Categorical Data: One-Sample Distributions

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.

Dichotomous Observations

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: contaminated soda

At any given time, soft-drink dispensers may harbor bacteria such as Chryseobacterium meningosepticum that can cause illness. 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 Wilson-adjusted sample proportion, $\tilde{p}$

The estimate, $\hat{p} = 0.167$, given in the previous example 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, $\tilde{p}$, is another estimate of the population proportion and is given by

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

The Wilson-adjusted sample proportion of contaminated dispensers is

$$\tilde{p} = \frac{5+2}{30+4} = 0.206.$$

- The Wilson-adjusted sample proportion is equivalent to computing the ordinary sample proportion $\hat{p}$ on an augmented sample: one that includes four extra observations of soft-drink dispensers with 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 later, 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 $\tilde{p}$.

Example: 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.

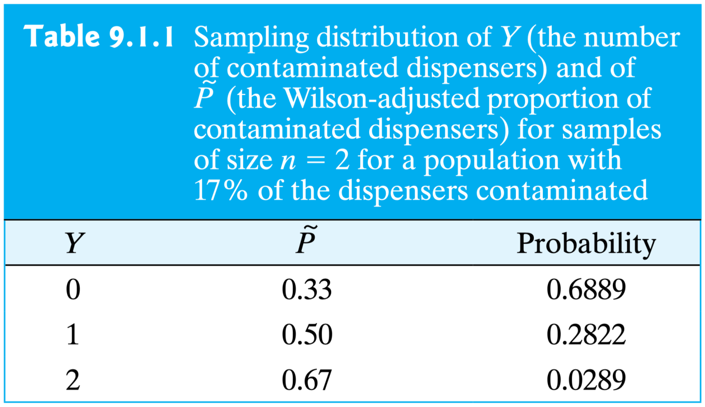


Table 9.1.1 Sampling distribution of Y (the number of contaminated dispensers) and of $\tilde{P}$ (the Wilson-adjusted proportion of contaminated dispensers) for samples of size $n = 2$ for a population with 17% of the dispensers contaminated

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

Example: 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? 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:

$$P(\tilde{p} = \frac{5+2}{20+4}) = P(Y = 5) = \binom{20}{5} 0.17^5 (1 - 0.17)^{20-5} = 0.1345$$

for $Y \sim B(20, 0.17)$. The population proportion p thus corresponds to the second parameter of a binomial distribution $B(n, p)$. The sampling distribution of $\tilde{p}$ is

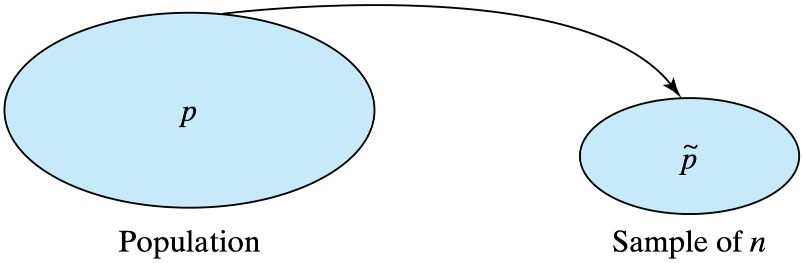
$$P(\tilde{p} = \frac{j+2}{n+4}) = P(Y = j) = \binom{n}{j} p^j (1-p)^{n-j}.$$

for $Y \sim B(n, p)$ and $j = 0, 1, \ldots, n$.

Confidence Interval for a Population Proportion

In Chapter 6 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.

- 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, which is our estimate of p.



When the sample size, n, is large, the sampling distribution of $\tilde{p}$ is **approximately normal** (recall the normal approximation to the binomial distribution); this approximation is related to the Central Limit Theorem.

Standard error of $\tilde{p}$

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

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

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: smoking during pregnancy

In the Pregnancy Risk Assessment Monitoring System survey, 999 women who had given birth were asked about their smoking habits. Smoking during the last 3 months of pregnancy was reported by 125 of those sampled. Thus, $\tilde{p}$ is

$$\frac{125+2}{999+4} = \frac{127}{1003} = 0.127;$$

the standard error is

$$\sqrt{\frac{0.127(1-0.127)}{999+4}} = 0.011.$$

95% confidence interval for p

- Once we have the standard error of $\tilde{p}$, we need to know how likely it is that $\tilde{p}$ will be close to p. The general process of constructing a confidence interval for a proportion is similar to that used in Chapter 6 to construct a confidence interval for a mean.
- When constructing a confidence interval for a mean, we multiplied the standard error by a t multiplier since

$$\frac{\bar{Y}-\mu}{\mathrm{SE}_{\bar{Y}}} \sim t_{n-1}.$$

- The sampling distribution of $\tilde{p}$ is approximately normal if the sample size, n, is large. That is

$$\frac{\tilde{p}-p}{\mathrm{SE}_{\tilde{p}}} \overset{\text{approx}}{\sim} N(0,1).$$

- It turns out that even for moderate or small samples, intervals based on $\tilde{p}$ and Z multipliers do a very good job of estimating the population proportion, p.

For a 95% confidence interval, the appropriate Z multiplier is $z_{0.025} = 1.960$. The approximate 95% confidence interval for a population proportion p is thus

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

For the smoking example, a 95% confidence interval for p is

$$0.127 \pm 1.96 \times 0.011$$

or $(0.105, 0.149)$. Thus, we are 95% confident that the proportion of smoking during the last 3 months of pregnancy is between 0.105 and 0.149 (i.e., between 10.5% and 14.9%).

Conditions for use of the Wilson 95% confidence interval for p

- In order for the Wilson confidence interval to be applicable, it must be reasonable to regard the data as a **random sample** from some population.
- In particular, it is important that the observations are **chosen independently** and that all items in the population have the same chance of being sampled.

- **The Wilson interval does not require large sample sizes to be valid.**

One-sided confidence intervals

Similar to Chapter 6, to construct a one-sided confidence interval, we replace the multiplier $z_{\alpha/2}$ by z_α. Since the population proportion p is always between 0 and 1 and $z_{0.05} = 1.645$, the upper one-side 95% confidence interval for p is

$$(0, \tilde{p} + 1.645 \times \mathrm{SE}_{\tilde{p}}),$$

and the lower one-sided 95% confidence interval for p is

$$(\tilde{p} - 1.645 \times \mathrm{SE}_{\tilde{p}}, 1).$$

Example: 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. 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

$$\sqrt{\frac{0.133(1-0.133)}{11+4}} = 0.088.$$

Thus, a 95% two-sided confidence interval for p is

$$0.133 \pm 1.96 \times 0.088$$

or $(-0.039, 0.305)$. 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%).

A upper one-sided 95% confidence interval for p is

$$(0, 0.133 + 1.645 \times 0.088)$$

or $(0, 0.278)$. That is, we are 95% confident that the probability of death is at most 27.8%.

$1-\alpha$ confidence interval for p

In order to construct intervals with other confidence coefficients, some modifications to the procedure are needed. In general, for a $1-\alpha$ confidence interval, the sample proportion $\tilde{p}$ is defined as

$$\tilde{p} = \frac{Y + 0.5 \times z_{\alpha/2}^2}{n + z_{\alpha/2}^2}$$

while the standard error is given by

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

where $z_{\alpha/2}$ denotes the $100(1-\alpha/2)$ percentile of the standard normal distribution and can be found in t Table at $\mathrm{df} = \infty$ (Recall from Chapter 6 that the t distribution with $\mathrm{df} = \infty$ is a standard normal distribution).

For a 95% confidence interval, $z_{\alpha/2} = z_{0.025} = 1.96$, so

$$\tilde{p} = \frac{Y + 0.5 \times 1.96^2}{n + 1.96^2} = \frac{Y + 1.92}{n + 3.84},$$

which we round off as

$$\frac{Y+2}{n+4}.$$

Similar arguments apply to the standard error for a 95% confidence interval.

In general, the $1-\alpha$ confidence interval for p is

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

the upper one-side $1-\alpha$ confidence interval for p is

$$({\color{red}0}, \tilde{p} + z_{\alpha} \times \mathrm{SE}_{\tilde{p}}),$$

and the lower one-sided $1-\alpha$ confidence interval for p is

$$(\tilde{p} - z_{\alpha} \times \mathrm{SE}_{\tilde{p}}, {\color{red}1}),$$

where

$$\tilde{p} = \frac{Y + 0.5 \times z_{\alpha/2}^2}{n + z_{\alpha/2}^2}, \quad \mathrm{SE}_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n + z_{\alpha/2}^2}}.$$

Vegetarians

In a survey of 136 students at a U.S. college, 19 of them said that they were vegetarians. Let us construct a 90% confidence interval for the proportion, p, of vegetarians in the population. The sample estimate of the proportion is

$$\tilde{p} = \frac{19 + 0.5 \times 1.645^2}{136 + 1.645^2} = \frac{19 + 1.35}{136 + 2.7} = 0.147$$

and the standard error is

$$\mathrm{SE}_{\tilde{p}} = \sqrt{\frac{0.147(1 - 0.147)}{136 + 2.7}} = 0.030.$$

A 90% confidence interval for p is

$$0.147 \pm 1.645 \times 0.030$$

or $(0.098, 0.196)$. Thus, we are 90% confident that between 9.8% and 19.6% of the population that was sampled are vegetarians.

Inference for Proportions: The Chi-square Goodness-of-Fit Test

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 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.

Example: 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 the figure and table below. The null hypothesis is that that deer show no preference to any particular type of burned/unburned habitat (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 (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.

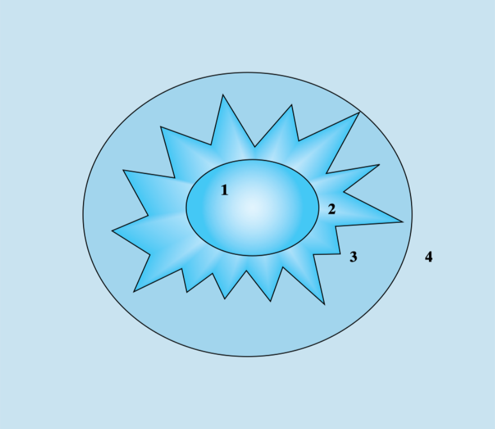


Table 9.4.1 Deer distribution

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

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? One approach is to examine the observed frequencies and compared with the expected frequencies.

Researchers observed a total of 75 deer in the 3,000-acre parcel: 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).

The null and alternative hypotheses are

$$\begin{aligned} H_0 : & P(\text{inner burn}) = 0.173, P(\text{inner edge}) = 0.070, \\ & P(\text{outer edge}) = 0.080, P(\text{outer unburned}) = 0.677. \end{aligned}$$

$$H_A : \text{ At least two of the hypothesized proportions differ from the null.}$$

The chi-square test statistic

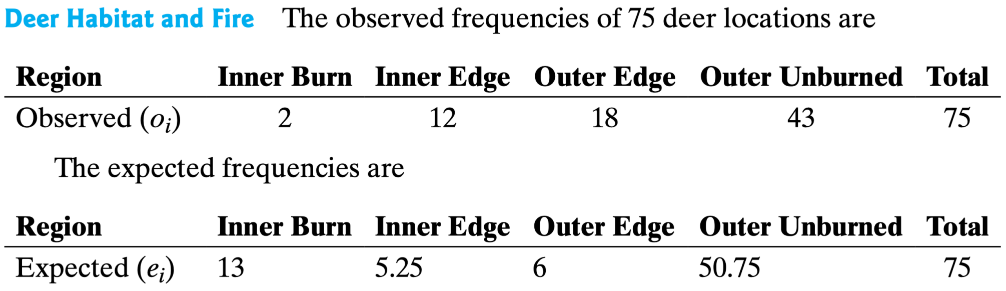
- The goodness-of-fit test is used to assess the compatibility of the data with H_0 that specifies the population proportions, or probabilities, of the various categories. The most widely used goodness-of-fit test is the chi-square test or χ^2 test (χ is the Greek letter "chi").
- For each category i, let o_i represent the observed frequency of the category and let e_i represent the expected frequency (the frequency that would be expected according to H_0).

- The chi-square test statistic is

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

  where the summation is over all k categories.

Consider the deer habitat and fire example, 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: $e_1 = 13$. The corresponding expected frequencies for the other regions are $e_2 = 5.25, e_3 = 6, e_4 = 50.75$.



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

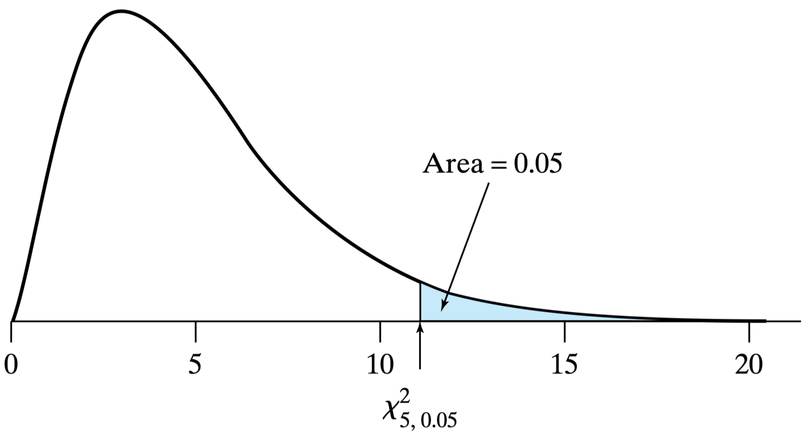
The χ^2 test statistic is

$$\begin{aligned} T &= \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i} \\ &= \frac{(2-13)^2}{13} + \frac{(12-5.25)^2}{5.25} + \frac{(18-6)^2}{6} + \frac{(43-50.75)^2}{50.75} \\ &= 43.2. \end{aligned}$$

The χ^2 distribution

From the way in which T is defined, it is clear that small values of T would indicate that the data agree with H_0, while large values of T would indicate disagreement. To make this idea precise, we need to know the distribution of the test statistic under the null hypothesis H_0.

It can be shown (using the methods of mathematical statistics) that, if **the sample size is large enough ($e_i \geq 5$ for all i)**, then the null distribution of T 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).



χ^2 Table gives critical values for the χ^2 distribution. For instance, for $\mathrm{df} = 5$, 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 the above figure.

The goodness-of-fit test

For the chi-square goodness-of-fit test, the null distribution of T is approximately a χ^2 distribution with $\mathrm{df} = k - 1$, where k equals the number of categories. Specifically,

$$T = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \overset{H_0}{\sim} \chi^2_{k-1}.$$

For example, for the setting presented in deer habitat and fire example 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**.

H_0 is rejected at the α level of significance if

$$p\text{-value } = P(\chi^2_{k-1} > T) < \alpha \text{ or } T > \chi^2_{k-1}(\alpha).$$

For the deer habitat and fire example, the observed χ^2 test statistic was $T = 43.2$. Because there are four categories, the degrees of freedom for the null distribution are calculated as $\mathrm{df} = 4 - 1 = 3$. From χ^2 Table with $\mathrm{df} = 3$ we find that $\chi^2_3(0.0001) = 21.11$. Since $T = 43.2$ is greater than 21.11, the upper tail area beyond 43.2 is less than 0.0001.Thus the p-value is less that 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.

Summary of Inference Methods for Categorical Data

- 95% confidence interval for p:

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

where

$$\tilde{p} = \frac{Y+2}{n+4}, \quad \mathrm{SE}_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}.$$

- General confidence interval for p:

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

  where

  $$\tilde{p} = \frac{Y + 0.5 \times z_{\alpha/2}^2}{n + z_{\alpha/2}^2}, \quad \mathrm{SE}_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n + z_{\alpha/2}^2}}.$$

- Goodness-of-fit test:
  - H_0 specifies the probability of each category
  - H_A: At least two of the hypothesized proportions differ from the null
  - Data: $o_i =$ the observed frequency of category i
  - Calculation of expected frequencies: $e_i = n \times$ probability specified for category i by H_0
  - Test statistic:

    $$T = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \overset{H_0}{\sim} \chi^2_{k-1}$$

  - The approximation is adequate if $e_i \geq 5$ for every category