared-error loss, $E(p_i|X_1, \ldots, X_k)$ minimizes the Bayes risk.

Problems

28. Consider the canopy gap closure data of Example 2.1. Determine a Bayes estimate for p. Explain how you obtained your prior distribution.

29. Consider the cardiac preservation data of Problem 16. Determine a Bayes estimate for p. Explain how you obtained your prior distribution.

30. Consider the silicone elastomer data of Problem 17. Determine a Bayes estimate for p. Explain how you obtained your prior distribution.

31. Consider the panic attack data of Problem 18. Determine a Bayes estimate for p. Explain how you obtained your prior distribution.

32. Consider the tempting fate data of Example 2.4. Determine a Bayes estimate for p. Explain how you obtained your prior distribution.

33. Consider the data on smokers in Example 2.5. Suppose the data from STA 5507 is not available so that you only have the data from the STA 4321 class. Determine a Bayes estimate for p. Explain how you obtained your prior distribution.

34. Consider the tomato data of Problem 23. Determine a Bayes estimate for $\mathbf{p} = (p_1, p_2, p_3, p_4)$. Explain how you obtained your prior distribution.

35. Consider the insurance data of Problem 24, Determine a Bayes estimate for $\mathbf{p} = (p_1, p_2, p_3, p_4)$. Explain how you obtained your prior distribution.

36. Describe three situations in which the costs associated with obtaining sample observations are exorbitant and thus in those situations the Bayesian approach is particularly appealing.