Chapter 3

Inference for categorical data

Chapter 3 provides a more complete framework for statistical techniques suitable for categorical data. We'll continue working with the normal model in the context of inference for proportions, and we'll also encounter a new technique and distribution suitable for working with frequency and contingency tables in Sections 3.3 and 3.4.

3.1 Inference for a single proportion

Before we get started, we'll introduce a little terminology and notation.

In the tappers-listeners study, one person tapped a tune on the table and the listener tried to guess the game. In this study, each game can be thought of as a **trial**. We could label each trial a **success** if the listener successfully guessed the tune, and we could label a trial a **failure** if the listener was unsuccessful.

Trial, success, and failure

A single event that leads to an outcome can be called a *trial*. If the trial has two possible outcomes, e.g. heads or tails when flipping a coin, we typically label one of those outcome a *success* and the other a *failure*. The choice of which outcome is labeled a success and which a failure is arbitrary, and it will not impact the results of our analyses.

When a proportion is recorded, it is common to use a 1 to represent a "success" and a 0 to represent a "failure" and then write down a **key** to communicate what each value represents. This notation is also convenient for calculations. For example, if we have 10 trials with 6 success (1's) and 4 failures (0's), the sample proportion can be computed using the mean of the zeros and ones:

$$\hat{p} = \frac{1+1+1+1+1+1+0+0+0+0}{10} = 0.6$$

Next we'll take a look at when we can apply our normal distribution framework to the distribution of the sample proportion, $\hat{p}$.

3.1.1 When the sample proportion is nearly normal

$\hat{p}$
sample proportion

p
population proportion

Conditions for when the sampling distribution of $\hat{p}$ is nearly normal

The sampling distribution for $\hat{p}$, taken from a sample of size n from a population with a true proportion p, is nearly normal when

1. the sample observations are independent and
2. we expected to see at least 10 successes and 10 failures in our sample, i.e. $np \geq 10$ and $n(1-p) \geq 10$. This is called the **success-failure condition**.

If these conditions are met, then the sampling distribution of $\hat{p}$ is nearly normal with mean p and standard error

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{3.1}$$

Typically we do not know the true proportion, p, so we substitute some value to check conditions and to estimate the standard error. For confidence intervals, usually $\hat{p}$ is used to check the success-failure condition and compute the standard error. For hypothesis tests, typically the null value p_0 is used in place of p. Examples are presented for each of these cases in Sections 3.1.2 and 3.1.3.

TIP: Reminder on checking independence of observations

If data come from a simple random sample and consist of less than 10% of the population, then the independence assumption is reasonable. Or, for example, if the data come from an experiment where each user was randomly assigned to the treatment or control group and users do not interact, then the observations in each group are typically independent.

3.1.2 Confidence intervals for a proportion

According to a New York Times / CBS News poll in June 2012, only about 44% of the American public approves of the job the Supreme Court is doing.[1] This poll included responses of 976 randomly sampled adults.

We want a confidence interval for the proportion of Americans who approve of the job the Supreme Court is doing. Our point estimate, based on a simple random sample of size $n = 976$ from the NYTimes/CBS poll, is $\hat{p} = 0.44$. To use our confidence interval formula from Section 2.8, we must first check whether the sampling distribution of $\hat{p}$ is nearly normal and calculate the standard error of the estimate.

The data are based on a simple random sample and consist of far fewer than 10% of the U.S. population, so independence is confirmed. The sample size must also be sufficiently large, which is checked via the success-failure condition: there were approximately $976\times\hat{p} = 429$ "successes" and $976\times(1-\hat{p}) = 547$ "failures" in the sample, both easily greater than 10.

With the conditions met, we are assured that the sampling distribution of $\hat{p}$ is nearly normal. Next, a standard error for $\hat{p}$ is needed, and then we can employ the usual method to construct a confidence interval.

[1] nytimes.com/2012/06/08/us/politics/44-percent-of-americans-approve-of-supreme-court-in-new-poll.html

⊙ **Guided Practice 3.2** Estimate the standard error of $\hat{p} = 0.44$ using Equation (3.1). Because p is unknown and the standard error is for a confidence interval, use $\hat{p}$ in place of p. [2]

● **Example 3.3** Construct a 95% confidence interval for p, the proportion of Americans who approve of the job the Supreme Court is doing.

Using the standard error estimate from Guided Practice 3.2, the point estimate 0.44, and $z^\star = 1.96$ for a 95% confidence interval, the confidence interval can be computed as

$$\text{point estimate} \ \pm \ z^\star SE \quad \to \quad 0.44 \ \pm \ 1.96 \times 0.016 \quad \to \quad (0.409, 0.471)$$

We are 95% confident that the true proportion of Americans who approve of the job of the Supreme Court (in June 2012) is between 0.409 and 0.471. At the time this poll was taken, we can say with high confidence that the job approval of the Supreme Court was below 50%.

Constructing a confidence interval for a proportion

- Verify the observations are independent and also verify the success-failure condition using $\hat{p}$ and n.
- If the conditions are met, then the Central Limit Theorem applies, and the sampling distribution of $\hat{p}$ is well-approximated by the normal model.
- Construct the standard error using $\hat{p}$ in place of p and apply the general confidence interval formula.

3.1.3 Hypothesis testing for a proportion

To apply the same normal distribution framework in the context of a hypothesis test for a proportion, the independence and success-failure conditions must also be satisfied. However, in a hypothesis test, the success-failure condition is checked using the null proportion: we verify np_0 and $n(1 - p_0)$ are at least 10, where p_0 is the null value.

⊙ **Guided Practice 3.4** Deborah Toohey is running for Congress, and her campaign manager claims she has more than 50% support from the district's electorate. Ms. Toohey's opponent claimed that Ms. Toohey has *less* than 50%. Set up a hypothesis test to evaluate who is right.[3]

[2] $SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.44(1-0.44)}{976}} = 0.016$

[3] We should run a two-sided.H_0: Ms. Toohey's support is 50%. $p = 0.50$. H_A: Ms. Toohey's support is either above or below 50%. $p \neq 0.50$.

● **Example 3.5** A newspaper collects a simple random sample of 500 likely voters in the district and estimates Toohey's support to be 52%. Does this provide convincing evidence for the claim of Toohey's manager at the 5% significance level?

Because this is a simple random sample that includes fewer than 10% of the population, the observations are independent. In a one-proportion hypothesis test, the success-failure condition is checked using the null proportion, $p_0 = 0.5$: $np_0 = n(1 - p_0) = 500 \times 0.5 = 250 > 10$. With these conditions verified, the normal model may be applied to $\hat{p}$.

Next the standard error can be computed. The null value is used again here, because this is a hypothesis test for a single proportion.

$$SE = \sqrt{\frac{p_0 \times (1 - p_0)}{n}} = \sqrt{\frac{0.5 \times (1 - 0.5)}{500}} = 0.022$$

A picture of the normal model is shown in Figure 3.1 with the p-value represented by both shaded tails. Based on the normal model, we can compute a test statistic as the Z score of the point estimate:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.52 - 0.50}{0.022} = 0.89$$

The right tail area is 0.1867, and the p-value is 2 × 0.1867 = 0.3734. Because the p-value is larger than 0.05, we do not reject the null hypothesis, and we do not find convincing evidence to support the campaign manager's claim.

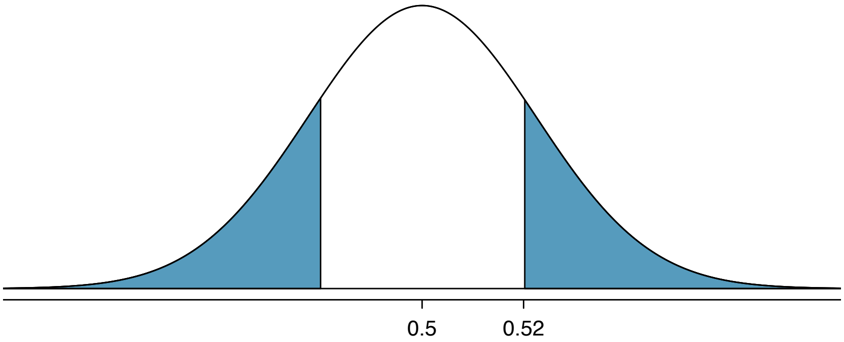

Figure 3.1: Sampling distribution of the sample proportion if the null hypothesis is true for Example 3.5. The p-value for the test is shaded.

Hypothesis test for a proportion
Set up hypotheses and verify the conditions using the null value, p_0, to ensure $\hat{p}$ is nearly normal under H_0. If the conditions hold, construct the standard error, again using p_0, and show the p-value in a drawing. Lastly, compute the p-value and evaluate the hypotheses.

3.1.4 Choosing a sample size when estimating a proportion

Frequently statisticians find themselves in a position to not only analyze data, but to help others determine how to most effectively collect data and also how much data should be collected. We can perform sample size calculations that are helpful in planning a study. Our task will be to identify an appropriate sample size that ensures the margin of error $ME = z^\star SE$ will be no larger than some value m. For example, we might be asked to find a sample size so the margin of error is no larger than $m = 0.04$, in which case, we write

$$z^\star SE \leq 0.04$$

Generally, we plug in a suitable value for $z^\star$ for the confidence level we plan to use, write in the formula for the standard error, and then solve for the sample size n. In the case of a single proportion, we use $\sqrt{p(1-p)/n}$ for the standard error (SE).

● **Example 3.6** If we are conducting a university survey to determine whether students support a $200 per year increase in fees to pay for a new football stadium, how big of a sample is needed to ensure the margin of error is less than 0.04 using a 95% confidence level?

For a 95% confidence level, the value $z^\star$ corresponds to 1.96, and we can write the margin of error expression as follows:

$$ME = z^\star SE = 1.96 \times \sqrt{\frac{p(1-p)}{n}} \leq 0.04$$

There are two unknowns in the equation: p and n. If we have an estimate of p, perhaps from a similar survey, we could use that value. If we have no such estimate, we must use some other value for p. The margin of error for a proportion is largest when p is 0.5, so we typically use this *worst case estimate* if no other estimate is available:

$$\begin{aligned} 1.96 \times \sqrt{\frac{0.5(1-0.5)}{n}} &\leq 0.04 \\ 1.96^2 \times \frac{0.5(1-0.5)}{n} &\leq 0.04^2 \\ 1.96^2 \times \frac{0.5(1-0.5)}{0.04^2} &\leq n \\ 600.25 &\leq n \end{aligned}$$

We would need at least 600.25 participants, which means we need 601 participants or more, to ensure the sample proportion is within 0.04 of the true proportion with 95% confidence. Notice that in such calculations, we always round up for the sample size!

As noted in the example, if we have an estimate of the proportion, we should use it in place of the worst case estimate of the proportion, 0.5.

⊙ **Guided Practice 3.7** A manager is about to oversee the mass production of a new tire model in her factory, and she would like to estimate what proportion of these tires will be rejected through quality control. The quality control team has monitored the last three tire models produced by the factory, failing 1.7% of tires in the first model, 6.2% of the second model, and 1.3% of the third model. The manager would like to examine enough tires to estimate the failure rate of the new tire model to within about 2% with a 90% confidence level.[4]

(a) There are three different failure rates to choose from. Perform the sample size computation for each separately, and identify three sample sizes to consider.

(b) The sample sizes vary widely. Which of the three would you suggest using? What would influence your choice?

⊙ **Guided Practice 3.8** A recent estimate of Congress' approval rating was 17%.[5] If we were to conduct a new poll and wanted an estimate with a margin of error smaller than about 0.04 with 95% confidence, how big of a sample should we use?[6]

3.2 Difference of two proportions

We would like to make conclusions about the difference in two population proportions ($p_1 - p_2$) using the normal model. In this section we consider three such examples. In the first, we compare the approval of the 2010 healthcare law under two different question phrasings. In the second application, a company weighs whether they should switch to a higher quality parts manufacturer. In the last example, we examine the cancer risk to dogs from the use of yard herbicides.

In our investigations, we first identify a reasonable point estimate of $p_1 - p_2$ based on the sample. You may have already guessed its form: $\hat{p}_1 - \hat{p}_2$. Next, in each example we verify that the point estimate follows the normal model by checking certain conditions; as before, these conditions relate to independence of observations and checking for sufficiently large sample size. Finally, we compute the estimate's standard error and apply our inferential framework.

[4] (a) For the 1.7% estimate of p, we estimate the appropriate sample size as follows:

$$1.65 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.65 \times \sqrt{\frac{0.017(1-0.017)}{n}} \le 0.02 \qquad \to \qquad n \ge 113.7$$

Using the estimate from the first model, we would suggest examining 114 tires (round up!). A similar computation can be accomplished using 0.062 and 0.013 for p: 396 and 88.

(b) We could examine which of the old models is most like the new model, then choose the corresponding sample size. Or if two of the previous estimates are based on small samples while the other is based on a larger sample, we should consider the value corresponding to the larger sample. (Answers will vary.)

[5] www.gallup.com/poll/155144/Congress-Approval-June.aspx

[6] We complete the same computations as before, except now we use 0.17 instead of 0.5 for p:

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.96 \times \sqrt{\frac{0.17(1-0.17)}{n}} \le 0.04 \qquad \to \qquad n \ge 338.8$$

A sample size of 339 or more would be reasonable.