Chapter 9

Categorical Data: One-Sample Distributions

Objectives

In this chapter we study categorical data. We will

- explore sampling distributions for estimators that describe dichotomous populations.
- demonstrate how to make and interpret confidence intervals for proportions.
- provide a method for finding an optimal sample size for estimating a proportion.
- show how and when to conduct a chi-square goodness-of-fit test.

9.1 Dichotomous Observations

In Chapter 5 we worked with problems involving numeric variables and examined the sampling distribution of the sample mean. In Chapter 6 we used the sampling distribution to explain how the sample mean tends to vary from the population mean and we constructed confidence intervals for the population mean. We begin this chapter by proceeding in a similar manner by first considering a simple dichotomous categorical variable (i.e., a categorical variable that has only two possible values) and the sampling distribution of the sample proportion. In Section 9.2 we will use the sampling distribution of the sample proportion to construct a confidence interval for a population proportion.

The Wilson-Adjusted Sample Proportion, $\widetilde{P}$

When sampling from a large dichotomous population, a natural estimate of the population proportion, p, is the sample proportion, $\hat{p} = y/n$, where y is the number of observations in the sample with the attribute of interest and n is the sample size.

Example 9.1.1

Contaminated Soda At any given time, soft-drink dispensers may harbor bacteria such as *Chryseobacterium meningosepticum* that can cause illness.[1] To estimate the proportion of contaminated soft-drink dispensers in a community in Virginia, researchers randomly sampled 30 dispensers and found 5 to be contaminated with *Chryseobacterium meningosepticum.* Thus the sample proportion of contaminated dispensers is

$$\hat{p} = \frac{5}{30} = 0.167$$

■

The estimate, $\hat{p} = 0.167$, given in Example 9.1.1 is a good estimate of the population proportion of contaminated soda dispensers, but it is not the only possible estimate. The Wilson-adjusted sample proportion, $\widetilde{p}$, is another estimate of the population proportion and is given by the formula in the following box.

Wilson-Adjusted Sample Proportion, $\tilde{p}$

$$\tilde{p} = \frac{y + 2}{n + 4}$$

Example 9.1.2

Contaminated Soda The Wilson-adjusted sample proportion of contaminated dispensers is

$$\tilde{p} = \frac{5 + 2}{30 + 4} = 0.206\text{*}$$ ■

As the previous example illustrates, $\tilde{P}$ is equivalent to computing the ordinary sample proportion $\hat{P}$ on an augmented sample: one that includes four extra observations of soft-drink dispensers—two that are contaminated and two that are not. This augmentation has the effect of biasing the estimate towards the value 1/2. Generally speaking we would like to avoid biased estimates, but as we shall see in Section 9.2, confidence intervals based on this biased estimate, $\tilde{P}$, actually are more reliable than those based on $\hat{P}$.

The Sampling Distribution of $\tilde{P}$

For random sampling from a large dichotomous population, we saw in Chapter 3 how to use the binomial distribution to calculate the probabilities of all the various possible sample compositions. These probabilities in turn determine the sampling distribution of $\tilde{P}$. An example follows.

Example 9.1.3

Contaminated Soda Suppose that in a certain region of the United States, 17% of all soft-drink dispensers are contaminated with *Chryseobacterium meningosepticum.* If we were to examine a random sample of two drink dispensers from this population of dispensers, then we will get either zero, one, or two contaminated machines. The probability that both dispensers are contaminated is $0.17 \times 0.17 = 0.0289$. The probability that neither are contaminated is $(1 - 0.17) \times (1 - 0.17) = 0.6889$. There are two ways to get a sample in which one machine is contaminated and one is not: The first could be contaminated, but not the second, or vice versa. Thus, the probability that exactly one machine is contaminated is

$$0.17 \times (1 - 0.17) + 0.17 \times (1 - 0.17) = 0.2822$$

If we let $\tilde{P}$ represent the Wilson-adjusted sample proportion of contaminated dispensers, then a sample that contains no contaminated dispensers has $\tilde{p} = \dfrac{0 + 2}{2 + 4} = 0.33$, which occurs with probability 0.6889. A sample that contains one contaminated machine has $\tilde{p} = \dfrac{1 + 2}{2 + 4} = 0.50$; this happens with probability 0.2822. Finally, a sample that contains two contaminated machines has $\tilde{p} = \dfrac{2 + 2}{2 + 4} = 0.67$, which occurs with probability 0.0289.[†] Thus, there is roughly a 69% chance that $\tilde{P}$ will equal 0.33, a 28% chance that $\tilde{P}$ will equal 0.50, and a 3% chance that $\tilde{P}$ will equal 0.67.

[*] In keeping with our convention, $\tilde{P}$ denotes a random variable, whereas $\tilde{p}$ denotes a particular number (such as 0.206 in this example).

[†] It is worth noting that with a small sample size ($n = 2$) the possible values of $\tilde{p}$ are 0.33, 0.50, and 0.67 while the possible values of $\hat{p}$ are 0.00, 0.50, and 1.00. This sheds some light as to why $\tilde{p}$ is a sensible estimator of the population proportion, particularly for small samples. With a small sample it is quite likely that one could obtain no contaminated machines even if a reasonable proportion of the population is contaminated. It would be unwise, with such a small sample, to assert that the population proportion of contaminated machines is 0.

This sampling distribution is given in Table 9.1.1 and Figure 9.1.1. ■

Table 9.1.1 Sampling distribution of Y (the number of contaminated dispensers) and of $\tilde{P}$ (the Wilson-adjusted proportion of contaminated dispensers) for samples of size $n = 2$ for a population with 17% of the dispensers contaminated

Y	$\tilde{P}$	Probability
0	0.33	0.6889
1	0.50	0.2822
2	0.67	0.0289

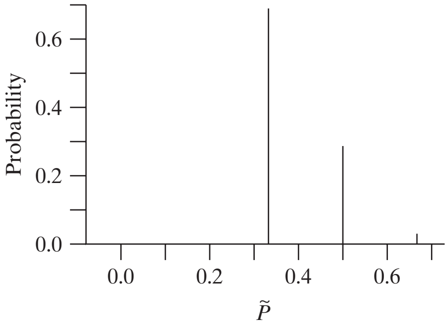


Figure 9.1.1 Sampling distribution of $\tilde{P}$ for $n = 2$ and $p = 0.17$

Example 9.1.4

Contaminated Soda and a Larger Sample Suppose we were to examine a sample of 20 dispensers from a population in which 17% are contaminated. How many contaminated dispensers might we expect to find in the sample? As was true in Example 9.1.3, this question can be answered in the language of probability. However, since $n = 20$ is rather large, we will not list each possible sample. Rather, we will make calculations using the binomial distribution with $n = 20$ and $p = 0.17$. For instance, let us calculate the probability that 5 dispensers in the sample would be contaminated and 15 would not:

$$\begin{aligned}\Pr\{5 \text{ contaminated, } 15 \text{ not contaminated}\} &= {}_{20}C_5(0.17)^5(0.83)^{15}\\ &= 15{,}504(0.17)^5(0.83)^{15}\\ &= 0.1345\end{aligned}$$

Letting $\tilde{P}$ represent the Wilson-adjusted sample proportion of contaminated dispensers, a sample that contains 5 contaminated dispensers has $\tilde{p} = \frac{5+2}{20+4} = 0.2917$. Thus, we have found that

$$\Pr\{\tilde{P} = 0.2917\} = 0.1345$$

The binomial distribution can be used to determine the entire sampling distribution of $\tilde{P}$. The distribution is displayed in Table 9.1.2 and as a probability histogram in Figure 9.1.2.

Table 9.1.2 Sampling distribution of Y, the number of successes, and of $\tilde{P}$, the Wilson-adjusted proportion of successes, when $n = 20$ and $p = 0.17$

Y	$\tilde{P}$	Probability	Y	$\tilde{P}$	Probability
0	0.0833	0.0241	11	0.5417	0.0001
1	0.1250	0.0986	12	0.5833	0.0000
2	0.1667	0.1919	13	0.6250	0.0000
3	0.2083	0.2358	14	0.6667	0.0000
4	0.2500	0.2053	15	0.7083	0.0000
5	0.2917	0.1345	16	0.7500	0.0000
6	0.3333	0.0689	17	0.7917	0.0000
7	0.3750	0.0282	18	0.8333	0.0000
8	0.4167	0.0094	19	0.8750	0.0000
9	0.4583	0.0026	20	0.9167	0.0000
10	0.5000	0.0006			

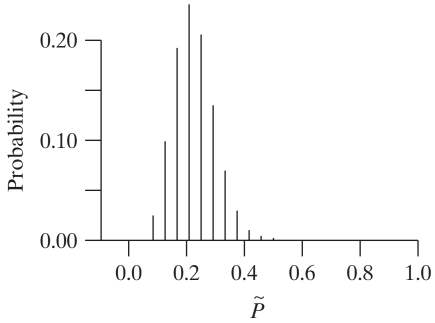


Figure 9.1.2 Sampling distribution of $\tilde{P}$ when $n = 20$ and $p = 0.17$

We can use this distribution to answer questions such as "If we take a random sample of size $n = 20$, what is the probability that no more than 5 will be contaminated?" Notice that this question can be asked in two equivalent ways: "What is $\Pr\{Y \leq 5\}$?" and "What is $\Pr\{\tilde{P} \leq 0.2917\}$?" The answer to either question is found by adding the first six probabilities in Table 9.1.2:

$$\begin{aligned}\Pr\{Y \leq 5\} &= \Pr\{\tilde{P} \leq 0.2917\}\\ &= 0.0241 + 0.0986 + 0.1919 + 0.2358 + 0.2053 + 0.1345\\ &= 0.8902\end{aligned}$$

■

Relationship to Statistical Inference

In making a statistical inference from a sample to the population, it is reasonable to use $\tilde{P}$ as our estimate of p. The sampling distribution of $\tilde{P}$ can be used to predict how much sampling error to expect in this estimate. For example, suppose we want to know whether the sampling error will be less than 5 percentage points, in other words, whether $\tilde{P}$ will be within ± 0.05 of p. We cannot predict for certain whether this event will occur, but we can find the probability of it happening, as illustrated in the following example.

Example 9.1.5

Contaminated Soda In the soda-dispenser example with $n = 20$, we see from Table 9.1.2 that

$$\begin{aligned}\Pr\{0.12 \leq \tilde{P} \leq 0.22\} &= 0.0986 + 0.1919 + 0.2358\\ &= 0.5263 \approx 0.53\end{aligned}$$

Thus, there is a 53% chance that, for a sample of size 20, $\tilde{P}$ will be within $\pm$ 0.05 of p. ■

Dependence on Sample Size

Just as the sampling distribution of $\overline{Y}$ depends on n, so does the sampling distribution of $\tilde{P}$. The larger the value of n, then the more likely it is $\tilde{P}$ will be close to p.* The following example illustrates this effect.

Example 9.1.5

Contaminated Soda Figure 9.1.3 shows the sampling distribution of $\tilde{P}$, for three different values of n, for the soft-drink dispenser population of Example 9.1.1. (Each sampling distribution is determined by a binomial distribution with $p = 0.17$.)

*This statement should not be interpreted too literally. As a function of n, the probability that $\tilde{P}$ is close to p has an overall increasing trend, but it can fluctuate somewhat.

Figure 9.1.3 Sampling distributions of $\widetilde{P}$ for $p = 0.17$ and various values of n

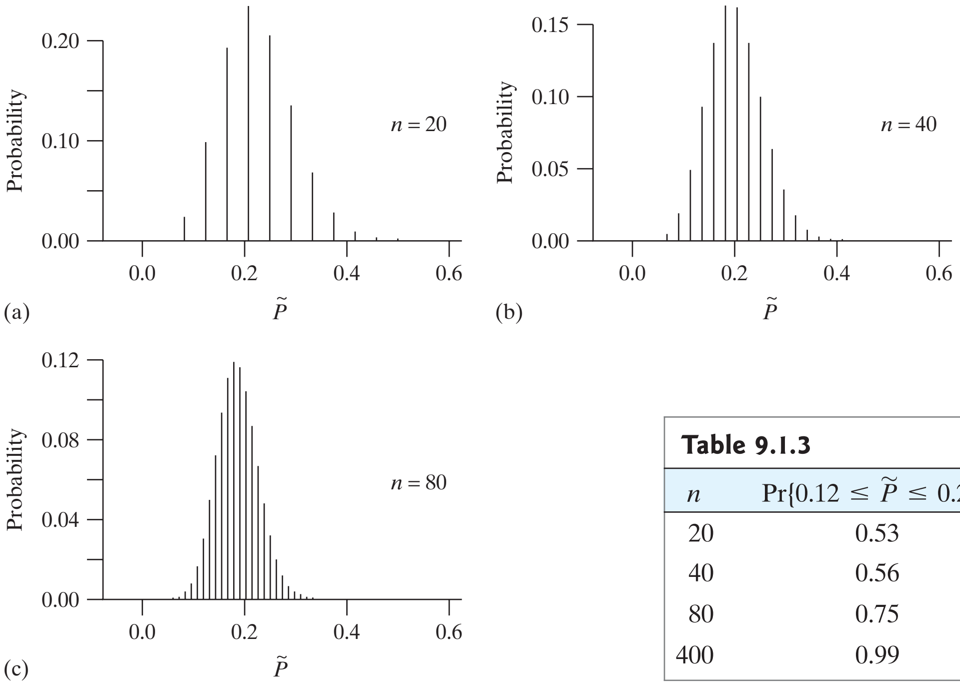


Table 9.1.3

n	$\Pr\{0.12 \le \widetilde{P} \le 0.22\}$
20	0.53
40	0.56
80	0.75
400	0.99

You can see from the figure that as n increases, the sampling distribution becomes more compressed around the value $p = 0.17$; thus, the probability that $\widetilde{P}$ is close to p tends to increase as n increases. For example, consider the probability that $\widetilde{P}$ is within ±5 percentage points of p. We saw in Example 9.1.5 that for $n = 20$ this probability is equal to 0.53; Table 9.1.3 and Figure 9.1.3 shows how the probability depends on n.

Note: A larger sample improves the probability that $\widetilde{P}$ will be close to p. We should be mindful, however, that the probability that $\widetilde{P}$ is exactly *equal* to p is very small for large n. In fact,

$$\Pr\{\widetilde{P} = 0.17\} = 0.110 \text{ for } n = 80^*$$

The value $\Pr\{0.12 \le \widetilde{P} \le 0.22\} = 0.75$ is the sum of many small probabilities, the largest of which is 0.110; you can see this effect clearly in Figure 9.1.3(c). ■

Exercises 9.1.1–9.1.10

9.1.1 Consider taking a random sample of size 3 from a population of persons who smoke and recording how many of them, if any, have lung cancer. Let $\widetilde{P}$ represent the Wilson-adjusted proportion of persons in the sample with lung cancer. What are the possible values in the sampling distribution of $\widetilde{P}$?

9.1.2 Suppose we are to draw a random sample of three individuals from a large population in which 37% of the individuals are mutants (as in Example 3.6.4). Let $\widetilde{P}$ represent the Wilson-adjusted proportion of mutants in the sample. Calculate the probability that $\widetilde{P}$ will be equal to

(a) 2/7 (b) 3/7

Is it possible to obtain a sample of three individuals for which $\widetilde{P}$ is zero? Explain.

9.1.3 Suppose we are to draw a random sample of five individuals from a large population in which 37% of the individuals are mutants (as in Example 3.6.4). Let $\widetilde{P}$ represent the Wilson-adjusted proportion of mutants in the sample.

(a) Use the results in Table 3.6.3 to determine the probability that $\widetilde{P}$ will be equal to

(i) 2/9 (ii) 3/9 (iii) 4/9
(iv) 5/9 (v) 6/9 (vi) 7/9

(b) Display the sampling distribution of $\widetilde{P}$ in a graph similar to Figure 9.1.1.

*For $n = 80$, $\widetilde{p} = 0.1677$ when $y = 12$, is the closest possible value to 0.17.

9.1.4 A new treatment for acquired immune deficiency syndrome (AIDS) is to be tested in a small clinical trial on 15 patients. The Wilson-adjusted proportion $\tilde{P}$ who respond to the treatment will be used as an estimate of the proportion p of (potential) responders in the entire population of AIDS patients. If in fact $p = 0.2$, and if the 15 patients can be regarded as a random sample from the population, find the probability that

(a) $\tilde{P} = 5/19$ (b) $\tilde{P} = 2/19$

9.1.5 In a certain forest, 25% of the white pine trees are infected with blister rust. Suppose a random sample of four white pine trees is to be chosen, and let $\tilde{P}$ be the Wilson-adjusted sample proportion of infected trees.

(a) Compute the probability that $\tilde{P}$ will be equal to

(i) 2/8 (ii) 3/8 (iii) 4/8 (iv) 5/8 (v) 6/8

(b) Display the sampling distribution of $\tilde{P}$ in a graph similar to Figure 9.1.1.

9.1.6 Refer to Exercise 9.1.5.

(a) Determine the sampling distribution of $\tilde{P}$ for samples of size $n = 8$ white pine trees from the same forest.

(b) Construct graphs of the sampling distributions of $\tilde{P}$ for $n = 4$ and for $n = 8$, using the same horizontal and vertical scales for both. Compare the two distributions visually. How do they differ?

9.1.7 The shell of the land snail *Limocolaria marfensiana* has two possible color forms: streaked and pallid. In a certain population of these snails, 60% of the individuals have streaked shells (as in Exercise 3.6.4). Suppose a random sample of six snails is to be chosen from the population; let $\tilde{p}$ be the Wilson-adjusted sample proportion of streaked snails. Find

(a) $\Pr\{\tilde{P} = 0.5\}$ (b) $\Pr\{\tilde{P} = 0.6\}$

(c) $\Pr\{\tilde{P} = 0.7\}$ (d) $\Pr\{0.5 \le \tilde{P} \le 0.7\}$

(e) the percentage of samples for which $\tilde{P}$ is within ±0.10 of p.

9.1.8 In a certain community, 17% of the soda dispensers are contaminated (as in Example 9.1.5). Suppose a random sample of five dispensers is to be chosen and the contamination observed. Let $\tilde{P}$ represent the Wilson-adjusted sample proportion contaminated dispensers.

(a) Compute the sampling distribution of $\tilde{P}$.

(b) Construct a histogram of the distribution found in part (a) and compare it visually with Figure 9.1.3. How do the two distributions differ?

9.1.9 Consider random sampling from a dichotomous population; let E be the event that $\tilde{P}$ is within ±.05 of p. In Example 9.1.5, we found that $\Pr\{E\} = 0.53$ for $n = 20$ and $p = 0.17$. Calculate $\Pr\{E\}$ for $n = 20$ and $p = 0.25$. (Perhaps surprisingly, the two probabilities are roughly equal.)

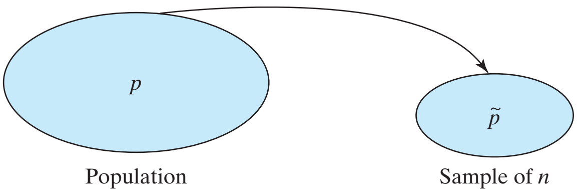
9.1.10 Consider taking a random sample of size 10 from the population of students at a certain college and asking each of the 10 students whether or not they smoke. In the context of this setting, explain what is meant by the sampling distribution of $\hat{P}$, the ordinary sample proportion.

9.2 Confidence Interval for a Population Proportion

In Section 6.3 we described confidence intervals when the observed variable is quantitative. Similar ideas can be used to construct confidence intervals in situations in which the variable is *categorical* and the parameter of interest is a population *proportion*. We assume that the data can be regarded as a random sample from some population. In this section we discuss construction of a confidence interval for a population proportion.

Consider a random sample of n categorical observations, and let us fix attention on one of the categories. For instance, suppose a geneticist observes n guinea pigs whose coat color can be either black, sepia, cream, or albino; let us fix attention on the category "black." Let p denote the population proportion of the category of interest, and let $\tilde{p}$ denote the Wilson-adjusted sample proportion (as in Section 9.1), which is our estimate of p. The situation is schematically represented in Figure 9.2.1.

Figure 9.2.1 Notation for population and sample proportion



How close to p is $\widetilde{P}$ likely to be? We saw in Section 9.1 that this question can be answered in terms of the sampling distribution of $\widetilde{P}$ (which in turn is computed from the binomial distribution). As we shall see, by using properties of the sampling distribution of $\widetilde{P}$, such as the standard error and $\widetilde{P}$'s approximately normal behavior under certain situations, we will be able to construct confidence statements for p. To construct the intervals we will use the same rationale used for numeric data in Section 6.3 where we constructed confidence statements for μ based on the properties of the sampling distribution of $\overline{Y}$.

Although a confidence interval for p can be constructed directly from the binomial distribution, for many practical situations a simple approximate method can be used instead. When the sample size, n, is large, the sampling distribution of $\widetilde{P}$ is approximately normal; this approximation is related to the Central Limit Theorem. If you review Figure 9.1.2, you will see that the sampling distributions resemble normal curves, especially the distribution with $n = 80$. (The approximation is described in detail in optional Section 5.4.) In Section 6.3 we stated that when the data come from a normal population, a 95% confidence interval for a population mean μ is constructed as

$$\bar{y} \pm t_{0.025}\mathrm{SE}_{\overline{Y}}$$

A confidence interval for a population proportion p is constructed analogously. We will use $\widetilde{P}$ as the center of a 95% confidence interval for p. In order to proceed we need to calculate the standard error for $\widetilde{P}$.

Standard Error of $\widetilde{P}$

The standard error of the estimate is found using the following formula.

Standard Error of $\widetilde{P}$ (for a 95% Confidence Interval)

$$\mathrm{SE}_{\widetilde{P}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

This formula for the standard error of the estimate looks similar to the formula for the standard error of a mean, but with $\sqrt{\tilde{p}(1-\tilde{p})}$ playing the role of s and with $n + 4$ in place of n.

Example 9.2.1

Smoking during Pregnancy As part of the National Survey of Family Growth, 496 women aged 20 to 24 who had given birth were asked about their smoking habits.[2] Smoking during pregnancy was reported by 78 of those sampled, which is 15.7 percent (78/496 = 0.157 or 15.7%). Thus, $\tilde{p}$ is $\frac{78+2}{496+4} = \frac{80}{500} = 0.16$; the standard error is $\sqrt{\frac{0.16(1-0.16)}{500}} = 0.016$ or 1.6%. A sample value of $\widetilde{P}$ is typically within ±2 standard errors of the population proportion p. Based on this standard error, we can expect that the proportion, p, of all women aged 20 to 24 who smoked during pregnancy is in the interval (0.128, 0.192) or (12.8%, 19.2%). A confidence interval for p makes this idea more precise. ■

95% Confidence Interval for p

Once we have the standard error of $\widetilde{P}$, we need to know how likely it is that $\widetilde{P}$ will be close to p. The general process of constructing a confidence interval for a proportion is similar to that used in Section 6.3 to construct a confidence interval for a mean. However, when constructing a confidence interval for a mean, we multiplied the standard error by a t multiplier. This was based on having a sample from a normal distribution. When dealing with proportion data we know that the population is not normal—there only are two values in the population!—but the Central Limit Theorem tells us that the sampling distribution of $\widetilde{P}$ is approximately normal if the sample size, n, is large. Moreover, it turns out that even for moderate or small samples, intervals based on $\widetilde{P}$ and Z multipliers do a very good job of estimating the population proportion, p.[3]

For a 95% confidence interval, the appropriate Z multiplier is $z_{0.025} = 1.960$. Thus, the approximate 95% confidence interval for a population proportion p is constructed as shown in the following box.*

95% Confidence Interval for p

95% confidence interval: $\widetilde{p} \pm 1.96\text{SE}_{\widetilde{P}}$

Example 9.2.2

Breast Cancer *BRCA1* is a gene that has been linked to breast cancer. Researchers used DNA analysis to search for *BRCA1* mutations in 169 women with family histories of breast cancer. Of the 169 women tested, 27 (16%) had *BRCA1* mutations.[4] Let p denote the probability that a woman with a family history of breast cancer will have a *BRCA1* mutation. For these data, $\widetilde{p} = \frac{27 + 2}{169 + 4} = 0.168$. The standard error for $\widetilde{P}$ is $\sqrt{\frac{0.168(1 - 0.168)}{169 + 4}} = 0.028$. Thus, a 95% confidence interval for p is

$$0.168 \pm (1.96)(0.028)$$

or

$$0.168 \pm 0.055$$

or

$$0.113 < p < 0.223$$

Thus, we are 95% confident that the probability of a BRCA1 mutation in a woman with a family history of breast cancer is between 0.113 and 0.223 (i.e., between 11.3% and 22.3%). ■

Note that the size of the standard error is inversely proportional to $\sqrt{n}$, as illustrated in the following example.

Example 9.2.3

Breast Cancer Suppose, as in Example 9.2.2, that a sample of n women with family histories of breast cancer contains 16% with *BRCA1* mutations. Then $\widetilde{p} \approx 0.168$ and

$$\text{SE}_{\widetilde{P}} \approx \sqrt{\frac{0.168(0.832)}{n + 4}}$$

We saw in Example 9.2.2 that if $n = 169$, then

$$\text{SE}_{\widetilde{P}} = 0.028$$

If $n = 4 \times 169 = 676$, then

$$\text{SE}_{\widetilde{P}} = 0.014$$

*Many statistics books present the confidence interval for a proportion as $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$ where $\hat{p} = y/n$. This commonly used interval is similar to the interval we present, particularly if n is large. For small or moderate sample sizes, the interval we present is more likely to cover the population proportion p. A technical discussion of the Wilson-interval using $\widetilde{P}$ is given in Appendix 9.1.

Thus, a sample with the same composition (that is, 16% with *BRCA1* mutations) but four times as large, would yield twice as much precision in the estimation of p. ■

The Wilson-adjusted sample proportion can be used to construct a confidence interval for p even when the sample size is small, as the following example illustrates.

Example 9.2.4

ECMO Extracorporeal membrane oxygenation (ECMO) is a potentially life-saving procedure that is used to treat newborn babies who suffer from severe respiratory failure. An experiment was conducted in which 11 babies were treated with ECMO; none of the 11 babies died.[5] Let p denote the probability of death for a baby treated with ECMO. The fact that none of the babies in the experiment died should not lead us to believe that the probability of death, p, is precisely zero—only that it is close to zero. The estimate given by $\tilde{p}$ is $2/15 = 0.133$. The standard error of $\tilde{p}$ is

$$\sqrt{\frac{0.133(0.867)}{15}} = 0.088^*$$

Thus, a 95% confidence interval for p is

$$0.133 \pm (1.96)(0.088)$$

or

$$0.133 \pm 0.172$$

or

$$-0.039 < p < 0.305$$

We know that p cannot be negative, so we state the confidence interval as (0, 0.305).

Thus, we are 95% confident that the probability of death in a newborn with severe respiratory failure who is treated with ECMO is between 0 and 0.305 (i.e., between 0% and 30.5%). ■

One-Sided Confidence Intervals

Most confidence intervals are of the form "estimate ± margin of error"; these are known as two-sided intervals. However, it is possible to construct a one-sided confidence interval, which is appropriate when only a lower bound, or only an upper bound, is of interest. The following example provides an illustration.

Example 9.2.5

ECMO—One-Sided Consider the ECMO data from Example 9.2.4, which are used to estimate the probability of death, p, in a newborn with severe respiratory failure. We know that p cannot be less than zero, but we might want to know how large p might be. Whereas a two-sided confidence interval is based on capturing the middle 95% of a standard normal distribution and thus uses the Z multipliers of ± 1.96, a one-sided 95% (upper) confidence interval uses the fact that $\Pr(-\infty < Z < 1.645) = 0.95$. Thus, the upper limit of the confidence interval is $\tilde{p} + 1.645 \times \mathrm{SE}_{\tilde{p}}$ and the lower limit of the interval is negative infinity. In this case we get

$$0.133 + (1.645)(0.088) = 0.133 + 0.145 = 0.278$$

as the upper limit. The resulting interval is $(-\infty, 0.278)$, but since p cannot be negative, we state the confidence interval as (0, 0.278). That is, we are 95% confident that the probability of death is at most 27.8%. ■

*Note that if we used the commonly presented method of $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ we would find that the standard error is zero, leading to a confidence interval of 0 ± 0. Such an interval would not seem to be very useful in practice!

Planning a Study to Estimate p

In Section 6.4 we discussed a method for choosing the sample size n so that a proposed study would have sufficient precision for its intended purpose. The approach depended on two elements: (1) a specification of the desired $\mathrm{SE}_{\overline{Y}}$ and (2) a preliminary guess of the SD. In the present context, when the observed variable is categorical, a similar approach can be used. If a desired value of $\mathrm{SE}_{\tilde{p}}$ is specified, and if a rough informed guess of $\tilde{p}$ is available, then the required sample size n can be determined from the following equation:

$$\text{Desired SE} = \sqrt{\frac{(\text{Guessed } \tilde{p})(1 - \text{Guessed } \tilde{p})}{n + 4}}$$

The following example illustrates the use of the method.

Example 9.2.6

Left-Handedness In a survey of English and Scottish college students, 40 of 400 male students were found to be left-handed.[6]

The sample estimate of the proportion is

$$\tilde{p} = \frac{40 + 2}{400 + 4} \approx 0.104$$

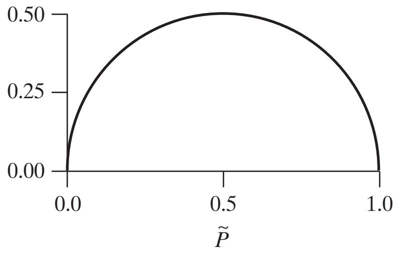
Suppose we regard these data as a pilot study and we now wish to plan a study large enough to estimate p with a standard error of one percentage point, that is, 0.01. We choose n to satisfy the following relation:

$$\sqrt{\frac{0.104(0.896)}{n + 4}} \leq 0.01$$

This equation is easily solved to give $n + 4 \geq 931.8$. We should plan a sample of 928 students. ■

Planning in Ignorance Suppose no preliminary informed guess of p is available. Remarkably, in this situation it is still possible to plan an experiment to achieve a desired value of $\mathrm{SE}_{\tilde{p}}$.* Such a "blind" plan depends on the fact that the crucial quantity $\sqrt{\tilde{p}(1 - \tilde{p})}$ is *largest* when $\tilde{p} = 0.5$; you can see this in the graph of Figure 9.2.2. It follows that a value of n calculated using "guessed $\tilde{p}$" $= 0.5$ will be *conservative*—that is, it will certainly be large enough. (Of course, it will be much larger than necessary if $\tilde{p}$ is really very different from 0.5.) The following example shows how such "worst-case" planning is used.

Figure 9.2.2 How $\sqrt{\tilde{p}(1 - \tilde{p})}$ depends on $\tilde{p}$



Example 9.2.7

Left-Handedness Suppose, as in Example 9.2.6, that we are planning a study of left-handedness and that we want $\mathrm{SE}_{\tilde{p}}$ to be 0.01, but suppose that we have no preliminary

*By contrast, it would not be possible if we were planning a study to estimate a population mean μ and we had no information whatsoever about the value of the SD.

information whatsoever. We can proceed as in Example 9.2.6, but using a guessed value of $\widetilde{p}$ of 0.5. Then we have

$$\sqrt{\frac{0.5(0.5)}{n + 4}} \leq 0.01$$

which means that $n + 4 \geq 2500$, so we need $n = 2{,}496$. Thus, a sample of 2,496 students would be adequate to estimate p with a standard error of 0.01, regardless of the actual value of p. (Of course, if $p = 0.1$, this value of n is much larger than is necessary.) ■

Exercises 9.2.1–9.2.13

9.2.1 A series of patients with bacterial wound infections were treated with the antibiotic Cefotaxime. Bacteriologic response (disappearance of the bacteria from the wound) was considered "satisfactory" in 84% of the patients.[7] Determine the standard error of $\widetilde{P}$, the Wilson-adjusted observed proportion of "satisfactory" responses, if the series contained

(a) 50 patients of whom 42 were considered "satisfactory."

(b) 200 patients of whom 168 were considered "satisfactory."

9.2.2 In an experiment with a certain mutation in the fruitfly *Drosophila*, n individuals were examined; of these, 20% were found to be mutants. Determine the standard error of $\widetilde{P}$ if

(a) $n = 100$ (20 mutants). (b) $n = 400$ (80 mutants).

9.2.3 Refer to Exercise 9.2.2. In each case ($n = 100$ and $n = 400$) construct a 95% confidence interval for the population proportion of mutants.

9.2.4 In a natural population of mice (*Mus musculus*) near Ann Arbor, Michigan, the coats of some individuals are white spotted on the belly. In a sample of 580 mice from the population, 28 individuals were found to have white-spotted bellies.[8] Construct a 95% confidence interval for the population proportion of this trait.

9.2.5 To evaluate the policy of routine vaccination of infants for whooping cough, adverse reactions were monitored in 339 infants who received their first injection of vaccine. Reactions were noted in 69 of the infants.[9]

(a) Construct a 95% confidence interval for the probability of an adverse reaction to the vaccine.

(b) Interpret the confidence interval from part (a). What does the interval say about whooping cough vaccinations?

(c) Using your interval from part (a), can we be confident that the probability of an adverse reaction to the vaccine is less than 0.25?

(d) What level of confidence is a associated with your answer to part (c)? (*Hint*: What is the associated one-sided interval confidence level?)

9.2.6 In a study of human blood types in nonhuman primates, a sample of 71 orangutans were tested and 14 were found to be blood type B.[10] Construct a 95% confidence interval for the relative frequency of blood type B in the orangutan population.

9.2.7 In populations of the snail *Cepaea*, the shells of some individuals have dark bands, while other individuals have unbanded shells.[11] Suppose that a biologist is planning a study to estimate the percentage of banded individuals in a certain natural population, and that she wants to estimate the percentage—which she anticipates will be in the neighborhood of 60%—with a standard error not to exceed 4 percentage points. How many snails should she plan to collect?

9.2.8 (Continuation of Exercise 9.2.7) What would the answer be if the anticipated percentage of banded snails were 50% rather than 60%?

9.2.9 The ability to taste the compound phenylthiocarbamide (PTC) is a genetically controlled trait in humans. In Europe and Asia, about 70% of people are "tasters."[12] Suppose a study is being planned to estimate the relative frequency of tasters in a certain Asian population, and it is desired that the standard error of the estimated relative frequency should be 0.01. How many people should be included in the study?

9.2.10 Refer to Exercise 9.2.9. Suppose a study is being planned for a part of the world for which the percentage of tasters is completely unknown, so that the 70% figure used in Exercise 9.2.9 is not applicable. What sample size is needed so that the standard error will be no larger than 0.01?

9.2.11 Refer to Exercise 9.2.9. Suppose the SE requirement is relaxed by a factor of 2—from 0.01 to 0.02. Would this reduce the required sample size by a factor of 2? Explain.

9.2.12 The "Luso" variety of wheat is resistant to the Hessian fly. In order to understand the genetic mechanism controlling this resistance, an agronomist plans to examine the progeny of a certain cross involving Luso and a nonresistant variety. Each progeny plant will be classified as resistant or susceptible and the agronomist will estimate the proportion of progeny that are resistant.[13] How many progeny does he need to classify in order to guarantee that the standard error of his estimate of this proportion will not exceed 0.05?

9.2.13 (Continuation of Exercise 9.2.12) Suppose the agronomist is considering two possible genetic mechanisms for the inheritance of resistance; the population ratio of resistant to susceptible progeny would be 1:1 under one mechanism and 3:1 under the other. If the agronomist uses the sample size determined in Exercise 9.2.12, can he be sure that a 95% confidence interval will exclude at least one of the mechanisms? That is, can he be sure that the confidence interval will *not* contain both 0.50 and 0.75? Explain.

9.3 Other Confidence Levels (Optional)

The procedure outlined in Section 9.2 can be used to construct 95% confidence intervals. In order to construct intervals with other confidence coefficients, some modifications to the procedure are needed. The first modification concerns $\tilde{p}$. For a 95% confidence interval we defined $\tilde{p}$ to be $\frac{y+2}{n+4}$. In general, for a confidence interval of level $100(1-\alpha)\%$, $\tilde{p}$ is defined as

$$\tilde{p} = \frac{y + 0.5(z^2_{\alpha/2})}{n + z^2_{\alpha/2}}$$

For a 95% confidence interval $z_{\alpha/2}$ is 1.96, so that $\tilde{p} = \frac{y + 0.5(1.96^2)}{n + 1.96^2}$. This is equal to $\frac{y+1.92}{n+3.84}$, which we rounded off as $\frac{y+2}{n+4}$. However, any confidence level can be used. As an example, for a 90% confidence interval, $\tilde{p} = \frac{y + 0.5(1.645^2)}{n + 1.645^2}$; this is equal to $\frac{y+1.35}{n+2.7}$.

The second modification concerns the standard error. For a 95% confidence interval we used $\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$ as the standard error term. In general, we use $\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+z^2_{\alpha/2}}}$ as the standard error term.

Finally, the Z multiplier must match the confidence level (1.645 for a 90% confidence interval, etc.). These can be found most easily from Table 4 with df $= \infty$. (Recall from Section 6.3 that the t distribution with df $= \infty$ is a normal (Z) distribution.) The following example illustrates these modifications.

Example 9.3.1

Left-Handedness In Example 9.2.6 we considered a survey of English and Scottish college students where 40 of 400 male students were found to be left-handed. Let us construct a 90% confidence interval for the proportion, p, of left-handed individuals in the population.[6]

The sample estimate of the proportion is

$$\tilde{p} = \frac{40 + 0.5(1.645^2)}{400 + 1.645^2} = \frac{40 + 1.35}{400 + 2.7} \approx 0.103$$

and the SE is

$$\sqrt{\frac{0.103(0.897)}{402.7}} = 0.015$$

A 90% confidence interval for p is

$$0.103 \pm (1.645)(0.015)$$

or

$$0.078 < p < 0.128$$

Thus, we are 90% confident that between 7.8% and 12.8% of the population that was sampled are left-handed. ■

Exercises 9.3.1–9.3.4

9.3.1 In a sample of 848 children aged 3 to 5 it was found that 3.7% of them had iron deficiency.[14] Use these data to construct a 90% confidence interval for the proportion of all 3- to 5-year-old children with iron deficiency.

9.3.2 Researchers tested patients with cardiac pacemakers to see if use of a cellular telephone interferes with the operation of the pacemaker. There were 959 tests conducted for one type of cellular telephone; interference with the pacemaker (detected with electrocardiographic monitoring) was found in 15.7% of these tests.[15]

(a) Use these data to construct an appropriate 90% confidence interval.

(b) The confidence interval from part (a) is a confidence interval for what quantity? Answer in the context of the setting.

9.3.3 Gene mutations have been found in patients with muscular dystrophy. In one study, it was found that there were defects in the gene coding of sarcoglycan proteins in 23 of 180 patients with limb-girdle muscular dystrophy.[16] Use these data to construct a 99% confidence interval for the corresponding population proportion.

9.3.4 In an ecological study of the Carolina Junco, 53 birds were captured from a certain population; of these, 40 were male.[17] Use these data to construct a 90% confidence interval for the proportion of male birds in the Carolina Junco population.

9.4 Inference for Proportions: The Chi-Square Goodness-of-Fit Test

In Section 9.2 we described methods for constructing confidence intervals when the observed variable is categorical. We now turn our attention to hypothesis testing for categorical data. We will begin by considering analysis of a single sample of categorical data. We assume that the data can be regarded as a random sample from some population and we will test a null hypothesis, H_0, that specifies the population proportions, or probabilities, of the various categories. Here is an example.

Example 9.4.1

Deer Habitat and Fire Does fire affect deer behavior? Six months after a fire burned 730 acres of homogenous deer habitat, researchers surveyed a 3,000-acre parcel surrounding the area, which they divided into four regions: the region near the heat of the burn (1), the inside edge of the burn (2), the outside edge of the burn (3), and the area outside of the burned area (4); see Figure 9.4.1 and Table 9.4.1.[18] The null hypothesis is that that deer show no preference to any particular type of burned/unburned habitat—that they are randomly distributed over the 3,000 acres. The alternative hypothesis is that the deer do show a preference for some of the regions—that they are not randomly distributed across all 3,000 acres.

Under the null hypothesis, if deer were randomly distributed over the 3,000 acres, then we would expect the counts of deer in the regions to be in proportion to the sizes of the regions. Expressing the null hypothesis numerically we have the following probabilities of sighting deer:

$$
\begin{aligned}
H_0\text{: } \Pr\{\text{inner burn}\} &= \frac{520}{3{,}000} = 0.173 \\
\Pr\{\text{inner edge}\} &= \frac{210}{3{,}000} = 0.070 \\
\Pr\{\text{outer edge}\} &= \frac{240}{3{,}000} = 0.080 \\
\Pr\{\text{outer unburned}\} &= \frac{2{,}030}{3{,}000} = 0.677
\end{aligned}
$$

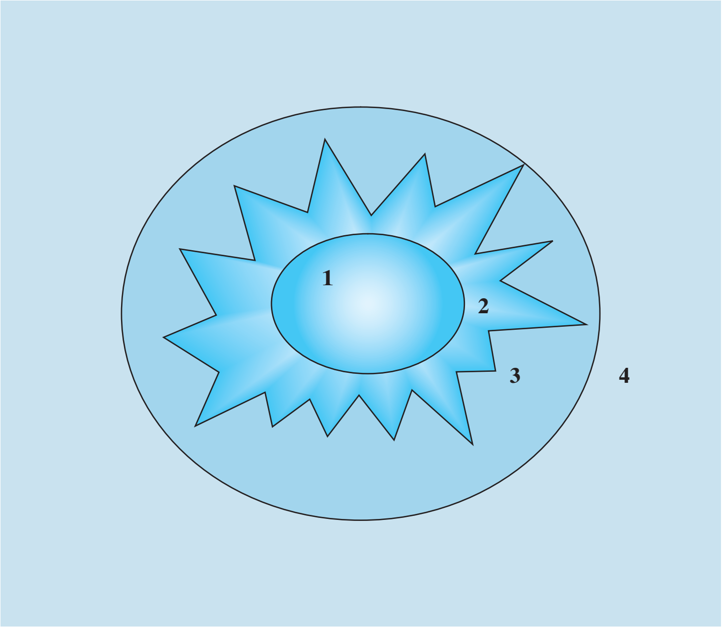


Figure 9.4.1 Schematic of three thousand-acre parcel with an interior 730-acre fire (not to scale)

Table 9.4.1 Deer distribution

Region	Acres	Proportion
1. Inner burn	520	0.173
2. Inner edge	210	0.070
3. Outer edge	240	0.080
4. Outer unburned	2,030	0.677
	3,000	1.000

Because the alternative hypothesis is not specific (it only states that the deer prefer some regions over others but doesn't indicate the nature of the preference), there is no simple symbolic way to express the alternative hypothesis. Thus, typically we do not use a symbolic representation. If we chose to express the alternative symbolically we could write:

$$H_A\text{: Pr\{inner burn\}} \neq 0.173, \text{and/or Pr\{inner edge\}} \neq 0.070, \text{and/or}$$
$$\text{Pr\{outer edge\}} \neq 0.080, \text{and/or Pr\{outer unburned\}} \neq 0.677$$ ■

Given a random sample of n categorical observations, how can one judge whether they provide evidence against a null hypothesis H_0 that specifies the probabilities of the categories? There are two complementary approaches to this question: The first considers an examination of the observed relative frequencies of each category while the second examines the frequencies directly. Considering the first method, the observed relative frequencies serve as estimates of the probabilities of the categories. The following notation for relative frequencies is useful: When a probability Pr{E} is estimated from observed data, the estimate is denoted by a hat ("^"); thus,

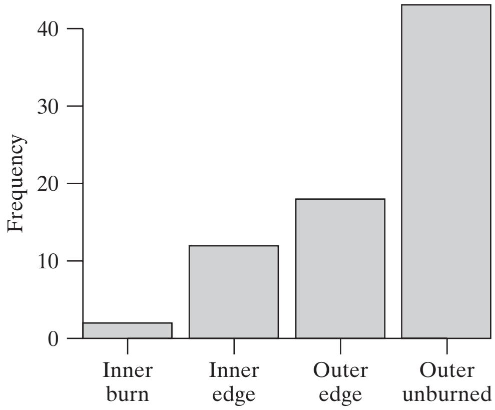
$$\hat{\Pr}\{E\} = \text{the estimated probability of event } E$$

Example 9.4.2

Deer Habitat and Fire Researchers observed a total of 75 deer in the 3,000-acre parcel described in Example 9.4.1: Two were in the region near the heat of the burn (Region 1), 12 were on the inside edge of the burn (Region 2), 18 were on the outside edge of the burn (Region 3), and 43 were outside of the burned area (Region 4).

These data are shown in Figure 9.4.2.

Figure 9.4.2 Bar chart of deer distribution data



The estimated category probabilities are

$$\hat{\Pr}\{\text{inner burn}\} = \frac{2}{75} = 0.027$$

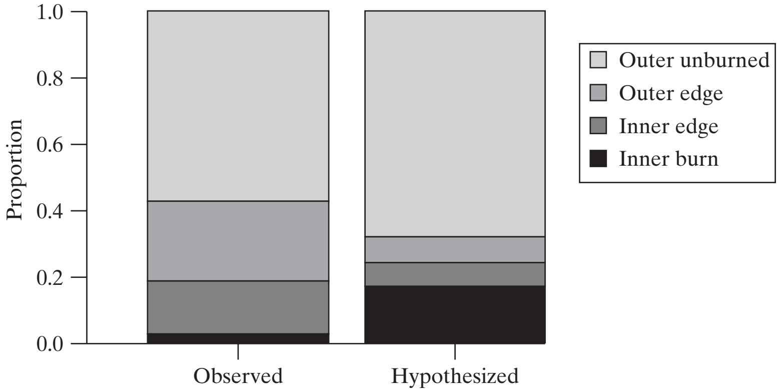
$$\hat{\Pr}\{\text{inner edge}\} = \frac{12}{75} = 0.160$$

$$\hat{\Pr}\{\text{outer edge}\} = \frac{18}{75} = 0.240$$

$$\hat{\Pr}\{\text{outer unburned}\} = \frac{43}{75} = 0.573$$

These estimated probabilities differ quite a bit from those in the model that is specified by H_0. Figure 9.4.3 shows stacked bar charts of the observed and hypothesized proportions. ■

Figure 9.4.3 Stacked bar charts of the deer proportions



The Chi-Square Statistic

The second approach, which considers the actual frequencies, is to use a statistical test, called a **goodness-of-fit test**, to assess the compatibility of the data with H_0. The most widely used goodness-of-fit test is the **chi-square test** or χ^2 test (χ is the Greek letter "chi").

The calculation of the chi-square test statistic is done in terms of the absolute, rather than the relative, frequencies of the categories. For each category level, i, let o_i represent the **observed frequency** of the category and let e_i represent the **expected frequency**—that is, the frequency that would be expected according to H_0. The e_i's are calculated by multiplying each probability specified in H_0 by n, as shown in Example 9.4.3.

Example 9.4.3

Deer Habitat and Fire Consider the null hypothesis specified in Example 9.4.1 and the data from Example 9.4.2. If the null hypothesis is true, then we expect 17.3% of the 75 deer to be in the inner burn region; 17.3% of 75 is 13.0:

$$\text{Inner burn: } e_1 = (0.173)(75) = 13.00$$

The corresponding expected frequencies for the other regions are

$$\text{Inner edge: } e_2 = (0.070)(75) = 5.25$$
$$\text{Outer edge: } e_3 = (0.080)(75) = 6.00$$
$$\text{Outer unburned: } e_4 = (0.677)(75) = 50.75$$

■

The test statistic for the chi-square goodness-of-fit test is then calculated from the o_i's and the e_i's using the formula given in the accompanying box with k equal to the number of category levels. Example 9.4.4 illustrates the calculation of the chi-square statistic.

The Chi-Square Statistic

$$\chi_s^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

where the summation is over all k categories.

Example 9.4.4

Deer Habitat and Fire The observed frequencies of 75 deer locations are

Region	Inner Burn	Inner Edge	Outer Edge	Outer Unburned	Total
Observed (o_i)	2	12	18	43	75

The expected frequencies are

Region	Inner Burn	Inner Edge	Outer Edge	Outer Unburned	Total
Expected (e_i)	13	5.25	6	50.75	75

Note that the sum of the expected frequencies is the same as the sum of the observed frequencies (75). The χ^2 statistic is

$$\chi_s^2 = \frac{(2-13)^2}{13} + \frac{(12-5.25)^2}{5.25} + \frac{(18-6)^2}{6} + \frac{(43-50.75)^2}{50.75}$$
$$= 43.2$$

■

Computational Note: In calculating a chi-square statistic the o_i's must be *absolute*, rather than relative, frequencies.

The χ^2 Distribution

From the way in which χ_s^2 is defined, it is clear that small values of χ_s^2 would indicate that the data agree with H_0, while large values of χ_s^2 would indicate disagreement. In order to base a statistical test on this agreement or disagreement, we need to know how much χ_s^2 may be affected by sampling variation.

We consider the null distribution of χ_s^2—that is, the sampling distribution that χ_s^2 follows if H_0 is true. It can be shown (using the methods of mathematical statistics) that, if the sample size is large enough, then the null distribution of χ_s^2 can be approximated by a distribution known as a **χ^2 distribution**. The form of a χ^2 distribution depends on a parameter called "degrees of freedom" (df). Figure 9.4.4 shows the χ^2 distribution with df $= 5$.

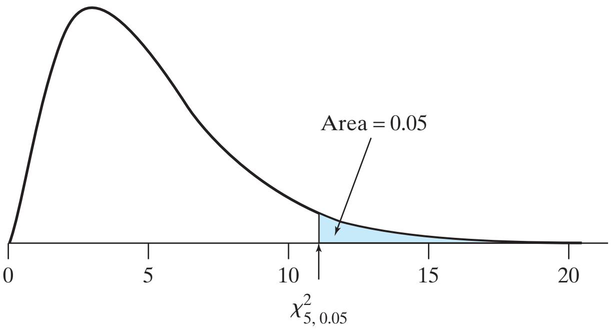


Figure 9.4.4 The χ^2 distribution with df $= 5$

Table 9 (at the end of this book) gives critical values for the χ^2 distribution. For instance, for df $= 5$, the 5% critical value is $\chi^2_{5,\,0.05} = 11.07$. This critical value corresponds to an area of 0.05 in the upper tail of the χ^2 distribution, as shown in Figure 9.4.4.

The Goodness-of-Fit Test

For the chi-square goodness-of-fit test we have presented, the null distribution of χ_s^2 is approximately a χ^2 distribution with*

$$\text{df} = k - 1, \text{where } k \text{ equals the number of categories}$$

For example, for the setting presented in Example 9.4.4 there are four categories so $k = 4$. The null hypothesis specifies the probabilities for each of the four categories. However, once the first three probabilities are specified, the last one is determined, since the four probabilities must sum to 1. There are four categories, but only three of them are "free"; the last one is constrained by the first three.

The test of H_0 is carried out using critical values from Table 9, as illustrated in the following example.

*The chi-square test can be extended to more general situations in which parameters are estimated from the data before the expected frequencies are calculated. In general, the degrees of freedom for the test are (number of categories) − (number of parameters estimated) −1. We are considering only the case in which there are no parameters to be estimated from the data.

Example 9.4.5

Deer Habitat and Fire For the deer habitat data of Example 9.4.4, the observed chi-square statistic was $\chi^2_s = 43.2$. Because there are four categories, the degrees of freedom for the null distribution are calculated as

$$\text{df} = 4 - 1 = 3$$

From Table 9 with df = 3 we find that $\chi^2_{3,\,0.0001} = 21.11$. Since $\chi^2_s = 43.2$ is greater than 21.11, the upper tail area beyond 43.2 is less than 0.0001. Thus, the P-value is less than 0.0001 and we have strong evidence against H_0 and in favor of the alternative hypothesis that the deer show preference for some areas over others. Upon comparing the observed and expected frequencies (or equivalently the hypothesized and estimated probabilities), we note that deer moved away from the burned and unburned regions (1) and (4) to be near the edge regions (2) and (3) (where there is likely to be new growth of vegetation yet proximity to old-growth shelter). ■

The chi-square test can be used with any number of categories. In Example 9.4.6 the test is applied to a variable with six categories.

Example 9.4.6

Flax Seeds Researchers studied a mutant type of flax seed that they hoped would produce oil for use in margarine and shortening. The amount of palmitic acid in the flax seed was an important factor in this research; a related factor was whether the seed was brown or was variegated. The seeds were classified into six combinations of palmitic acid and color, as shown in Table 9.4.2.[19] According to a hypothesized (Mendelian) genetic model, the six combinations should occur in a 3:6:3:1:2:1 ratio. That is, brown and low acid level should occur with probability 3/16, brown and intermediate acid level should occur with probability 6/16, and so on. The null hypothesis is that the model is correct; the alternative hypothesis is that the model is incorrect. The χ^2 statistic is

$$\begin{aligned}\chi^2_s &= \frac{(15-13.5)^2}{13.5} + \frac{(26-27)^2}{27} + \frac{(15-13.5)^2}{13.5} + \frac{(0-4.5)^2}{4.5} + \frac{(8-9)^2}{9} + \frac{(8-4.5)^2}{4.5} \\ &= 7.71\end{aligned}$$

Table 9.4.2 Flax seed distribution

Color	Acid level	Observed (o_i)	Expected (e_i)
Brown	Low	15	13.5
Brown	Intermediate	26	27
Brown	High	15	13.5
Variegated	Low	0	4.5
Variegated	Intermediate	8	9
Variegated	High	8	4.5
Total		72	72

The χ^2 test has $6 - 1 = 5$ degrees of freedom. From Table 9 with df = 5, we find that $\chi^2_{5,\,0.20} = 7.29$ and $\chi^2_{5,\,0.10} = 9.24$. Thus, the P-value is bracketed as $0.10 < P\text{-value} < 0.20$. If the level of α chosen for the test is 0.10 or smaller, then the P-value is larger than α and we would not reject H_0. We conclude that there is no significant evidence that the data are inconsistent with the Mendelian model. (Note that we have not necessarily demonstrated that the Mendelian model is correct, only that we cannot reject this model.) ■

Note that the critical values for the chi-square test do not depend on the sample size, n. However, the test procedure *is* affected by n, through the value of the chi-square statistic. If we change the size of a sample while keeping its percentage composition fixed, then χ_s^2 varies directly as the sample size, n. For instance, imagine appending a replicate of a sample to the sample itself. Then the expanded sample would have twice as many observations as the original, but they would be in the same relative proportions. The value of each o_i would be doubled, the value of each e_i would be doubled, and so the value of χ^2 would be doubled [because in each term of χ_s^2 the numerator $(o_i - e_i)^2$ would be multiplied by 4, and the denominator e_i would be multiplied by 2]. That is, the value of χ_s^2 would go up by a factor of 2, despite the fact that the pattern in the data stayed the same! In this way, an increased sample size magnifies any discrepancy between what is observed and what is expected under the null hypothesis.

Compound Hypotheses and Directionality

Let us examine the goodness-of-fit null hypothesis more closely. In a two-sample comparison such as a t test, the null hypothesis contains exactly one assertion—for instance, that two population means are equal. By contrast, a goodness-of-fit null hypothesis can contain more than one assertion. Such a null hypothesis may be called a **compound null hypothesis**. An example follows.

Example 9.4.7 Deer Habitat and Fire The null hypothesis of Example 9.4.1 is

$$H_0\text{: } \Pr\{\text{inner burn}\} = 0.173, \Pr\{\text{inner edge}\} = 0.070, \Pr\{\text{outer edge}\} = 0.080, \\ \Pr\{\text{outer unburned}\} = 0.677$$

This is a compound hypothesis because it makes three independent assertions, namely

$$\Pr\{\text{inner burn}\} = 0.173,\ \Pr\{\text{inner edge}\} = 0.070, \text{and}\ \Pr\{\text{outer edge}\} = 0.080$$

Note that the fourth assertion ($\Pr\{\text{outer unburned}\} = 0.677$) is not an independent assertion because it follows from the other three. ■

When the null hypothesis is compound, the chi-square test has two special features. First, the alternative hypothesis is necessarily nondirectional. Second, if H_0 is rejected, the test does not yield a directional conclusion. (However, if H_0 is rejected, then an examination of the observed proportions will sometimes show an interesting pattern of departure from H_0, as in Example 9.4.5.)

When H_0 is compound, the chi-square test is nondirectional in nature (perhaps "omnidirectional" would be a better term) because the chi-square statistic measures deviations from H_0 in all directions. Statistical methods are available that do yield directional conclusions and that can handle directional alternatives, but such methods are beyond the scope of this book.

Dichotomous Variables

If the categorical variable analyzed by a goodness-of-fit test is dichotomous, then the null hypothesis is not compound, and directional alternatives and directional conclusions do not pose any particular difficulty.*

Directional Conclusion

The following example illustrates the directional conclusion.

Example 9.4.8

Deer Habitat, Fire, and Two Regions Suppose that the deer habitat data of Example 9.4.1 had been presented as being from only two regions, A and B, where region A is the area at the edge of the fire, which combines regions (2) and (3), and region B is the remainder of the parcel, combining regions (1) and (4). There were 30 deer seen in region A and 45 deer seen in region B. Is this evidence that deer prefer one region over the other?

An appropriate null hypothesis is

$$H_0\text{: Pr}\{\text{region A}\} = \frac{450}{3{,}000} = 0.15,\ \text{Pr}\{\text{region B}\} = \frac{2{,}550}{3{,}000} = 0.85$$

This hypothesis is not compound because it contains only one independent assertion. (Note that the second assertion—Pr{region B} = 0.85—is redundant; it follows from the first.)

Let us test H_0 against the nondirectional alternative

$$H_A\text{: Pr}\{\text{region A}\} \neq 0.15$$

The observed and expected frequencies are shown in Table 9.4.3.

Table 9.4.3 Deer habitat data for two regions

	A	B	Total
Observed	30	45	75
Expected	11.25	63.75	75

The data yield $\chi_s^2 = 36.8$ and from Table 9 we find that $P < 0.0001$. Even at $\alpha = 0.0001$ we would reject H_0 and find that there is sufficient evidence to conclude that the population of deer prefers one region over the other. Comparing the observed and expected counts we observe that they prefer region A over region B.

To recapitulate, the directional conclusion in Example 9.4.8 is legitimate because we know that if H_0 is false, then necessarily either Pr{region A} < 0.15 or Pr{region A} > 0.15. By contrast, in Example 9.4.7 H_0 may be false but Pr{outer unburned} may still be equal to 0.677; the chi-square analysis does not determine which of the probabilities are not as specified by H_0.

*When the data are dichotomous, there is an alternative to the goodness-of-fit test that is known as the Z test for a single proportion. The calculations used in the Z test look quite different from those of the goodness-of-fit test but, in fact, the two tests are mathematically equivalent. However, unlike the goodness-of-fit test, which can handle any number of categories, the Z test can be used only when the data are limited to two categories. Thus, we do not present it here.

Directional Alternative A chi-square goodness-of-fit test against a directional alternative (when the observed variable is dichotomous) uses the familiar two-step procedure:

Step 1 Check directionality (see if the data deviate from H_0 in the direction specified by H_A).

(a) If not, the P-value is greater than 0.50.

(b) If so, proceed to step 2.

Step 2 The P-value is half what it would be if H_A were nondirectional.

The following example illustrates the procedure.

Example 9.4.9

Harvest Moon Festival Can people who are close to death postpone dying until after a symbolically meaningful occasion? Researchers studied death from natural causes among elderly Chinese women (over age 75) living in California. They chose to study the time around the Harvest Moon Festival because (1) the date of the traditional Chinese festival changes somewhat from year to year, making it less likely that a time-of-year effect would be confounded with the effect they were studying and (2) it is a festival in which the role of the oldest woman in the family is very important.

Previous research had suggested that there might be a decrease in the mortality rate among elderly Chinese women immediately prior to the festival, with a corresponding increase afterwards. The researchers found that over a period of several years there were 33 deaths in the group in the week preceding the Harvest Moon Festival and 70 deaths in the week following the festival.[20] How strongly does this support the interpretation that people can prolong life until a symbolically meaningful event?

We may formulate null and alternative hypotheses as follows:

H_0: Given that an elderly Chinese woman dies within one week of the Harvest Moon Festival, she is equally likely to die before the festival or after the festival.

H_A: Given that an elderly Chinese woman dies within one week of the Harvest Moon Festival, she is more likely to die after the festival than before the festival.

These hypotheses can be translated as

$$H_0\text{: Pr\{die after festival\}} = \frac{1}{2}$$

$$H_A\text{: Pr\{die after festival\}} > \frac{1}{2}$$

where it is understood that Pr{die after festival} is the probability of death after the festival, given that the woman dies within one week before or after the festival. The observed and expected frequencies are shown in Table 9.4.4.

Table 9.4.4 Harvest Moon Festival data

	Before	After	Total
Observed	33	70	103
Expected	51.5	51.5	103

From the data on the 103 deaths, we first note that the data do, indeed, deviate from H_0 in the direction specified by H_A, because the observed relative frequency of deaths after the festival is 70/103, which is greater than 1/2. The value of the chi-square statistic is $\chi^2_s = 13.3$; from Table 9 we see that the P-value would have been bracketed between 0.0001 and 0.001 had H_A been nondirectional. However, for the directional alternative hypothesis specified in this test, we bracket the P-value as $0.00005 < P\text{-value} < 0.0005$. We conclude that the evidence is very strong that the death rate among elderly Chinese women goes up after the festival.* ■

Exercises 9.4.1–9.4.13

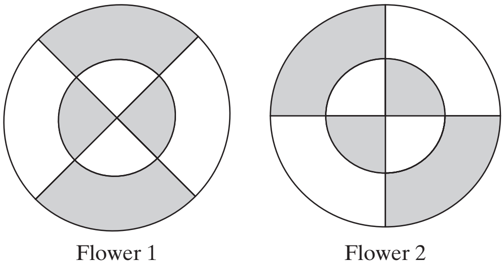
9.4.1 A cross between white and yellow summer squash gave progeny of the following colors:[21]

COLOR	WHITE	YELLOW	GREEN
Number of progeny	155	40	10

Are these data consistent with the 12:3:1 ratio predicted by a certain genetic model? Use a chi-square test at $\alpha = 0.10$.

9.4.2 Refer to Exercise 9.4.1. Suppose the sample had the same composition but was 10 times as large: 1,550 white, 400 yellow, and 100 green progeny. Would the data be consistent with the 12:3:1 model?

9.4.3 How do bees recognize flowers? As part of a study of this question, researchers used the following two artificial "flowers":[22]



The experiment was conducted as a series of trials on individual bees; each trial consisted of presenting a bee with both flowers and observing which flower it landed on first. (Flower 1 was sometimes on the left and sometimes on the right.) During the "training" trials, flower 1 contained a sucrose solution and flower 2 did not; thus, the bee was trained to prefer flower 1. During the testing trials, neither flower contained sucrose. In 25 testing trials with a particular bee, the bee chose flower 1 twenty times and flower 2 five times.

Use a goodness-of-fit test to assess the evidence that the bee could remember and distinguish the flower patterns. Use a directional alternative and let $\alpha = 0.05$.

9.4.4 At a midwestern hospital there was a total of 932 births in 20 consecutive weeks. Of these births, 216 occurred on weekends.[23] Do these data reveal more than chance deviation from random timing of the births? (Test for goodness of fit, with two categories of births: weekday and weekend. Use a nondirectional alternative and let $\alpha = 0.05$.)

9.4.5 In a breeding experiment, white chickens with small combs were mated and produced 190 offspring of the types shown in the accompanying table.[24] Are these data consistent with the Mendelian expected ratios of 9:3:3:1 for the four types? Use a chi-square test at $\alpha = 0.10$.

TYPE	NUMBER OF OFFSPRING
White feathers, small comb	111
White feathers, large comb	37
Dark feathers, small comb	34
Dark feathers, large comb	8
Total	190

9.4.6 Among n babies born in a certain city, 51% were boys.[25] Suppose we want to test the hypothesis that the true probability of a boy is $\frac{1}{2}$. Calculate the value of χ^2_s, and bracket the P-value for testing against a nondirectional alternative, if

(a) $n = 1{,}000$

(b) $n = 5{,}000$

(c) $n = 10{,}000$

*Based on these results, one might jump to the conclusion that this festival should be canceled to protect elderly Chinese women. As this study is only observational, however, we must not jump to causal conclusions!

9.4.7 In an agronomy experiment peanuts with shriveled seeds were crossed with normal peanuts. The genetic model that the agronomists were considering predicted that the ratio of normal to shriveled progeny would be 3:1. They obtained 95 normal and 54 shriveled progeny.[26] Do these data support the hypothesized model?

Conduct a chi-square test with $\alpha = 0.05$. Use a nondirectional alternative.

9.4.8 An experimental design using litter-matching was employed to test a certain drug for cancer-causing potential. From each of 50 litters of rats, three females were selected; one of these three, chosen at random, received the test drug, and the other two were kept as controls. During a two-year observation period, the time of occurrence of a tumor, and/or death from various causes, was recorded for each animal. One way to analyze the data is to note simply which rat (in each triplet) developed a tumor first. Some triplets were uninformative on this point because either (a) none of the three littermates developed a tumor, or (b) a rat developed a tumor after its littermate had died from some other cause. The results for the 50 triplets are shown in the table.[27] Use a goodness-of-fit test to evaluate the evidence that the drug causes cancer. Use a directional alternative and let $\alpha = 0.01$. State your conclusion from part (a) in the context of this setting. (*Hint*: Use only the 20 triplets that provide complete information.)

	NUMBER OF TRIPLETS
Tumor first in the treated rat	12
Tumor first in one of the two control rats	8
No tumor	23
Death from another cause	7
Total	50

9.4.9 A study of color vision in squirrels used an apparatus containing three small translucent panels that could be separately illuminated. The animals were trained to choose, by pressing a lever, the panel that appeared different from the other two. (During these "training" trials, the panels differed in brightness, rather than color.) Then the animals were tested for their ability to discriminate between various colors. In one series of "testing" trials on a single animal, one of the panels was red and the other two were white; the location of the red panel was varied randomly from trial to trial. In 75 trials, the animal chose correctly 45 times and incorrectly 30 times.[28] How strongly does this support the interpretation that the animal can discriminate between the two colors?

(a) Test the null hypothesis that the animal cannot discriminate red from white. Use a directional alternative and let $\alpha = 0.02$.

(b) Why is a directional alternative appropriate in this case?

9.4.10 Scientists have used Mongolian gerbils when conducting neurological research. A certain breed of these gerbils was crossed and gave progeny of the following colors:[29]

COLOR	BLACK	BROWN	WHITE
Number of progeny	40	59	42

Are these data consistent with the 1:2:1 ratio predicted by a certain genetic model? Use a chi-square test at $\alpha = 0.05$.

9.4.11 Each of 36 men was asked to touch the foreheads of three women, one of whom was their romantic partner, while blindfolded. The two "decoy" women were the same age, height, and weight as the man's partner. Of the 36 men tested, 18 were able to correctly identify their partner.[30] Do the data provide sufficient evidence to conclude that men can do better than they would do by merely guessing?

Conduct an appropriate test.

9.4.12 Geneticists studying the inheritance pattern of cowpea plants classified the plants in one experiment according to the nature of their leaves. The data follow:[31]

TYPE	I	II	III
Number	179	44	23

Test the null hypothesis that the three types occur with probabilities 12/16, 3/16, and 1/16. Use a chi-square test with $\alpha = 0.10$.

9.4.13 In the snapdragon (*Antirrhinum majus*), individual plants can be red flowered, pink flowered, or white flowered. According to a certain Mendelian genetic model, self-pollination of pink-flowered plants should produce progeny that are red, pink, and white in the ratio 1:2:1. A geneticist self-pollinated pink-flowered snapdragon plants and produced 234 progeny with the following colors:[32]

TYPE	RED	PINK	WHITE
Number	54	122	58

Test the null hypothesis that the three colors occur with probabilities 1/4, 1/2, and 1/4. Use a chi-square test with $\alpha = 0.10$.

9.5 Perspective and Summary

In this chapter we have discussed inference for categorical data, including confidence intervals and hypothesis tests. The procedures that we have developed, which are summarized next, can be applied if (1) the data can be regarded as a random sample from a large population and (2) the observations are independent.

Summary of Inference Methods for Categorical Data

95% Confidence interval for *p*

$$\tilde{p} \pm 1.96 \times \mathrm{SE}_{\tilde{p}}$$

where

$$\tilde{p} = \frac{y + 2}{n + 4}$$

and

$$\mathrm{SE}_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + 4}}$$

General confidence interval for *p*

$$\tilde{p} \pm z_{\alpha/2} \times \mathrm{SE}_{\tilde{p}}$$

where

$$\tilde{p} = \frac{y + 0.5(z^2_{\alpha/2})}{n + z^2_{\alpha/2}}$$

$$\mathrm{SE}_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + z^2_{\alpha/2}}}$$

Goodness-of-fit test

Data:

o_i = the observed frequency of category i

Null hypothesis:

H_0 specifies the probability of each category.*

Calculation of expected frequencies:

$e_i = n \times$ Probability specified for category i by H_0

Test statistic:

$$\chi^2_s = \sum_{i=i}^{k} \frac{(o_i - e_i)^2}{e_i}$$

*A slightly modified form of the goodness-of-fit test can be used to test a hypothesis that merely constrains the probabilities rather than specifying them exactly. An example would be testing the fit of a binomial distribution to data (see optional Section 3.9). The details of this test are beyond the scope of this text.

Null distribution (approximate):

$$\chi^2 \text{ distribution with df} = k - 1$$

where k = the number of categories
This approximation is adequate if $e_i \geq 5$ for every category.

Supplementary Exercises 9.S.1–9.S.22

9.S.1 In a certain population, 83% of the people have Rh-positive blood type.[33] Suppose a random sample of $n = 10$ people is to be chosen from the population and let $\tilde{P}$ represent the Wilson-adjusted proportion of Rh-positive people in the sample. Find

(a) $\Pr\{\tilde{P} = 0.714\}$

(b) $\Pr\{\tilde{P} = 0.786\}$

9.S.2 In a population of flatworms (*Planaria*) living in a certain pond, one in five individuals is adult and four are juvenile.[34] An ecologist plans to count the adults in a random sample of 16 flatworms from the pond; she will then use $\tilde{P}$, the Wilson-adjusted sample proportion of adults in the sample, as her estimate of p, the proportion of adults in the pond population. Find

(a) $\Pr\{\tilde{P} = p\}$

(b) $\Pr\{p - 0.05 \leq \tilde{P} \leq p + 0.05\}$

9.S.3 In a study of environmental effects upon reproduction, 123 female adult white-tailed deer from the central Adirondack area were captured and 97 were found to be pregnant.[35] Construct a 95% confidence interval for the proportion of females pregnant in this deer population.

9.S.4 Refer to Exercise 9.S.3. Which of the conditions for validity of the confidence interval might have been violated in this study?

9.S.5 A sample of 32 breastfed infants found that 2 of them developed iron deficiency by age 5.5 months.[36]

(a) Use these data to construct an appropriate 90% confidence interval.

(b) What conditions are necessary for the confidence interval from part (a) to be valid?

(c) Interpret your confidence interval from part (a) in the context of this setting. That is, what do the numbers in the confidence interval tell us about iron deficiency in breastfed infants?

9.S.6 A certain California winery produces 720,000 bottles of wine each year. Suppose you want to estimate the proportion of those bottles that have cork taint (i.e., the wine is spoiled due to a failure of the cork). Suppose that 4% of all corked wine has cork taint. Using this as a preliminary guess of p, how many bottles of wine would need to be included in a random sample if you want the standard error of your estimate to be less than or equal to 1 percentage point?[37]

9.S.7 Refer to Exercise 9.S.6. Suppose you do not trust that the 4% taint rate for wines in general is a useful guess for this particular winery.

(a) Suppose that, based on previous years of data at this winery, about 10% of the wines have had cork taint. How many bottles would need to be included in a random sample if you want the standard error of your estimate to be less than or equal to 1 percentage point?

(b) How many bottles would need to be included in a random sample if you want the standard error of your estimate to be less than or equal to 1 percentage point, no matter what the value of p is?

9.S.8 When male mice are grouped, one of them usually becomes dominant over the others. In order to see how a parasitic infection might affect the competition for dominance, male mice were housed in groups, three mice to a cage; two mice in each cage received a mild dose of the parasitic worm *H. polygyrus.* Two weeks later, criteria such as the relative absence of tail wounds were used to identify the dominant mouse in each cage. It was found that the uninfected mouse had become dominant in 15 of 30 cages.[38] Is this evidence that the parasitic infection tends to inhibit the development of dominant behavior? Use a goodness-of-fit test against a directional alternative. Let $\alpha = 0.05$. (*Hint*: The observational unit in this experiment is not an individual mouse, but a cage of three mice.)

9.S.9 Are mice right-handed or left-handed? In a study of this question, 320 mice of a highly inbred strain were tested for paw preference by observing which forepaw—right or left—they used to retrieve food from a narrow tube. Each animal was tested 50 times, for a total of 320 × 50 = 16,000 observations. The results were as follows:[39]

	RIGHT	LEFT
Number of observations	7,871	8,129

Suppose we assign an expected frequency of 8,000 to each category and perform a goodness-of-fit test; we find that $\chi^2_s = 4.16$, so that at $\alpha = 0.05$ we would reject the hypothesis of a 1:1 ratio and find that there is sufficient evidence to conclude that mice of this strain are (slightly) biased toward use of the left paw. This analysis contains a fatal flaw. What is it?

9.S.10 As part of the study of the inheritance pattern of cowpea plants, geneticists classified the plants in one experiment according to whether the plants had one leaf or three. The data follow:[40]

NUMBER OF LEAVES	1	3
Number of plants	74	61

Test the null hypothesis that the two types of plants occur with equal probabilities. Use a nondirectional alternative and let $\alpha = 0.05$.

9.S.11 People who harvest wild mushrooms sometimes accidentally eat the toxic "death cap" mushroom, *Amanita phalloides.* In reviewing 205 European cases of death-cap poisoning from 1971 through 1980, researchers found that 45 of the victims had died.[41] Conduct a test to compare this mortality to the 30% mortality that was recorded before 1970. Let the alternative hypothesis be that mortality has decreased with time and let $\alpha = 0.05$.

9.S.12 The appearance of leaf pigment glands in the seedling stage of cotton plants is genetically controlled. According to one theory of the control mechanism, the population ratio of glandular to glandless plants resulting from a certain cross should be 11:5; according to another theory it should be 13:3. In one experiment, the cross produced 89 glandular and 36 glandless plants.[42] Use goodness-of-fit tests (at $\alpha = 0.10$) to determine whether these data are consistent with

(a) the 11:5 theory

(b) the 13:3 theory

9.S.13 (Continuation of 9.S.12)

(a) If the 11:5 and 13:3 ratios are the only two reasonable theories to consider, would you have compelling evidence that the theory you selected in Exercise 9.S.12 is the correct theory? Explain.

(b) If there are also other possible theoretical ratios that weren't considered, would you have compelling evidence that the theory you selected in Exercise 9.S.12 is the correct theory? Explain.

9.S.14 When fleeing a predator, the minnow *Fundulus notti* will often head for shore and jump onto the bank. In a study of spatial orientation in this fish, individuals were caught at various locations and later tested in an artificial pool to see which direction they would choose when released: Would they swim in a direction which, at their place of capture, would have led toward shore? The following are the directional choices (±45°) of 50 fish tested under cloudy skies:[43]

Toward shore	18
Away from shore	12
Along shore to the right	13
Along shore to the left	7

Use chi-square tests at $\alpha = .05$ to test the hypothesis that directional choice under cloudy skies is random,

(a) using the four categories listed in the table.

(b) collapsing to two categories—"toward shore" and "away from or along shore"—and using a directional H_A.

(*Note:* Although the chi-square test is valid in this setting, it should be noted that more powerful tests are available for analysis of orientation data.)[44]

9.S.15 Refer to the cortex-weight data of Exercise 8.4.4.

(a) Use a goodness-of-fit test to test the hypothesis that the environmental manipulation has no effect. As in Exercise 8.4.4, use a directional alternative and let $\alpha = 0.05$. (This exercise shows how, by a shift of viewpoint, the sign test can be reinterpreted as a goodness-of-fit test. Of course, the chi-square goodness-of-fit test described in this chapter can be used only if the number of observations is large enough.)

(b) Is the number of observations large enough for the test in part (a) to be valid?

9.S.16 A biologist wanted to know if the cowpea weevil has a preference for one type of bean over others as a place to lay eggs. She put equal amounts of four types of seeds into a jar and added adult cowpea weevils. After a few days she observed the following data:[45]

TYPE OF BEAN	NUMBER OF EGGS
Pinto	167
Cowpea	176
Navy beans	174
Northern beans	194

Do these data provide evidence of a preference for some types of beans over others? That is, are the data consistent with the claim that the eggs are distributed randomly among the four types of bean?

9.S.17 An experiment was conducted in which two types of acorn squash were crossed. According to a genetic model, 1/2 of the resulting plants should have dark stems and dark fruit, 1/4 should have light stems and light fruit, and 1/4 should have light stems and plain fruit. The actual data were 220, 129, and 105 for these three categories.[46] Do these data refute this model? Conduct a chi-square test with $\alpha = 0.10$.

9.S.18 Each of 36 men was asked to touch the backs of the hands of three women, one of whom was the man's romantic partner, while blindfolded. The two "decoy" women were the same age, height, and weight as the man's partner.[30] Of the 36 men tested, 16 were able to correctly identify their partner. Do the data provide sufficient evidence that the men are able to sense their partners better than guessing would predict? Conduct a goodness-of-fit test of the data, using $\alpha = 0.05$.

9.S.19 In a study of resistance to a certain soybean virus, biologists cross fertilized two soybean cultivars. They expected to get a 3:1 ratio of resistant to susceptible plants. The observed data were 58 resistant and 26 susceptible plants.[47] Are these data significantly inconsistent with the expected 3:1 ratio? Conduct a test, using $\alpha = 0.10$; use a nondirectional alternative.

9.S.20 A group of 1,438 sexually active patients were counseled on condom use and the risk of contracting a sexually transmitted disease (STD). After six months, 103 of the patients had new STDs.[48] Construct a 95% confidence interval for the probability of contracting an STD within six months after being part of a counseling program like the one used in this study.

9.S.21 (Continuation of 9.S.20) Suppose that for (uncounseled) sexually active individuals the probability of acquiring an STD in a six-month period is 10%.

(a) Using your interval computed in Exercise 9.S.21, is there compelling evidence that the six-month STD probability is different for those who receive counseling?

(b) Using the data from Exercise 9.S.21, conduct a nondirectional chi-square test to determine if the six-month STD rate is different for counseled individuals compared to the uncounseled population.

(c) Do your answers to parts (a) and (b) agree? Explain.