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).