Chapter 6

Inference for categorical data

Chapter 6 introduces inference in the setting of categorical data. We use these methods to answer questions like the following:

- What proportion of the American public approves of the job the Supreme Court is doing?
- The Pew Research Center conducted a poll about support for the 2010 health care law, and they used two forms of the survey question. Each respondent was randomly given one of the two questions. What is the difference in the support for respondents under the two question orderings?

We will find that the methods we learned in previous chapters are very useful in these settings. For example, sample proportions are well characterized by a nearly normal distribution when certain conditions are satisfied, making it possible to employ the usual confidence interval and hypothesis testing tools. In other instances, such as those with contingency tables or when sample size conditions are not met, we will use a different distribution, though the core ideas remain the same.

6.1 Inference for a single 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 adults.

6.1.1 Identifying when the sample proportion is nearly normal

A sample proportion can be described as a sample mean. If we represent each "success" as a 1 and each "failure" as a 0, then the sample proportion is the mean of these numerical outcomes:

$$\hat{p} = \frac{0 + 1 + 1 + \cdots + 0}{976} = 0.44$$

The distribution of $\hat{p}$ is nearly normal when the distribution of 0's and 1's is not too strongly skewed for the sample size. The most common guideline for sample size and skew when

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

working with proportions is to ensure that we expect to observe a minimum number of successes and failures, typically at least 10 of each.

$\hat{p}$
sample proportion

p
population proportion

Conditions for the sampling distribution of $\hat{p}$ being 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

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{6.1}$$

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 – that is, the proportion claimed in the null hypothesis – is used in place of p. Examples are presented for each of these cases in Sections 6.1.2 and 6.1.3.

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. Alternatively, if the data come from a random process, we must evaluate the independence condition more carefully.

6.1.2 Confidence intervals for a proportion

We may 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 sample of size $n = 976$ from the NYTimes/CBS poll, is $\hat{p} = 0.44$. To use the general confidence interval formula from Section 4.5, we must check the conditions to ensure that the sampling distribution of $\hat{p}$ is nearly normal. We also must determine 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.

⊙ **Exercise 6.2** Estimate the standard error of $\hat{p} = 0.44$ using Equation (6.1). Because p is unknown and the standard error is for a confidence interval, use $\hat{p}$ in place of p.[2]

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

● **Example 6.3** 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 Exercise 6.2, the point estimate 0.44, and $z^\star = 1.96$ for a 95% confidence interval, the confidence interval may be computed as

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

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. If the proportion has not changed since this poll, than we can say with high confidence that the job approval of the Supreme Court is 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, the sampling distribution of $\hat{p}$ may be well-approximated by the normal model.
- Construct the standard error using $\hat{p}$ in place of p and apply the general confidence interval formula.

6.1.3 Hypothesis testing for a proportion

To apply the normal distribution framework in the context of a hypothesis test for a proportion, the independence and success-failure conditions must be satisfied. 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.

⊙ **Exercise 6.4** Deborah Toohey is running for Congress, and her campaign manager claims she has more than 50% support from the district's electorate. Set up a one-sided hypothesis test to evaluate this claim.[3]

● **Example 6.5** 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.

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

[3] Is there convincing evidence that the campaign manager is correct? $H_0 : p = 0.50$, $H_A : p > 0.50$.

A picture of the normal model is shown in Figure 6.1 with the p-value represented by the shaded region. Based on the normal model, the test statistic can be computed 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 upper tail area, representing the p-value, is 0.1867. 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.

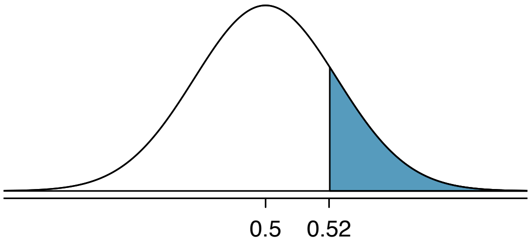


Figure 6.1: Sampling distribution of the sample proportion if the null hypothesis is true for Example 6.5. The p-value for the test is shaded.

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.

6.1.4 Choosing a sample size when estimating a proportion

We first encountered sample size computations in Section 4.6, which considered the case of estimating a single mean. We found that these computations were helpful in planning a study to control the size of the standard error of a point estimate. The task was to find a sample size n so that the sample mean would be within some margin of error m of the actual mean with a certain level of confidence. For example, the margin of error for a point estimate using 95% confidence can be written as $1.96 \times SE$. We set up a general equation to represent the problem:

$$ME = z^{\star} SE \leq m$$

where ME represented the actual margin of error and $z^{\star}$ was chosen to correspond to the confidence level. The standard error formula is specified to correspond to the particular setting. For instance, in the case of means, the standard error was given as $\sigma/\sqrt{n}$. In the case of a single proportion, we use $\sqrt{p(1-p)/n}$ for the standard error.

Planning a sample size before collecting data is equally important when estimating a proportion. For instance, 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 be sure the margin of error is less than 0.04 using a 95% confidence level?

● **Example 6.6** Find the smallest sample size n so that the margin of error of the point estimate $\hat{p}$ will be no larger than $m = 0.04$ when using a 95% confidence interval.

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

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

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. It turns out that the margin of error is largest when p is 0.5, so we typically use this *worst case estimate* if no other estimate is available:

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

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.

No estimate of the true proportion is required in sample size computations for a proportion, whereas an estimate of the standard deviation is always needed when computing a sample size for a margin of error for the sample mean. However, if we have an estimate of the proportion, we should use it in place of the worst case estimate of the proportion, 0.5.

⊙ **Exercise 6.7** 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?

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

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

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.)

⊙ **Exercise 6.8** A recent estimate of Congress' approval rating was 17%.[5] What sample size does this estimate suggest we should use for a margin of error of 0.04 with 95% confidence?[6]

6.2 Difference of two proportions

We would like to make conclusions about the difference in two population proportions: $p_1 - p_2$. We consider three 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. Finally, we compute the estimate's standard error and apply our inferential framework.

6.2.1 Sample distribution of the difference of two proportions

We must check two conditions before applying the normal model to $\hat{p}_1 - \hat{p}_2$. First, the sampling distribution for each sample proportion must be nearly normal, and secondly, the samples must be independent. Under these two conditions, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ may be well approximated using the normal model.

Conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to be normal

The difference $\hat{p}_1 - \hat{p}_2$ tends to follow a normal model when

- each proportion separately follows a normal model, and
- the two samples are independent of each other.

The standard error of the difference in sample proportions is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \qquad (6.9)$$

where p_1 and p_2 represent the population proportions, and n_1 and n_2 represent the sample sizes.

For the difference in two means, the standard error formula took the following form:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2}$$

The standard error for the difference in two proportions takes a similar form. The reasons behind this similarity are rooted in the probability theory of Section 2.4, which is described for this context in Exercise 5.14 on page 221.

[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:

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

A sample size of 339 or more would be reasonable.

	Sample size (n_i)	Approve law (%)	Disapprove law (%)	Other
"people who cannot afford it will receive financial help from the government" is given second	771	47	49	3
"people who do not buy it will pay a penalty" is given second	732	34	63	3

Table 6.2: Results for a Pew Research Center poll where the ordering of two statements in a question regarding healthcare were randomized.

6.2.2 Intervals and tests for $p_1 - p_2$

In the setting of confidence intervals, the sample proportions are used to verify the success-failure condition and also compute standard error, just as was the case with a single proportion.

● **Example 6.10** The way a question is phrased can influence a person's response. For example, Pew Research Center conducted a survey with the following question:[7]

> As you may know, by 2014 nearly all Americans will be required to have health insurance. [People who do not buy insurance will pay a penalty] while [People who cannot afford it will receive financial help from the government]. Do you approve or disapprove of this policy?

For each randomly sampled respondent, the statements in brackets were randomized: either they were kept in the order given above, or the two statements were reversed. Table 6.2 shows the results of this experiment. Create and interpret a 90% confidence interval of the difference in approval.

First the conditions must be verified. Because each group is a simple random sample from less than 10% of the population, the observations are independent, both within the samples and between the samples. The success-failure condition also holds for each sample. Because all conditions are met, the normal model can be used for the point estimate of the difference in support, where p_1 corresponds to the original ordering and p_2 to the reversed ordering:

$$\hat{p}_1 - \hat{p}_2 = 0.47 - 0.34 = 0.13$$

The standard error may be computed from Equation (6.9) using the sample proportions:

$$SE \approx \sqrt{\frac{0.47(1-0.47)}{771} + \frac{0.34(1-0.34)}{732}} = 0.025$$

For a 90% confidence interval, we use $z^\star = 1.65$:

$$\text{point estimate} \ \pm\ z^\star SE \quad \to \quad 0.13 \ \pm\ 1.65 \times 0.025 \quad \to \quad (0.09, 0.17)$$

We are 90% confident that the approval rating for the 2010 healthcare law changes between 9% and 17% due to the ordering of the two statements in the survey question. The Pew Research Center reported that this modestly large difference suggests that the opinions of much of the public are still fluid on the health insurance mandate.

[7] www.people-press.org/2012/03/26/public-remains-split-on-health-care-bill-opposed-to-mandate/. Sample sizes for each polling group are approximate.

⊙ **Exercise 6.11** A remote control car company is considering a new manufacturer for wheel gears. The new manufacturer would be more expensive but their higher quality gears are more reliable, resulting in happier customers and fewer warranty claims. However, management must be convinced that the more expensive gears are worth the conversion before they approve the switch. If there is strong evidence of a more than 3% improvement in the percent of gears that pass inspection, management says they will switch suppliers, otherwise they will maintain the current supplier. Set up appropriate hypotheses for the test.[8]

● **Example 6.12** The quality control engineer from Exercise 6.11 collects a sample of gears, examining 1000 gears from each company and finds that 899 gears pass inspection from the current supplier and 958 pass inspection from the prospective supplier. Using these data, evaluate the hypothesis setup of Exercise 6.11 using a significance level of 5%.

First, we check the conditions. The sample is not necessarily random, so to proceed we must assume the gears are all independent; for this sample we will suppose this assumption is reasonable, but the engineer would be more knowledgeable as to whether this assumption is appropriate. The success-failure condition also holds for each sample. Thus, the difference in sample proportions, $0.958 - 0.899 = 0.059$, can be said to come from a nearly normal distribution.

The standard error can be found using Equation (6.9):

$$SE = \sqrt{\frac{0.958(1-0.958)}{1000} + \frac{0.899(1-0.899)}{1000}} = 0.0114$$

In this hypothesis test, the sample proportions were used. We will discuss this choice more in Section 6.2.3.

Next, we compute the test statistic and use it to find the p-value, which is depicted in Figure 6.3.

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.059 - 0.03}{0.0114} = 2.54$$

Using the normal model for this test statistic, we identify the right tail area as 0.006. Since this is a one-sided test, this single tail area is also the p-value, and we reject the null hypothesis because 0.006 is less than 0.05. That is, we have statistically significant evidence that the higher quality gears actually do pass inspection more than 3% as often as the currently used gears. Based on these results, management will approve the switch to the new supplier.

[8] H_0: The higher quality gears will pass inspection no more than 3% more frequently than the standard quality gears. $p_{highQ} - p_{standard} = 0.03$. H_A: The higher quality gears will pass inspection more than 3% more often than the standard quality gears. $p_{highQ} - p_{standard} > 0.03$.

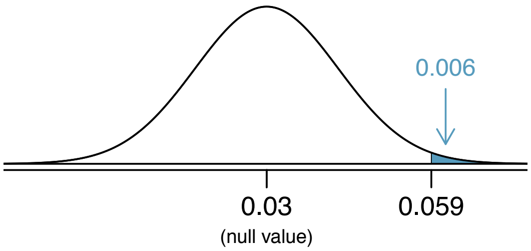


Figure 6.3: Distribution of the test statistic if the null hypothesis was true. The p-value is represented by the shaded area.

6.2.3 Hypothesis testing when $H_0 : p_1 = p_2$

Here we use a new example to examine a special estimate of standard error when $H_0 : p_1 = p_2$. We investigate whether there is an increased risk of cancer in dogs that are exposed to the herbicide 2,4-dichlorophenoxyacetic acid (2,4-D). A study in 1994 examined 491 dogs that had developed cancer and 945 dogs as a control group.[9] Of these two groups, researchers identified which dogs had been exposed to 2,4-D in their owner's yard. The results are shown in Table 6.4.

	cancer	no cancer
2,4-D	191	304
no 2,4-D	300	641

Table 6.4: Summary results for cancer in dogs and the use of 2,4-D by the dog's owner.

⊙ **Exercise 6.13** Is this study an experiment or an observational study?[10]

⊙ **Exercise 6.14** Set up hypotheses to test whether 2,4-D and the occurrence of cancer in dogs are related. Use a one-sided test and compare across the cancer and no cancer groups.[11]

[9] Hayes HM, Tarone RE, Cantor KP, Jessen CR, McCurnin DM, and Richardson RC. 1991. Case-Control Study of Canine Malignant Lymphoma: Positive Association With Dog Owner's Use of 2, 4-Dichlorophenoxyacetic Acid Herbicides. Journal of the National Cancer Institute 83(17):1226-1231.

[10] The owners were not instructed to apply or not apply the herbicide, so this is an observational study. This question was especially tricky because one group was called the *control group*, which is a term usually seen in experiments.

[11] Using the proportions within the cancer and no cancer groups may seem odd. We intuitively may desire to compare the fraction of dogs with cancer in the 2,4-D and no 2,4-D groups, since the herbicide is an explanatory variable. However, the cancer rates in each group do not necessarily reflect the cancer rates in reality due to the way the data were collected. For this reason, computing cancer rates may greatly alarm dog owners.
H_0: the proportion of dogs with exposure to 2,4-D is the same in "cancer" and "no cancer" dogs, $p_c - p_n = 0$.
H_A: dogs with cancer are more likely to have been exposed to 2,4-D than dogs without cancer, $p_c - p_n > 0$.

● **Example 6.15** Are the conditions met to use the normal model and make inference on the results?

(1) It is unclear whether this is a random sample. However, if we believe the dogs in both the cancer and no cancer groups are representative of each respective population and that the dogs in the study do not interact in any way, then we may find it reasonable to assume independence between observations. (2) The success-failure condition holds for each sample.

Under the assumption of independence, we can use the normal model and make statements regarding the canine population based on the data.

In your hypotheses for Exercise 6.14, the null is that the proportion of dogs with exposure to 2,4-D is the same in each group. The point estimate of the difference in sample proportions is $\hat{p}_c - \hat{p}_n = 0.067$. To identify the p-value for this test, we first check conditions (Example 6.15) and compute the standard error of the difference:

$$SE = \sqrt{\frac{p_c(1-p_c)}{n_c} + \frac{p_n(1-p_n)}{n_n}}$$

In a hypothesis test, the distribution of the test statistic is always examined as though the null hypothesis is true, i.e. in this case, $p_c = p_n$. The standard error formula should reflect this equality in the null hypothesis. We will use p to represent the common rate of dogs that are exposed to 2,4-D in the two groups:

$$SE = \sqrt{\frac{p(1-p)}{n_c} + \frac{p(1-p)}{n_n}}$$

We don't know the exposure rate, p, but we can obtain a good estimate of it by *pooling* the results of both samples:

$$\hat{p} = \frac{\text{\# of ``successes''}}{\text{\# of cases}} = \frac{191 + 304}{191 + 300 + 304 + 641} = 0.345$$

This is called the **pooled estimate** of the sample proportion, and we use it to compute the standard error when the null hypothesis is that $p_1 = p_2$ (e.g. $p_c = p_n$ or $p_c - p_n = 0$). We also typically use it to verify the success-failure condition.

Pooled estimate of a proportion

When the null hypothesis is $p_1 = p_2$, it is useful to find the pooled estimate of the shared proportion:

$$\hat{p} = \frac{\text{number of ``successes''}}{\text{number of cases}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

Here $\hat{p}_1 n_1$ represents the number of successes in sample 1 since

$$\hat{p}_1 = \frac{\text{number of successes in sample 1}}{n_1}$$

Similarly, $\hat{p}_2 n_2$ represents the number of successes in sample 2.

TIP: Use the pooled proportion estimate when $\mathbf{H_0 : p_1 = p_2}$
When the null hypothesis suggests the proportions are equal, we use the pooled proportion estimate ($\hat{p}$) to verify the success-failure condition and also to estimate the standard error:

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}} \qquad (6.16)$$

⊙ **Exercise 6.17** Using Equation (6.16), $\hat{p} = 0.345$, $n_1 = 491$, and $n_2 = 945$, verify the estimate for the standard error is $SE = 0.026$. Next, complete the hypothesis test using a significance level of 0.05. Be certain to draw a picture, compute the p-value, and state your conclusion in both statistical language and plain language.[12]

6.3 Testing for goodness of fit using chi-square (special topic)

In this section, we develop a method for assessing a null model when the data are binned. This technique is commonly used in two circumstances:

- Given a sample of cases that can be classified into several groups, determine if the sample is representative of the general population.
- Evaluate whether data resemble a particular distribution, such as a normal distribution or a geometric distribution.

Each of these scenarios can be addressed using the same statistical test: a chi-square test.

In the first case, we consider data from a random sample of 275 jurors in a small county. Jurors identified their racial group, as shown in Table 6.5, and we would like to determine if these jurors are racially representative of the population. If the jury is representative of the population, then the proportions in the sample should roughly reflect the population of eligible jurors, i.e. registered voters.

Race	White	Black	Hispanic	Other	Total
Representation in juries	205	26	25	19	275
Registered voters	0.72	0.07	0.12	0.09	1.00

Table 6.5: Representation by race in a city's juries and population.

While the proportions in the juries do not precisely represent the population proportions, it is unclear whether these data provide convincing evidence that the sample is not representative. If the jurors really were randomly sampled from the registered voters, we might expect small differences due to chance. However, unusually large differences may provide convincing evidence that the juries were not representative.

[12] Compute the test statistic:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.067 - 0}{0.026} = 2.58$$

We leave the picture to you. Looking up $Z = 2.58$ in the normal probability table: 0.9951. However this is the lower tail, and the upper tail represents the p-value: $1 - 0.9951 = 0.0049$. We reject the null hypothesis and conclude that dogs getting cancer and owners using 2,4-D are associated.

A second application, assessing the fit of a distribution, is presented at the end of this section. Daily stock returns from the S&P500 for the years 1990-2011 are used to assess whether stock activity each day is independent of the stock's behavior on previous days.

In these problems, we would like to examine all bins simultaneously, not simply compare one or two bins at a time, which will require us to develop a new test statistic.

6.3.1 Creating a test statistic for one-way tables

● **Example 6.18** Of the people in the city, 275 served on a jury. If the individuals are randomly selected to serve on a jury, about how many of the 275 people would we expect to be white? How many would we expect to be black?

About 72% of the population is white, so we would expect about 72% of the jurors to be white: $0.72 \times 275 = 198$.

Similarly, we would expect about 7% of the jurors to be black, which would correspond to about $0.07 \times 275 = 19.25$ black jurors.

⊙ **Exercise 6.19** Twelve percent of the population is Hispanic and 9% represent other races. How many of the 275 jurors would we expect to be Hispanic or from another race? Answers can be found in Table 6.6.

Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected counts	198	19.25	33	24.75	275

Table 6.6: Actual and expected make-up of the jurors.

The sample proportion represented from each race among the 275 jurors was not a precise match for any ethnic group. While some sampling variation is expected, we would expect the sample proportions to be fairly similar to the population proportions if there is no bias on juries. We need to test whether the differences are strong enough to provide convincing evidence that the jurors are not a random sample. These ideas can be organized into hypotheses:

H_0: The jurors are a random sample, i.e. there is no racial bias in who serves on a jury, and the observed counts reflect natural sampling fluctuation.

H_A: The jurors are not randomly sampled, i.e. there is racial bias in juror selection.

To evaluate these hypotheses, we quantify how different the observed counts are from the expected counts. Strong evidence for the alternative hypothesis would come in the form of unusually large deviations in the groups from what would be expected based on sampling variation alone.

6.3.2 The chi-square test statistic

In previous hypothesis tests, we constructed a test statistic of the following form:

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

This construction was based on (1) identifying the difference between a point estimate and an expected value if the null hypothesis was true, and (2) standardizing that difference using the standard error of the point estimate. These two ideas will help in the construction of an appropriate test statistic for count data.

Our strategy will be to first compute the difference between the observed counts and the counts we would expect if the null hypothesis was true, then we will standardize the difference:

$$Z_1 = \frac{\text{observed white count} - \text{null white count}}{\text{SE of observed white count}}$$

The standard error for the point estimate of the count in binned data is the square root of the count under the null.[13] Therefore:

$$Z_1 = \frac{205 - 198}{\sqrt{198}} = 0.50$$

The fraction is very similar to previous test statistics: first compute a difference, then standardize it. These computations should also be completed for the black, Hispanic, and other groups:

Black

$$Z_2 = \frac{26 - 19.25}{\sqrt{19.25}} = 1.54$$

Hispanic

$$Z_3 = \frac{25 - 33}{\sqrt{33}} = -1.39$$

Other

$$Z_4 = \frac{19 - 24.75}{\sqrt{24.75}} = -1.16$$

We would like to use a single test statistic to determine if these four standardized differences are irregularly far from zero. That is, Z_1, Z_2, Z_3, and Z_4 must be combined somehow to help determine if they – as a group – tend to be unusually far from zero. A first thought might be to take the absolute value of these four standardized differences and add them up:

$$|Z_1| + |Z_2| + |Z_3| + |Z_4| = 4.58$$

Indeed, this does give one number summarizing how far the actual counts are from what was expected. However, it is more common to add the squared values:

$$Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 = 5.89$$

Squaring each standardized difference before adding them together does two things:

- Any standardized difference that is squared will now be positive.
- Differences that already look unusual – e.g. a standardized difference of 2.5 – will become much larger after being squared.

The test statistic X^2, which is the sum of the Z^2 values, is generally used for these reasons. We can also write an equation for X^2 using the observed counts and null counts:

$$X^2 = \frac{(\text{observed count}_1 - \text{null count}_1)^2}{\text{null count}_1} + \cdots + \frac{(\text{observed count}_4 - \text{null count}_4)^2}{\text{null count}_4}$$

X^2
chi-square
test statistic

[13] Using some of the rules learned in earlier chapters, we might think that the standard error would be $np(1-p)$, where n is the sample size and p is the proportion in the population. This would be correct if we were looking only at one count. However, we are computing many standardized differences and adding them together. It can be shown – though not here – that the square root of the count is a better way to standardize the count differences.

The final number X^2 summarizes how strongly the observed counts tend to deviate from the null counts. In Section 6.3.4, we will see that if the null hypothesis is true, then X^2 follows a new distribution called a *chi-square distribution*. Using this distribution, we will be able to obtain a p-value to evaluate the hypotheses.

6.3.3 The chi-square distribution and finding areas

The **chi-square distribution** is sometimes used to characterize data sets and statistics that are always positive and typically right skewed. Recall the normal distribution had two parameters – mean and standard deviation – that could be used to describe its exact characteristics. The chi-square distribution has just one parameter called **degrees of freedom (df)**, which influences the shape, center, and spread of the distribution.

⊙ **Exercise 6.20** Figure 6.7 shows three chi-square distributions. (a) How does the center of the distribution change when the degrees of freedom is larger? (b) What about the variability (spread)? (c) How does the shape change?[14]

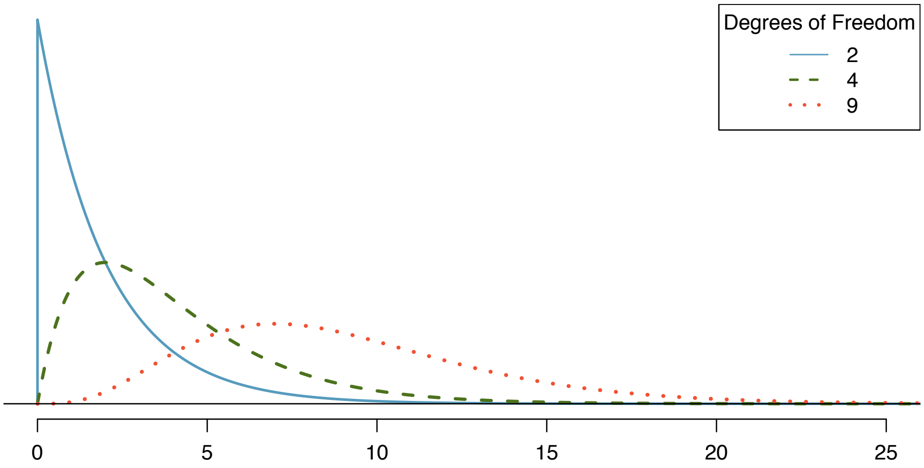


Figure 6.7: Three chi-square distributions with varying degrees of freedom.

Figure 6.7 and Exercise 6.20 demonstrate three general properties of chi-square distributions as the degrees of freedom increases: the distribution becomes more symmetric, the center moves to the right, and the variability inflates.

Our principal interest in the chi-square distribution is the calculation of p-values, which (as we have seen before) is related to finding the relevant area in the tail of a distribution. To do so, a new table is needed: the **chi-square table**, partially shown in Table 6.8. A more complete table is presented in Appendix B.3 on page 412. This table is very similar to the t table from Sections 5.3 and 5.4: we identify a range for the area, and we examine a particular row for distributions with different degrees of freedom. One important difference from the t table is that the chi-square table only provides upper tail values.

[14] (a) The center becomes larger. If we look carefully, we can see that the center of each distribution is equal to the distribution's degrees of freedom. (b) The variability increases as the degrees of freedom increases. (c) The distribution is very strongly skewed for $df = 2$, and then the distributions become more symmetric for the larger degrees of freedom $df = 4$ and $df = 9$. We would see this trend continue if we examined distributions with even more larger degrees of freedom.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	2	2.41	**3.22**	**4.61**	5.99	7.82	9.21	10.60	13.82
	3	*3.66*	*4.64*	6.25	*7.81*	*9.84*	*11.34*	*12.84*	*16.27*
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

Table 6.8: A section of the chi-square table. A complete table is in Appendix B.3 on page 412.

Example 6.21 Figure 6.9(a) shows a chi-square distribution with 3 degrees of freedom and an upper shaded tail starting at 6.25. Use Table 6.8 to estimate the shaded area.

This distribution has three degrees of freedom, so only the row with 3 degrees of freedom (df) is relevant. This row has been italicized in the table. Next, we see that the value – 6.25 – falls in the column with upper tail area 0.1. That is, the shaded upper tail of Figure 6.9(a) has area 0.1.

Example 6.22 We rarely observe the *exact* value in the table. For instance, Figure 6.9(b) shows the upper tail of a chi-square distribution with 2 degrees of freedom. The bound for this upper tail is at 4.3, which does not fall in Table 6.8. Find the approximate tail area.

The cutoff 4.3 falls between the second and third columns in the 2 degrees of freedom row. Because these columns correspond to tail areas of 0.2 and 0.1, we can be certain that the area shaded in Figure 6.9(b) is between 0.1 and 0.2.

Example 6.23 Figure 6.9(c) shows an upper tail for a chi-square distribution with 5 degrees of freedom and a cutoff of 5.1. Find the tail area.

Looking in the row with 5 df, 5.1 falls below the smallest cutoff for this row (6.06). That means we can only say that the area is *greater than 0.3*.

Exercise 6.24 Figure 6.9(d) shows a cutoff of 11.7 on a chi-square distribution with 7 degrees of freedom. Find the area of the upper tail.[15]

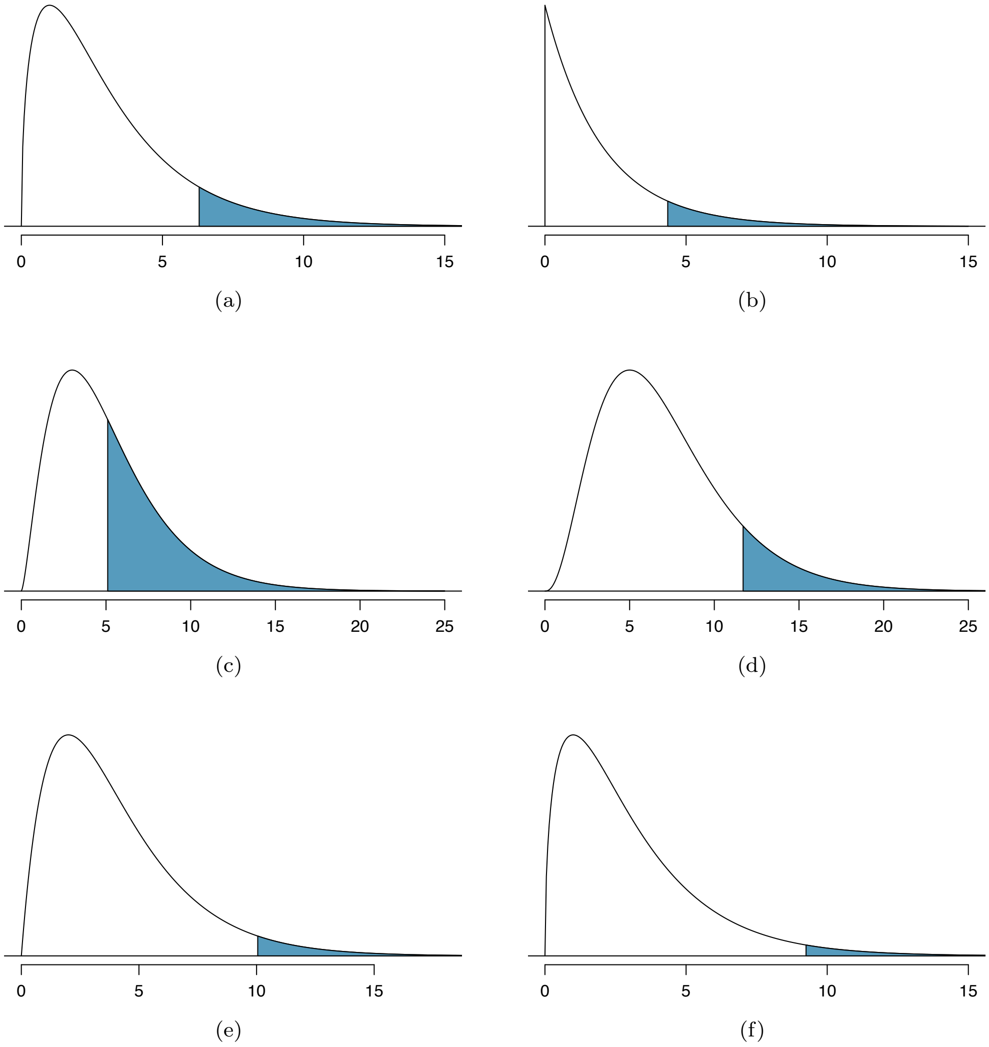
Exercise 6.25 Figure 6.9(e) shows a cutoff of 10 on a chi-square distribution with 4 degrees of freedom. Find the area of the upper tail.[16]

Exercise 6.26 Figure 6.9(f) shows a cutoff of 9.21 with a chi-square distribution with 3 df. Find the area of the upper tail.[17]

[15] The value 11.7 falls between 9.80 and 12.02 in the 7 df row. Thus, the area is between 0.1 and 0.2.

[16] The area is between 0.02 and 0.05.

[17] Between 0.02 and 0.05.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 6.9: **(a)** Chi-square distribution with 3 degrees of freedom, area above 6.25 shaded. **(b)** 2 degrees of freedom, area above 4.3 shaded. **(c)** 5 degrees of freedom, area above 5.1 shaded. **(d)** 7 degrees of freedom, area above 11.7 shaded. **(e)** 4 degrees of freedom, area above 10 shaded. **(f)** 3 degrees of freedom, area above 9.21 shaded.

6.3.4 Finding a p-value for a chi-square distribution

In Section 6.3.2, we identified a new test statistic (X^2) within the context of assessing whether there was evidence of racial bias in how jurors were sampled. The null hypothesis represented the claim that jurors were randomly sampled and there was no racial bias. The alternative hypothesis was that there was racial bias in how the jurors were sampled.

We determined that a large X^2 value would suggest strong evidence favoring the alternative hypothesis: that there was racial bias. However, we could not quantify what the chance was of observing such a large test statistic ($X^2 = 5.89$) if the null hypothesis actually was true. This is where the chi-square distribution becomes useful. If the null hypothesis was true and there was no racial bias, then X^2 would follow a chi-square distribution, with three degrees of freedom in this case. Under certain conditions, the statistic X^2 follows a chi-square distribution with $k - 1$ degrees of freedom, where k is the number of bins.

● **Example 6.27** How many categories were there in the juror example? How many degrees of freedom should be associated with the chi-square distribution used for X^2?

In the jurors example, there were $k = 4$ categories: white, black, Hispanic, and other. According to the rule above, the test statistic X^2 should then follow a chi-square distribution with $k - 1 = 3$ degrees of freedom if H_0 is true.

Just like we checked sample size conditions to use the normal model in earlier sections, we must also check a sample size condition to safely apply the chi-square distribution for X^2. Each expected count must be at least 5. In the juror example, the expected counts were 198, 19.25, 33, and 24.75, all easily above 5, so we can apply the chi-square model to the test statistic, $X^2 = 5.89$.

● **Example 6.28** If the null hypothesis is true, the test statistic $X^2 = 5.89$ would be closely associated with a chi-square distribution with three degrees of freedom. Using this distribution and test statistic, identify the p-value.

The chi-square distribution and p-value are shown in Figure 6.10. Because larger chi-square values correspond to stronger evidence against the null hypothesis, we shade the upper tail to represent the p-value. Using the chi-square table in Appendix B.3 or the short table on page 277, we can determine that the area is between 0.1 and 0.2. That is, the p-value is larger than 0.1 but smaller than 0.2. Generally we do not reject the null hypothesis with such a large p-value. In other words, the data do not provide convincing evidence of racial bias in the juror selection.

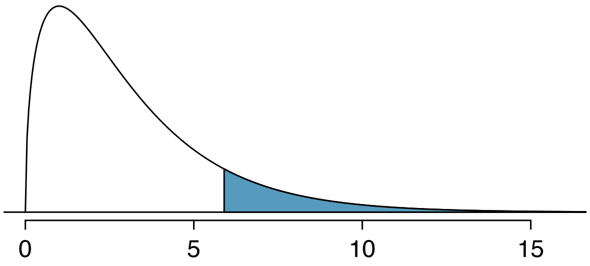


Figure 6.10: The p-value for the juror hypothesis test is shaded in the chi-square distribution with $df = 3$.

Chi-square test for one-way table

Suppose we are to evaluate whether there is convincing evidence that a set of observed counts O_1, O_2, ..., O_k in k categories are unusually different from what might be expected under a null hypothesis. Call the *expected counts* that are based on the null hypothesis E_1, E_2, ..., E_k. If each expected count is at least 5 and the null hypothesis is true, then the test statistic below follows a chi-square distribution with $k - 1$ degrees of freedom:

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k}$$

The p-value for this test statistic is found by looking at the upper tail of this chi-square distribution. We consider the upper tail because larger values of X^2 would provide greater evidence against the null hypothesis.

TIP: Conditions for the chi-square test

There are three conditions that must be checked before performing a chi-square test:

Independence. Each case that contributes a count to the table must be independent of all the other cases in the table.

Sample size / distribution. Each particular scenario (i.e. cell count) must have at least 5 expected cases.

Degrees of freedom We only apply the chi-square technique when the table is associated with a chi-square distribution with 2 or more degrees of freedom.

Failing to check conditions may affect the test's error rates.

When examining a table with just two bins, pick a single bin and use the one-proportion methods introduced in Section 6.1.

6.3.5 Evaluating goodness of fit for a distribution

Section 3.3 would be useful background reading for this example, but it is not a prerequisite.

We can apply our new chi-square testing framework to the second problem in this section: evaluating whether a certain statistical model fits a data set. Daily stock returns from the S&P500 for 1990-2011 can be used to assess whether stock activity each day is independent of the stock's behavior on previous days. This sounds like a very complex question, and it is, but a chi-square test can be used to study the problem. We will label each day as `Up` or `Down` (`D`) depending on whether the market was up or down that day. For example, consider the following changes in price, their new labels of up and down, and then the number of days that must be observed before each `Up` day:

Change in price	2.52	-1.46	0.51	-4.07	3.36	1.10	-5.46	-1.03	-2.99	1.71
Outcome	Up	D	Up	D	Up	Up	D	D	D	Up
Days to Up	1	-	2	-	2	1	-	-	-	4

If the days really are independent, then the number of days until a positive trading day should follow a geometric distribution. The geometric distribution describes the probability of waiting for the k^{th} trial to observe the first success. Here each up day (Up) represents a success, and down (D) days represent failures. In the data above, it took only one day until the market was up, so the first wait time was 1 day. It took two more days before we observed our next `Up` trading day, and two more for the third `Up` day. We would like to determine if these counts (1, 2, 2, 1, 4, and so on) follow the geometric distribution. Table 6.11 shows the number of waiting days for a positive trading day during 1990-2011 for the S&P500.

Days	1	2	3	4	5	6	7+	Total
Observed	1532	760	338	194	74	33	17	2948

Table 6.11: Observed distribution of the waiting time until a positive trading day for the S&P500, 1990-2011.

We consider how many days one must wait until observing an `Up` day on the S&P500 stock exchange. If the stock activity was independent from one day to the next and the probability of a positive trading day was constant, then we would expect this waiting time to follow a *geometric distribution*. We can organize this into a hypothesis framework:

H_0: The stock market being up or down on a given day is independent from all other days. We will consider the number of days that pass until an `Up` day is observed. Under this hypothesis, the number of days until an `Up` day should follow a geometric distribution.

H_A: The stock market being up or down on a given day is not independent from all other days. Since we know the number of days until an `Up` day would follow a geometric distribution under the null, we look for deviations from the geometric distribution, which would support the alternative hypothesis.

There are important implications in our result for stock traders: if information from past trading days is useful in telling what will happen today, that information may provide an advantage over other traders.

We consider data for the S&P500 from 1990 to 2011 and summarize the waiting times in Table 6.12 and Figure 6.13. The S&P500 was positive on 53.2% of those days.

Because applying the chi-square framework requires expected counts to be at least 5, we have *binned* together all the cases where the waiting time was at least 7 days to ensure each expected count is well above this minimum. The actual data, shown in the *Observed* row in Table 6.12, can be compared to the expected counts from the *Geometric Model* row. The method for computing expected counts is discussed in Table 6.12. In general, the expected counts are determined by (1) identifying the null proportion associated with each

Days	1	2	3	4	5	6	7+	Total
Observed	1532	760	338	194	74	33	17	2948
Geometric Model	1569	734	343	161	75	35	31	2948

Table 6.12: Distribution of the waiting time until a positive trading day. The expected counts based on the geometric model are shown in the last row. To find each expected count, we identify the probability of waiting D days based on the geometric model ($P(D) = (1 - 0.532)^{D-1}(0.532)$) and multiply by the total number of streaks, 2948. For example, waiting for three days occurs under the geometric model about $0.468^2 \times 0.532 = 11.65\%$ of the time, which corresponds to $0.1165 \times 2948 = 343$ streaks.

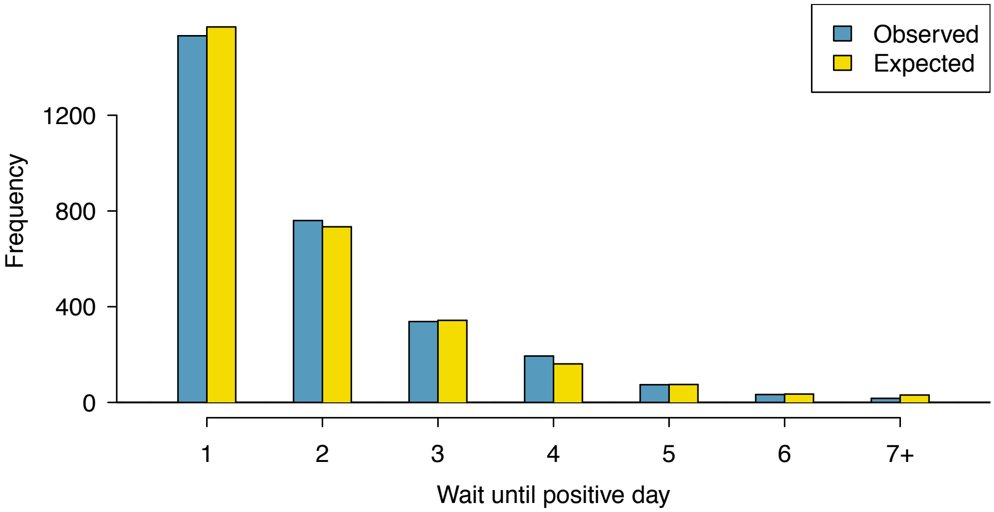


Figure 6.13: Side-by-side bar plot of the observed and expected counts for each waiting time.

bin, then (2) multiplying each null proportion by the total count to obtain the expected counts. That is, this strategy identifies what proportion of the total count we would expect to be in each bin.

● **Example 6.29** Do you notice any unusually large deviations in the graph? Can you tell if these deviations are due to chance just by looking?

It is not obvious whether differences in the observed counts and the expected counts from the geometric distribution are significantly different. That is, it is not clear whether these deviations might be due to chance or whether they are so strong that the data provide convincing evidence against the null hypothesis. However, we can perform a chi-square test using the counts in Table 6.12.

⊙ **Exercise 6.30** Table 6.12 provides a set of count data for waiting times ($O_1 = 1532$, $O_2 = 760$, ...) and expected counts under the geometric distribution ($E_1 = 1569$, $E_2 = 734$, ...). Compute the chi-square test statistic, X^2.[18]

⊙ **Exercise 6.31** Because the expected counts are all at least 5, we can safely apply the chi-square distribution to X^2. However, how many degrees of freedom should we use?[19]

● **Example 6.32** If the observed counts follow the geometric model, then the chi-square test statistic $X^2 = 15.08$ would closely follow a chi-square distribution with $df = 6$. Using this information, compute a p-value.

Figure 6.14 shows the chi-square distribution, cutoff, and the shaded p-value. If we look up the statistic $X^2 = 15.08$ in Appendix B.3, we find that the p-value is between 0.01 and 0.02. In other words, we have sufficient evidence to reject the notion that

[18] $X^2 = \frac{(1532-1569)^2}{1569} + \frac{(760-734)^2}{734} + \cdots + \frac{(17-31)^2}{31} = 15.08$

[19] There are $k = 7$ groups, so we use $df = k - 1 = 6$.

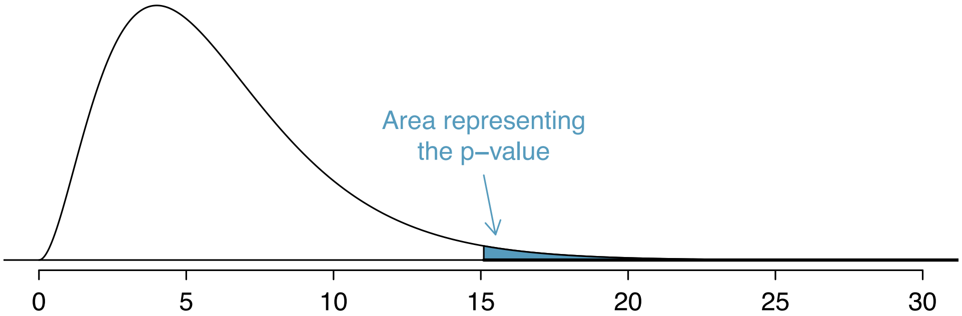


Figure 6.14: Chi-square distribution with 6 degrees of freedom. The p-value for the stock analysis is shaded.

the wait times follow a geometric distribution, i.e. trading days are not independent and past days may help predict what the stock market will do today.

Example 6.33 In Example 6.32, we rejected the null hypothesis that the trading days are independent. Why is this so important?

Because the data provided strong evidence that the geometric distribution is not appropriate, we reject the claim that trading days are independent. While it is not obvious how to exploit this information, it suggests there are some hidden patterns in the data that could be interesting and possibly useful to a stock trader.

6.4 Testing for independence in two-way tables (special topic)

Google is constantly running experiments to test new search algorithms. For example, Google might test three algorithms using a sample of 10,000 google.com search queries. Table 6.15 shows an example of 10,000 queries split into three algorithm groups.[20] The group sizes were specified before the start of the experiment to be 5000 for the current algorithm and 2500 for each test algorithm.

Search algorithm	current	test 1	test 2	Total
Counts	5000	2500	2500	10000

Table 6.15: Google experiment breakdown of test subjects into three search groups.

[20] Google regularly runs experiments in this manner to help improve their search engine. It is entirely possible that if you perform a search and so does your friend, that you will have different search results. While the data presented in this section resemble what might be encountered in a real experiment, these data are simulated.

● **Example 6.34** What is the ultimate goal of the Google experiment? What are the null and alternative hypotheses, in regular words?

The ultimate goal is to see whether there is a difference in the performance of the algorithms. The hypotheses can be described as the following:

H_0: The algorithms each perform equally well.

H_A: The algorithms do not perform equally well.

In this experiment, the explanatory variable is the search algorithm. However, an outcome variable is also needed. This outcome variable should somehow reflect whether the search results align with the user's interests. One possible way to quantify this is to determine whether (1) the user clicked one of the links provided and did not try a new search, or (2) the user performed a related search. Under scenario (1), we might think that the user was satisfied with the search results. Under scenario (2), the search results probably were not relevant, so the user tried a second search.

Table 6.16 provides the results from the experiment. These data are very similar to the count data in Section 6.3. However, now the different combinations of two variables are binned in a *two-way* table. In examining these data, we want to evaluate whether there is strong evidence that at least one algorithm is performing better than the others. To do so, we apply a chi-square test to this two-way table. The ideas of this test are similar to those ideas in the one-way table case. However, degrees of freedom and expected counts are computed a little differently than before.

Search algorithm	current	test 1	test 2	Total
No new search	3511	1749	1818	7078
New search	1489	751	682	2922
Total	5000	2500	2500	10000

Table 6.16: Results of the Google search algorithm experiment.

What is so different about one-way tables and two-way tables?
A one-way table describes counts for each outcome in a single variable. A two-way table describes counts for *combinations* of outcomes for two variables. When we consider a two-way table, we often would like to know, are these variables related in any way? That is, are they dependent (versus independent)?

The hypothesis test for this Google experiment is really about assessing whether there is statistically significant evidence that the choice of the algorithm affects whether a user performs a second search. In other words, the goal is to check whether the `search` variable is independent of the `algorithm` variable.

6.4.1 Expected counts in two-way tables

● **Example 6.35** From the experiment, we estimate the proportion of users who were satisfied with their initial search (no new search) as 7078/10000 = 0.7078. If there really is no difference among the algorithms and 70.78% of people are satisfied with the search results, how many of the 5000 people in the "current algorithm" group would be expected to not perform a new search?

About 70.78% of the 5000 would be satisfied with the initial search:

$$0.7078 \times 5000 = 3539 \text{ users}$$

That is, if there was no difference between the three groups, then we would expect 3539 of the current algorithm users not to perform a new search.

⊙ **Exercise 6.36** Using the same rationale described in Example 6.35, about how many users in each test group would not perform a new search if the algorithms were equally helpful?[21]

We can compute the expected number of users who would perform a new search for each group using the same strategy employed in Example 6.35 and Exercise 6.36. These expected counts were used to construct Table 6.17, which is the same as Table 6.16, except now the expected counts have been added in parentheses.

Search algorithm	current		test 1		test 2		Total
No new search	3511	**(3539)**	1749	**(1769.5)**	1818	**(1769.5)**	7078
New search	1489	**(1461)**	751	**(730.5)**	682	**(730.5)**	2922
Total	5000		2500		2500		10000

Table 6.17: The observed counts and the **(expected counts)**.

The examples and exercises above provided some help in computing expected counts. In general, expected counts for a two-way table may be computed using the row totals, column totals, and the table total. For instance, if there was no difference between the groups, then about 70.78% of each column should be in the first row:

$$\begin{aligned} 0.7078 \times (\text{column 1 total}) &= 3539 \\ 0.7078 \times (\text{column 2 total}) &= 1769.5 \\ 0.7078 \times (\text{column 3 total}) &= 1769.5 \end{aligned}$$

Looking back to how the fraction 0.7078 was computed – as the fraction of users who did not perform a new search (7078/10000) – these three expected counts could have been computed as

$$\begin{aligned} \left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 1 total}) &= 3539 \\ \left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 2 total}) &= 1769.5 \\ \left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 3 total}) &= 1769.5 \end{aligned}$$

This leads us to a general formula for computing expected counts in a two-way table when we would like to test whether there is strong evidence of an association between the column variable and row variable.

[21] We would expect $0.7078 * 2500 = 1769.5$. It is okay that this is a fraction.

Computing expected counts in a two-way table

To identify the expected count for the i^{th} row and j^{th} column, compute

$$\text{Expected Count}_{\text{row } i, \text{ col } j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{table total}}$$

6.4.2 The chi-square test for two-way tables

The chi-square test statistic for a two-way table is found the same way it is found for a one-way table. For each table count, compute

$$\begin{array}{ll}
\text{General formula} & \dfrac{(\text{observed count } - \text{ expected count})^2}{\text{expected count}} \\
\text{Row 1, Col 1} & \dfrac{(3511 - 3539)^2}{3539} = 0.222 \\
\text{Row 1, Col 2} & \dfrac{(1749 - 1769.5)^2}{1769.5} = 0.237 \\
\vdots & \vdots \\
\text{Row 2, Col 3} & \dfrac{(682 - 730.5)^2}{730.5} = 3.220
\end{array}$$

Adding the computed value for each cell gives the chi-square test statistic X^2:

$$X^2 = 0.222 + 0.237 + \cdots + 3.220 = 6.120$$

Just like before, this test statistic follows a chi-square distribution. However, the degrees of freedom are computed a little differently for a two-way table.[22] For two way tables, the degrees of freedom is equal to

$$df = (\text{number of rows minus 1}) \times (\text{number of columns minus 1})$$

In our example, the degrees of freedom parameter is

$$df = (2 - 1) \times (3 - 1) = 2$$

If the null hypothesis is true (i.e. the algorithms are equally useful), then the test statistic $X^2 = 6.12$ closely follows a chi-square distribution with 2 degrees of freedom. Using this information, we can compute the p-value for the test, which is depicted in Figure 6.18.

Computing degrees of freedom for a two-way table

When applying the chi-square test to a two-way table, we use

$$df = (R - 1) \times (C - 1)$$

where R is the number of rows in the table and C is the number of columns.

[22] Recall: in the one-way table, the degrees of freedom was the number of cells minus 1.

		Congress		
	Obama	Democrats	Republicans	Total
Approve	842	736	541	2119
Disapprove	616	646	842	2104
Total	1458	1382	1383	4223

Table 6.19: Pew Research poll results of a March 2012 poll.

TIP: Use two-proportion methods for 2-by-2 contingency tables
When analyzing 2-by-2 contingency tables, use the two-proportion methods introduced in Section 6.2.

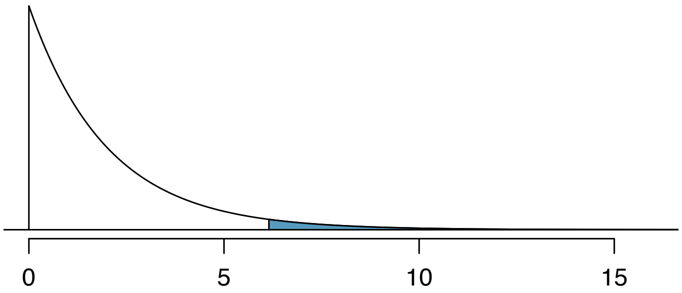


Figure 6.18: Computing the p-value for the Google hypothesis test.

● **Example 6.37** Compute the p-value and draw a conclusion about whether the search algorithms have different performances.

Looking in Appendix B.3 on page 412, we examine the row corresponding to 2 degrees of freedom. The test statistic, $X^2 = 6.120$, falls between the fourth and fifth columns, which means the p-value is between 0.02 and 0.05. Because we typically test at a significance level of $\alpha = 0.05$ and the p-value is less than 0.05, the null hypothesis is rejected. That is, the data provide convincing evidence that there is some difference in performance among the algorithms.

● **Example 6.38** Table 6.19 summarizes the results of a Pew Research poll.[23] We would like to determine if there are actually differences in the approval ratings of Barack Obama, Democrats in Congress, and Republicans in Congress. What are appropriate hypotheses for such a test?

H_0: There is no difference in approval ratings between the three groups.

H_A: There is some difference in approval ratings between the three groups, e.g. perhaps Obama's approval differs from Democrats in Congress.

[23] See the Pew Research website: www.people-press.org/2012/03/14/romney-leads-gop-contest-trails-in-matchup-with-obama. The counts in Table 6.19 are approximate.

Exercise 6.39 A chi-square test for a two-way table may be used to test the hypotheses in Example 6.38. As a first step, compute the expected values for each of the six table cells.[24]

Exercise 6.40 Compute the chi-square test statistic.[25]

Exercise 6.41 Because there are 2 rows and 3 columns, the degrees of freedom for the test is $df = (2-1) \times (3-1) = 2$. Use $X^2 = 106.4$, $df = 2$, and the chi-square table on page 412 to evaluate whether to reject the null hypothesis.[26]

6.5 Small sample hypothesis testing for a proportion (special topic)

In this section we develop inferential methods for a single proportion that are appropriate when the sample size is too small to apply the normal model to $\hat{p}$. Just like the methods related to the t distribution, these methods can also be applied to large samples.

6.5.1 When the success-failure condition is not met

People providing an organ for donation sometimes seek the help of a special "medical consultant". These consultants assist the patient in all aspects of the surgery, with the goal of reducing the possibility of complications during the medical procedure and recovery. Patients might choose a consultant based in part on the historical complication rate of the consultant's clients. One consultant tried to attract patients by noting the average complication rate for liver donor surgeries in the US is about 10%, but her clients have only had 3 complications in the 62 liver donor surgeries she has facilitated. She claims this is strong evidence that her work meaningfully contributes to reducing complications (and therefore she should be hired!).

Exercise 6.42 We will let p represent the true complication rate for liver donors working with this consultant. Estimate p using the data, and label this value $\hat{p}$.[27]

Example 6.43 Is it possible to assess the consultant's claim using the data provided?

No. The claim is that there is a causal connection, but the data are observational. Patients who hire this medical consultant may have lower complication rates for other reasons.

[24] The expected count for row one / column one is found by multiplying the row one total (2119) and column one total (1458), then dividing by the table total (4223): $\frac{2119\times1458}{3902} = 731.6$. Similarly for the first column and the second row: $\frac{2104\times1458}{4223} = 726.4$. Column 2: 693.5 and 688.5. Column 3: 694.0 and 689.0

[25] For each cell, compute $\frac{(\text{obs}-\text{exp})^2}{exp}$. For instance, the first row and first column: $\frac{(842-731.6)^2}{731.6} = 16.7$. Adding the results of each cell gives the chi-square test statistic: $X^2 = 16.7 + \cdots + 34.0 = 106.4$.

[26] The test statistic is larger than the right-most column of the $df = 2$ row of the chi-square table, meaning the p-value is less than 0.001. That is, we reject the null hypothesis because the p-value is less than 0.05, and we conclude that Americans' approval has differences among Democrats in Congress, Republicans in Congress, and the president.

[27] The sample proportion: $\hat{p} = 3/62 = 0.048$

While it is not possible to assess this causal claim, it is still possible to test for an association using these data. For this question we ask, could the low complication rate of $\hat{p} = 0.048$ be due to chance?

Exercise 6.44 Write out hypotheses in both plain and statistical language to test for the association between the consultant's work and the true complication rate, p, for this consultant's clients.[28]

Example 6.45 In the examples based on large sample theory, we modeled $\hat{p}$ using the normal distribution. Why is this not appropriate here?

The independence assumption may be reasonable if each of the surgeries is from a different surgical team. However, the success-failure condition is not satisfied. Under the null hypothesis, we would anticipate seeing $62 \times 0.10 = 6.2$ complications, not the 10 required for the normal approximation.

The uncertainty associated with the sample proportion should not be modeled using the normal distribution. However, we would still like to assess the hypotheses from Exercise 6.44 in absence of the normal framework. To do so, we need to evaluate the possibility of a sample value ($\hat{p}$) this far below the null value, $p_0 = 0.10$. This possibility is usually measured with a p-value.

The p-value is computed based on the null distribution, which is the distribution of the test statistic if the null hypothesis is true. Supposing the null hypothesis is true, we can compute the p-value by identifying the chance of observing a test statistic that favors the alternative hypothesis at least as strongly as the observed test statistic. This can be done using simulation.

6.5.2 Generating the null distribution and p-value by simulation

We want to identify the sampling distribution of the test statistic ($\hat{p}$) if the null hypothesis was true. In other words, we want to see how the sample proportion changes due to chance alone. Then we plan to use this information to decide whether there is enough evidence to reject the null hypothesis.

Under the null hypothesis, 10% of liver donors have complications during or after surgery. Suppose this rate was really no different for the consultant's clients. If this was the case, we could *simulate* 62 clients to get a sample proportion for the complication rate from the null distribution.

Each client can be simulated using a deck of cards. Take one red card, nine black cards, and mix them up. Then drawing a card is one way of simulating the chance a patient has a complication *if the true complication rate is 10%* for the data. If we do this 62 times and compute the proportion of patients with complications in the simulation, $\hat{p}_{sim}$, then this sample proportion is exactly a sample from the null distribution.

An undergraduate student was paid \$2 to complete this simulation. There were 5 simulated cases with a complication and 57 simulated cases without a complication, i.e. $\hat{p}_{sim} = 5/62 = 0.081$.

[28] H_0: There is no association between the consultant's contributions and the clients' complication rate. In statistical language, $p = 0.10$. H_A: Patients who work with the consultant tend to have a complication rate lower than 10%, i.e. $p < 0.10$.

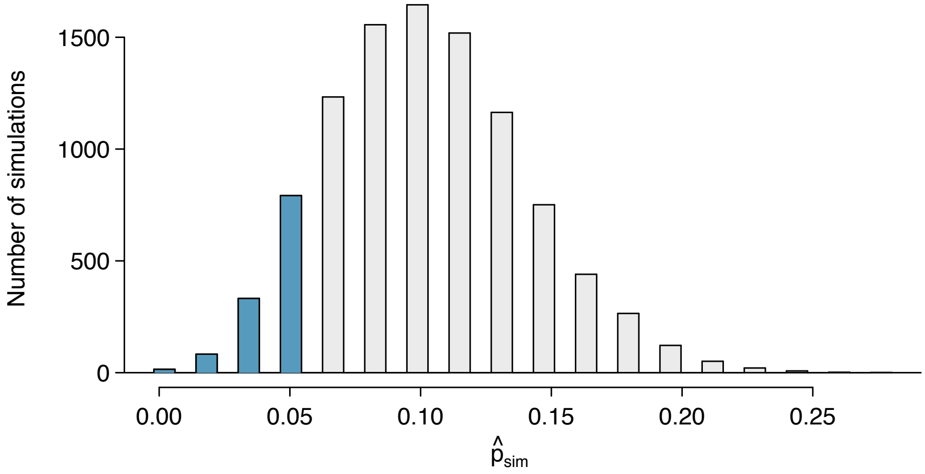


Figure 6.20: The null distribution for $\hat{p}$, created from 10,000 simulated studies. The left tail, representing the p-value for the hypothesis test, contains 12.22% of the simulations.

● **Example 6.46** Is this one simulation enough to determine whether or not we should reject the null hypothesis from Exercise 6.44? Explain.

No. To assess the hypotheses, we need to see a distribution of many $\hat{p}_{sim}$, not just a *single* draw from this sampling distribution.

One simulation isn't enough to get a sense of the null distribution; many simulation studies are needed. Roughly 10,000 seems sufficient. However, paying someone to simulate 10,000 studies by hand is a waste of time and money. Instead, simulations are typically programmed into a computer, which is much more efficient.

Figure 6.20 shows the results of 10,000 simulated studies. The proportions that are equal to or less than $\hat{p} = 0.048$ are shaded. The shaded areas represent sample proportions under the null distribution that provide at least as much evidence as $\hat{p}$ favoring the alternative hypothesis. There were 1222 simulated sample proportions with $\hat{p}_{sim} \leq 0.048$. We use these to construct the null distribution's left-tail area and find the p-value:

$$\text{left tail } = \frac{\text{Number of observed simulations with } \hat{p}_{sim} \leq \ 0.048}{10000} \tag{6.47}$$

Of the 10,000 simulated $\hat{p}_{sim}$, 1222 were equal to or smaller than $\hat{p}$. Since the hypothesis test is one-sided, the estimated p-value is equal to this tail area: 0.1222.

⊙ **Exercise 6.48** Because the estimated p-value is 0.1222, which is larger than the significance level 0.05, we do not reject the null hypothesis. Explain what this means in plain language in the context of the problem.[29]

⊙ **Exercise 6.49** Does the conclusion in Exercise 6.48 imply there is no real association between the surgical consultant's work and the risk of complications? Explain.[30]

[29] There isn't sufficiently strong evidence to support an association between the consultant's work and fewer surgery complications.

[30] No. It might be that the consultant's work is associated with a reduction but that there isn't enough data to convincingly show this connection.

One-sided hypothesis test for p with a small sample

The p-value is always derived by analyzing the null distribution of the test statistic. The normal model poorly approximates the null distribution for $\hat{p}$ when the success-failure condition is not satisfied. As a substitute, we can generate the null distribution using simulated sample proportions ($\hat{p}_{sim}$) and use this distribution to compute the tail area, i.e. the p-value.

We continue to use the same rule as before when computing the p-value for a two-sided test: double the single tail area, which remains a reasonable approach even when the sampling distribution is asymmetric. However, this can result in p-values larger than 1 when the point estimate is very near the mean in the null distribution; in such cases, we write that the p-value is 1. Also, very large p-values computed in this way (e.g. 0.85), may also be slightly inflated.

Exercise 6.48 said the p-value is *estimated*. It is not exact because the simulated null distribution itself is not exact, only a close approximation. However, we can generate an exact null distribution and p-value using the binomial model from Section 3.4.

6.5.3 Generating the exact null distribution and p-value

The number of successes in n independent cases can be described using the binomial model, which was introduced in Section 3.4. Recall that the probability of observing exactly k successes is given by

$$P(k \text{ successes}) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{6.50}$$

where p is the true probability of success. The expression $\binom{n}{k}$ is read as n *choose* k, and the exclamation points represent factorials. For instance, 3! is equal to $3 \times 2 \times 1 = 6$, 4! is equal to $4 \times 3 \times 2 \times 1 = 24$, and so on (see Section 3.4).

The tail area of the null distribution is computed by adding up the probability in Equation (6.50) for each k that provides at least as strong of evidence favoring the alternative hypothesis as the data. If the hypothesis test is one-sided, then the p-value is represented by a single tail area. If the test is two-sided, compute the single tail area and double it to get the p-value, just as we have done in the past.

Example 6.51 Compute the exact p-value to check the consultant's claim that her clients' complication rate is below 10%.

Exactly $k = 3$ complications were observed in the $n = 62$ cases cited by the consultant. Since we are testing against the 10% national average, our null hypothesis is $p = 0.10$. We can compute the p-value by adding up the cases where there are 3 or fewer complications:

$$\begin{aligned}
\text{p-value} &= \sum_{j=0}^{3} \binom{n}{j} p^j (1-p)^{n-j} \\
&= \sum_{j=0}^{3} \binom{62}{j} 0.1^j (1-0.1)^{62-j} \\
&= \binom{62}{0} 0.1^0 (1-0.1)^{62-0} + \binom{62}{1} 0.1^1 (1-0.1)^{62-1} \\
&\quad + \binom{62}{2} 0.1^2 (1-0.1)^{62-2} + \binom{62}{3} 0.1^3 (1-0.1)^{62-3} \\
&= 0.0015 + 0.0100 + 0.0340 + 0.0755 \\
&= 0.1210
\end{aligned}$$

This exact p-value is very close to the p-value based on the simulations (0.1222), and we come to the same conclusion. We do not reject the null hypothesis, and there is not statistically significant evidence to support the association.

If it were plotted, the exact null distribution would look almost identical to the simulated null distribution shown in Figure 6.20 on page 290.

6.5.4 Using simulation for goodness of fit tests

Simulation methods may also be used to test goodness of fit. In short, we simulate a new sample based on the purported bin probabilities, then compute a chi-square test statistic X^2_{sim}. We do this many times (e.g. 10,000 times), and then examine the distribution of these simulated chi-square test statistics. This distribution will be a very precise null distribution for the test statistic X^2 if the probabilities are accurate, and we can find the upper tail of this null distribution, using a cutoff of the observed test statistic, to calculate the p-value.

Example 6.52 Section 6.3 introduced an example where we considered whether jurors were racially representative of the population. Would our findings differ if we used a simulation technique?

Since the minimum bin count condition was satisfied, the chi-square distribution is an excellent approximation of the null distribution, meaning the results should be very similar. Figure 6.21 shows the simulated null distribution using 100,000 simulated X^2_{sim} values with an overlaid curve of the chi-square distribution. The distributions are almost identical, and the p-values are essentially indistinguishable: 0.115 for the simulated null distribution and 0.117 for the theoretical null distribution.

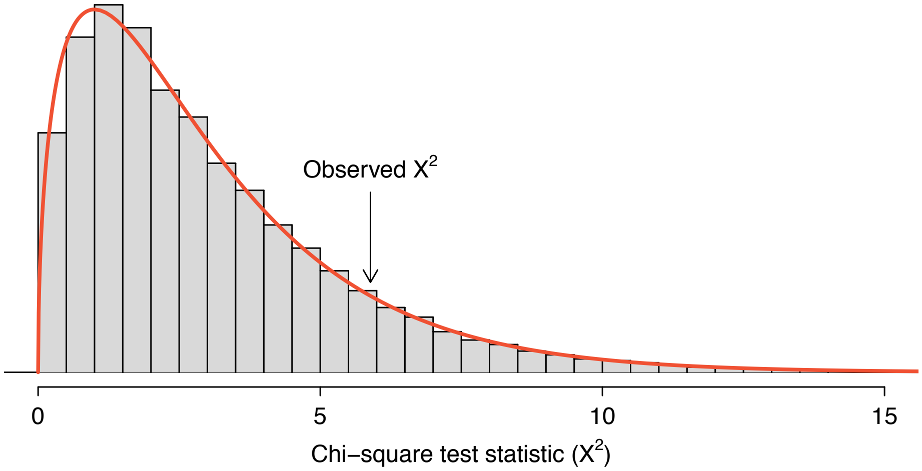


Figure 6.21: The precise null distribution for the juror example from Section 6.3 is shown as a histogram of simulated X^2_{sim} statistics, and the theoretical chi-square distribution is also shown.

6.6 Hypothesis testing for two proportions (special topic)

Cardiopulmonary resuscitation (CPR) is a procedure commonly used on individuals suffering a heart attack when other emergency resources are not available. This procedure is helpful in maintaining some blood circulation, but the chest compressions involved can also cause internal injuries. Internal bleeding and other injuries complicate additional treatment efforts following arrival at a hospital. For instance, blood thinners may be used to help release a clot that is causing the heart attack. However, the blood thinner would negatively affect an internal injury. Here we consider an experiment for patients who underwent CPR for a heart attack and were subsequently admitted to a hospital.[31] These patients were randomly divided into a treatment group where they received a blood thinner or the control group where they did not receive a blood thinner. The outcome variable of interest was whether the patients survived for at least 24 hours.

● **Example 6.53** Form hypotheses for this study in plain and statistical language. Let p_c represent the true survival rate of people who do not receive a blood thinner (corresponding to the control group) and p_t represent the survival rate for people receiving a blood thinner (corresponding to the treatment group).

We are interested in whether the blood thinners are helpful or harmful, so this should be a two-sided test.

H_0: Blood thinners do not have an overall survival effect, i.e. the survival proportions are the same in each group. $p_t - p_c = 0$.

H_A: Blood thinners do have an impact on survival. $p_t - p_c \neq 0$.

[31] *Efficacy and safety of thrombolytic therapy after initially unsuccessful cardiopulmonary resuscitation: a prospective clinical trial*, by Böttiger et al., The Lancet, 2001.

6.6.1 Large sample framework for a difference in two proportions

There were 50 patients in the experiment who did not receive the blood thinner and 40 patients who did. The study results are shown in Table 6.22.

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

Table 6.22: Results for the CPR study. Patients in the treatment group were given a blood thinner, and patients in the control group were not.

⊙ **Exercise 6.54** What is the observed survival rate in the control group? And in the treatment group? Also, provide a point estimate of the difference in survival proportions of the two groups: $\hat{p}_t - \hat{p}_c$.[32]

According to the point estimate, for patients who have undergone CPR outside of the hospital, an additional 13% of these patients survive when they are treated with blood thinners. However, we wonder if this difference could be easily explainable by chance. We'd like to investigate this using a large sample framework, but we first need to check the conditions for such an approach.

● **Example 6.55** Can the point estimate of the difference in survival proportions be adequately modeled using a normal distribution?

We will assume the patients are independent, which is probably reasonable. The success-failure condition is also satisfied. Since the proportions are equal under the null, we can compute the pooled proportion, $\hat{p} = (11 + 14)/(50 + 40) = 0.278$, for checking conditions. We find the expected number of successes (13.9, 11.1) and failures (36.1, 28.9) are above 10. The normal model is reasonable.

While we can apply a normal framework as an approximation to find a p-value, we might keep in mind that the expected number of successes is only 13.9 in one group and 11.1 in the other. Below we conduct an analysis relying on the large sample normal theory. We will follow up with a small sample analysis and compare the results.

● **Example 6.56** Assess the hypotheses presented in Example 6.53 using a large sample framework. Use a significance level of $\alpha = 0.05$.

We suppose the null distribution of the sample difference follows a normal distribution with mean 0 (the null value) and a standard deviation equal to the standard error of the estimate. The null hypothesis in this case would be that the two proportions are the same, so we compute the standard error using the pooled standard error formula from Equation (6.16) on page 273:

$$SE = \sqrt{\frac{p(1-p)}{n_t} + \frac{p(1-p)}{n_c}} \approx \sqrt{\frac{0.278(1-0.278)}{40} + \frac{0.278(1-0.278)}{50}} = 0.095$$

[32] Observed control survival rate: $p_c = \frac{11}{50} = 0.22$. Treatment survival rate: $p_t = \frac{14}{40} = 0.35$. Observed difference: $\hat{p}_t - \hat{p}_c = 0.35 - 0.22 = 0.13$.

where we have used the pooled estimate $\left(\hat{p} = \frac{11+14}{50+40} = 0.278\right)$ in place of the true proportion, p.

The null distribution with mean zero and standard deviation 0.095 is shown in Figure 6.23. We compute the tail areas to identify the p-value. To do so, we use the Z score of the point estimate:

$$Z = \frac{(\hat{p}_t - \hat{p}_c) - \text{null value}}{SE} = \frac{0.13 - 0}{0.095} = 1.37$$

If we look this Z score up in Appendix B.1, we see that the right tail has area 0.0853. The p-value is twice the single tail area: 0.176. This p-value does not provide convincing evidence that the blood thinner helps. Thus, there is insufficient evidence to conclude whether or not the blood thinner helps or hurts. (Remember, we never "accept" the null hypothesis – we can only reject or fail to reject.)

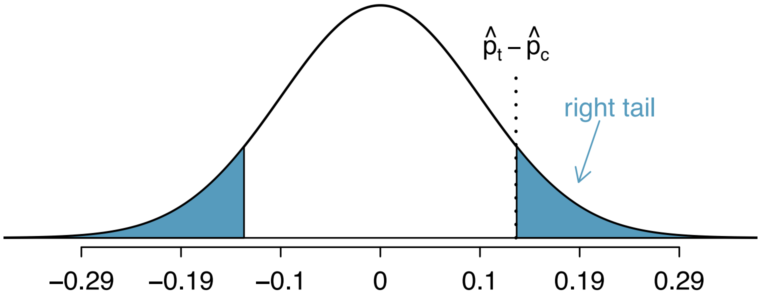


Figure 6.23: The null distribution of the point estimate $\hat{p}_t - \hat{p}_c$ under the large sample framework is a normal distribution with mean 0 and standard deviation equal to the standard error, in this case $SE = 0.095$. The p-value is represented by the shaded areas.

The p-value 0.176 relies on the normal approximation. We know that when the samples sizes are large, this approximation is quite good. However, when the sample sizes are relatively small as in this example, the approximation may only be adequate. Next we develop a simulation technique, apply it to these data, and compare our results. In general, the small sample method we develop may be used for any size sample, small or large, and should be considered as more accurate than the corresponding large sample technique.

6.6.2 Simulating a difference under the null distribution

The ideas in this section were first introduced in the optional Section 1.8 on page 42. For the interested reader, this earlier section provides a more in-depth discussion.

Suppose the null hypothesis is true. Then the blood thinner has no impact on survival and the 13% difference was due to chance. In this case, we can simulate *null* differences that are due to chance using a *randomization technique*.[33] By randomly assigning "fake treatment" and "fake control" stickers to the patients' files, we could get a new grouping – one that is completely due to chance. The expected difference between the two proportions under this simulation is zero.

[33] The test procedure we employ in this section is formally called a **permutation test**.

We run this simulation by taking 40 `treatment_fake` and 50 `control_fake` labels and randomly assigning them to the patients. The label counts of 40 and 50 correspond to the number of treatment and control assignments in the actual study. We use a computer program to randomly assign these labels to the patients, and we organize the simulation results into Table 6.24.

	Survived	Died	Total
`control_fake`	15	35	50
`treatment_fake`	10	30	40
Total	25	65	90

Table 6.24: Simulated results for the CPR study under the null hypothesis. The labels were randomly assigned and are independent of the outcome of the patient.

⊙ **Exercise 6.57** What is the difference in survival rates between the two fake groups in Table 6.24? How does this compare to the observed 13% in the real groups?[34]

The difference computed in Exercise 6.57 represents a draw from the null distribution of the sample differences. Next we generate many more simulated experiments to build up the null distribution, much like we did in Section 6.5.2 to build a null distribution for a one sample proportion.

Caution: Simulation in the two proportion case requires that the null difference is zero

The technique described here to simulate a difference from the null distribution relies on an important condition in the null hypothesis: there is no connection between the two variables considered. In some special cases, the null difference might not be zero, and more advanced methods (or a large sample approximation, if appropriate) would be necessary.

6.6.3 Null distribution for the difference in two proportions

We build up an approximation to the null distribution by repeatedly creating tables like the one shown in Table 6.24 and computing the sample differences. The null distribution from 10,000 simulations is shown in Figure 6.25.

● **Example 6.58** Compare Figures 6.23 and 6.25. How are they similar? How are they different?

The shapes are similar, but the simulated results show that the continuous approximation of the normal distribution is not very good. We might wonder, how close are the p-values?

[34] The difference is $\hat{p}_{t,fake} - \hat{p}_{c,fake} = \frac{10}{40} - \frac{15}{50} = -0.05$, which is closer to the null value $p_0 = 0$ than what we observed.

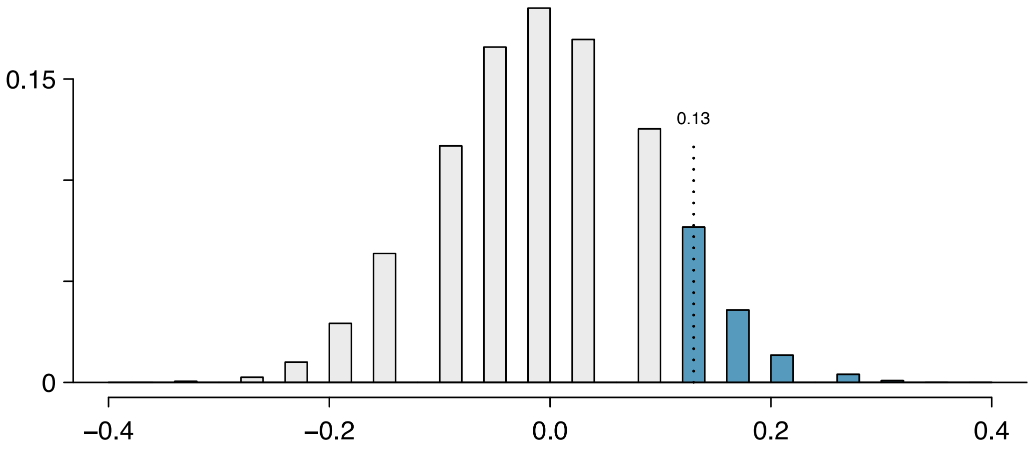


Figure 6.25: An approximation of the null distribution of the point estimate, $\hat{p}_t - \hat{p}_c$. The p-value is twice the right tail area.

⊙ **Exercise 6.59** The right tail area is about 0.13. (It is only a coincidence that we also have $\hat{p}_t - \hat{p}_c = 0.13$.) The p-value is computed by doubling the right tail area: 0.26. How does this value compare with the large sample approximation for the p-value?[35]

In general, small sample methods produce more accurate results since they rely on fewer assumptions. However, they often require some extra work or simulations. For this reason, many statisticians use small sample methods only when conditions for large sample methods are not satisfied.

6.6.4 Randomization for two-way tables and chi-square

Randomization methods may also be used for the contingency tables. In short, we create a randomized contingency table, then compute a chi-square test statistic X^2_{sim}. We repeat this many times using a computer, and then we examine the distribution of these simulated test statistics. This randomization approach is valid for any sized sample, and it will be more accurate for cases where one or more expected bin counts do not meet the minimum threshold of 5. When the minimum threshold is met, the simulated null distribution will very closely resemble the chi-square distribution. As before, we use the upper tail of the null distribution to calculate the p-value.

[35] The approximation in this case is fairly poor (p-values: 0.174 vs. 0.26), though we come to the same conclusion. The data do not provide convincing evidence showing the blood thinner helps or hurts patients.

6.7 Exercises

6.7.1 Inference for a single proportion

6.1 Vegetarian college students. Suppose that 8% of college students are vegetarians. Determine if the following statements are true or false, and explain your reasoning.

(a) The distribution of the sample proportions of vegetarians in random samples of size 60 is approximately normal since $n \geq 30$.

(b) The distribution of the sample proportions of vegetarian college students in random samples of size 50 is right skewed.

(c) A random sample of 125 college students where 12% are vegetarians would be considered unusual.

(d) A random sample of 250 college students where 12% are vegetarians would be considered unusual.

(e) The standard error would be reduced by one-half if we increased the sample size from 125 to 250.

6.2 Young Americans, Part I. About 77% of young adults think they can achieve the American dream. Determine if the following statements are true or false, and explain your reasoning.[36]

(a) The distribution of sample proportions of young Americans who think they can achieve the American dream in samples of size 20 is left skewed.

(b) The distribution of sample proportions of young Americans who think they can achieve the American dream in random samples of size 40 is approximately normal since $n \geq 30$.

(c) A random sample of 60 young Americans where 85% think they can achieve the American dream would be considered unusual.

(d) A random sample of 120 young Americans where 85% think they can achieve the American dream would be considered unusual.

6.3 Orange tabbies. Suppose that 90% of orange tabby cats are male. Determine if the following statements are true or false, and explain your reasoning.

(a) The distribution of sample proportions of random samples of size 30 is left skewed.

(b) Using a sample size that is 4 times as large will reduce the standard error of the sample proportion by one-half.

(c) The distribution of sample proportions of random samples of size 140 is approximately normal.

(d) The distribution of sample proportions of random samples of size 280 is approximately normal.

6.4 Young Americans, Part II. About 25% of young Americans have delayed starting a family due to the continued economic slump. Determine if the following statements are true or false, and explain your reasoning.[37]

(a) The distribution of sample proportions of young Americans who have delayed starting a family due to the continued economic slump in random samples of size 12 is right skewed.

(b) In order for the the distribution of sample proportions of young Americans who have delayed starting a family due to the continued economic slump to be approximately normal, we need random samples where the sample size is at least 40.

(c) A random sample of 50 young Americans where 20% have delayed starting a family due to the continued economic slump would be considered unusual.

(d) A random sample of 150 young Americans where 20% have delayed starting a family due to the continued economic slump would be considered unusual.

(e) Tripling the sample size will reduce the standard error of the sample proportion by one-third.

[36] A. Vaughn. "Poll finds young adults optimistic, but not about money". In: *Los Angeles Times* (2011).

[37] Demos.org. "The State of Young America: The Poll". In: (2011).

6.5 Prop 19 in California. In a 2010 Survey USA poll, 70% of the 119 respondents between the ages of 18 and 34 said they would vote in the 2010 general election for Prop 19, which would change California law to legalize marijuana and allow it to be regulated and taxed. At a 95% confidence level, this sample has an 8% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.[38]

(a) We are 95% confident that between 62% and 78% of the California voters in this sample support Prop 19.

(b) We are 95% confident that between 62% and 78% of all California voters between the ages of 18 and 34 support Prop 19.

(c) If we considered many random samples of 119 California voters between the ages of 18 and 34, and we calculated 95% confidence intervals for each, 95% of them will include the true population proportion of Californians who support Prop 19.

(d) In order to decrease the margin of error to 4%, we would need to quadruple (multiply by 4) the sample size.

(e) Based on this confidence interval, there is sufficient evidence to conclude that a majority of California voters between the ages of 18 and 34 support Prop 19.

6.6 2010 Healthcare Law. On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 46% of 1,012 Americans agree with this decision. At a 95% confidence level, this sample has a 3% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.[39]

(a) We are 95% confident that between 43% and 49% of Americans in this sample support the decision of the U.S. Supreme Court on the 2010 healthcare law.

(b) We are 95% confident that between 43% and 49% of Americans support the decision of the U.S. Supreme Court on the 2010 healthcare law.

(c) If we considered many random samples of 1,012 Americans, and we calculated the sample proportions of those who support the decision of the U.S. Supreme Court, 95% of those sample proportions will be between 43% and 49%.

(d) The margin of error at a 90% confidence level would be higher than 3%.

6.7 Fireworks on July 4^{th}. In late June 2012, Survey USA published results of a survey stating that 56% of the 600 randomly sampled Kansas residents planned to set off fireworks on July 4^{th}. Determine the margin of error for the 56% point estimate using a 95% confidence level.[40]

6.8 Elderly drivers. In January 2011, The Marist Poll published a report stating that 66% of adults nationally think licensed drivers should be required to retake their road test once they reach 65 years of age. It was also reported that interviews were conducted on 1,018 American adults, and that the margin of error was 3% using a 95% confidence level.[41]

(a) Verify the margin of error reported by The Marist Poll.

(b) Based on a 95% confidence interval, does the poll provide convincing evidence that *more than* 70% of the population think that licensed drivers should be required to retake their road test once they turn 65?

[38] Survey USA, Election Poll #16804, data collected July 8-11, 2010.

[39] Gallup, Americans Issue Split Decision on Healthcare Ruling, data collected June 28, 2012.

[40] Survey USA, News Poll #19333, data collected on June 27, 2012.

[41] Marist Poll, Road Rules: Re-Testing Drivers at Age 65?, March 4, 2011.

6.9 Life after college. We are interested in estimating the proportion of graduates at a mid-sized university who found a job within one year of completing their undergraduate degree. Suppose we conduct a survey and find out that 348 of the 400 randomly sampled graduates found jobs. The graduating class under consideration included over 4500 students.

(a) Describe the population parameter of interest. What is the value of the point estimate of this parameter?

(b) Check if the conditions for constructing a confidence interval based on these data are met.

(c) Calculate a 95% confidence interval for the proportion of graduates who found a job within one year of completing their undergraduate degree at this university, and interpret it in the context of the data.

(d) What does "95% confidence" mean?

(e) Now calculate a 99% confidence interval for the same parameter and interpret it in the context of the data.

(f) Compare the widths of the 95% and 99% confidence intervals. Which one is wider? Explain.

6.10 Life rating in Greece. Greece has faced a severe economic crisis since the end of 2009. A Gallup poll surveyed 1,000 randomly sampled Greeks in 2011 and found that 25% of them said they would rate their lives poorly enough to be considered "suffering".[42]

(a) Describe the population parameter of interest. What is the value of the point estimate of this parameter?

(b) Check if the conditions required for constructing a confidence interval based on these data are met.

(c) Construct a 95% confidence interval for the proportion of Greeks who are "suffering".

(d) Without doing any calculations, describe what would happen to the confidence interval if we decided to use a higher confidence level.

(e) Without doing any calculations, describe what would happen to the confidence interval if we used a larger sample.

6.11 Study abroad. A survey on 1,509 high school seniors who took the SAT and who completed an optional web survey between April 25 and April 30, 2007 shows that 55% of high school seniors are fairly certain that they will participate in a study abroad program in college.[43]

(a) Is this sample a representative sample from the population of all high school seniors in the US? Explain your reasoning.

(b) Let's suppose the conditions for inference are met. Even if your answer to part (a) indicated that this approach would not be reliable, this analysis may still be interesting to carry out (though not report). Construct a 90% confidence interval for the proportion of high school seniors (of those who took the SAT) who are fairly certain they will participate in a study abroad program in college, and interpret this interval in context.

(c) What does "90% confidence" mean?

(d) Based on this interval, would it be appropriate to claim that the majority of high school seniors are fairly certain that they will participate in a study abroad program in college?

[42] Gallup World, More Than One in 10 "Suffering" Worldwide, data collected throughout 2011.

[43] studentPOLL, College-Bound Students' Interests in Study Abroad and Other International Learning Activities, January 2008.

6.12 Legalization of marijuana, Part I. The 2010 General Social Survey asked 1,259 US residents: "Do you think the use of marijuana should be made legal, or not?" 48% of the respondents said it should be made legal.[44]

(a) Is 48% a sample statistic or a population parameter? Explain.

(b) Construct a 95% confidence interval for the proportion of US residents who think marijuana should be made legal, and interpret it in the context of the data.

(c) A critic points out that this 95% confidence interval is only accurate if the statistic follows a normal distribution, or if the normal model is a good approximation. Is this true for these data? Explain.

(d) A news piece on this survey's findings states, "Majority of Americans think marijuana should be legalized." Based on your confidence interval, is this news piece's statement justified?

6.13 Public option, Part I. A *Washington Post* article from 2009 reported that "support for a government-run health-care plan to compete with private insurers has rebounded from its summertime lows and wins clear majority support from the public." More specifically, the article says "seven in 10 Democrats back the plan, while almost nine in 10 Republicans oppose it. Independents divide 52 percent against, 42 percent in favor of the legislation." (6% responded with "other".) There were were 819 Democrats, 566 Republicans and 783 Independents surveyed.[45]

(a) A political pundit on TV claims that a majority of Independents oppose the health care public option plan. Do these data provide strong evidence to support this statement?

(b) Would you expect a confidence interval for the proportion of Independents who oppose the public option plan to include 0.5? Explain.

6.14 The Civil War. A national survey conducted in 2011 among a simple random sample of 1,507 adults shows that 56% of Americans think the Civil War is still relevant to American politics and political life.[46]

(a) Conduct a hypothesis test to determine if these data provide strong evidence that the majority of the Americans think the Civil War is still relevant.

(b) Interpret the p-value in this context.

(c) Calculate a 90% confidence interval for the proportion of Americans who think the Civil War is still relevant. Interpret the interval in this context, and comment on whether or not the confidence interval agrees with the conclusion of the hypothesis test.

6.15 Browsing on the mobile device. A 2012 survey of 2,254 American adults indicates that 17% of cell phone owners do their browsing on their phone rather than a computer or other device.[47]

(a) According to an online article, a report from a mobile research company indicates that 38 percent of Chinese mobile web users only access the internet through their cell phones.[48] Conduct a hypothesis test to determine if these data provide strong evidence that the proportion of Americans who only use their cell phones to access the internet is different than the Chinese proportion of 38%.

(b) Interpret the p-value in this context.

(c) Calculate a 95% confidence interval for the proportion of Americans who access the internet on their cell phones, and interpret the interval in this context.

[44] National Opinion Research Center, General Social Survey, 2010.

[45] D. Balz and J. Cohen. "Most support public option for health insurance, poll finds". In: *The Washington Post* (2009).

[46] Pew Research Center Publications, Civil War at 150: Still Relevant, Still Divisive, data collected between March 30 - April 3, 2011.

[47] Pew Internet, Cell Internet Use 2012, data collected between March 15 - April 13, 2012.

[48] S. Chang. "The Chinese Love to Use Feature Phone to Access the Internet". In: *M.I.C Gadget* (2012).

6.16 Is college worth it? Part I. Among a simple random sample of 331 American adults who do not have a four-year college degree and are not currently enrolled in school, 48% said they decided not to go to college because they could not afford school.[49]

(a) A newspaper article states that only a minority of the Americans who decide not to go to college do so because they cannot afford it and uses the point estimate from this survey as evidence. Conduct a hypothesis test to determine if these data provide strong evidence supporting this statement.

(b) Would you expect a confidence interval for the proportion of American adults who decide not to go to college because they cannot afford it to include 0.5? Explain.

6.17 Taste test. Some people claim that they can tell the difference between a diet soda and a regular soda in the first sip. A researcher wanting to test this claim randomly sampled 80 such people. He then filled 80 plain white cups with soda, half diet and half regular through random assignment, and asked each person to take one sip from their cup and identify the soda as diet or regular. 53 participants correctly identified the soda.

(a) Do these data provide strong evidence that these people are able to detect the difference between diet and regular soda, in other words, are the results significantly better than just random guessing?

(b) Interpret the p-value in this context.

6.18 Is college worth it? Part II. Exercise 6.16 presents the results of a poll where 48% of 331 Americans who decide to not go to college do so because they cannot afford it.

(a) Calculate a 90% confidence interval for the proportion of Americans who decide to not go to college because they cannot afford it, and interpret the interval in context.

(b) Suppose we wanted the margin of error for the 90% confidence level to be about 1.5%. How large of a survey would you recommend?

6.19 College smokers. We are interested in estimating the proportion of students at a university who smoke. Out of a random sample of 200 students from this university, 40 students smoke.

(a) Calculate a 95% confidence interval for the proportion of students at this university who smoke, and interpret this interval in context. (Reminder: check conditions)

(b) If we wanted the margin of error to be no larger than 2% at a 95% confidence level for the proportion of students who smoke, how big of a sample would we need?

6.20 Legalize Marijuana, Part II. As discussed in Exercise 6.12, the 2010 General Social Survey reported a sample where about 48% of US residents thought marijuana should be made legal. If we wanted to limit the margin of error of a 95% confidence interval to 2%, about how many Americans would we need to survey ?

6.21 Public option, Part II. Exercise 6.13 presents the results of a poll evaluating support for the health care public option in 2009, reporting that 52% of Independents in the sample opposed the public option. If we wanted to estimate this number to within 1% with 90% confidence, what would be an appropriate sample size?

6.22 Acetaminophen and liver damage. It is believed that large doses of acetaminophen (the active ingredient in over the counter pain relievers like Tylenol) may cause damage to the liver. A researcher wants to conduct a study to estimate the proportion of acetaminophen users who have liver damage. For participating in this study, he will pay each subject $20 and provide a free medical consultation if the patient has liver damage.

(a) If he wants to limit the margin of error of his 98% confidence interval to 2%, what is the minimum amount of money he needs to set aside to pay his subjects?

(b) The amount you calculated in part (a) is substantially over his budget so he decides to use fewer subjects. How will this affect the width of his confidence interval?

[49] Pew Research Center Publications, Is College Worth It?, data collected between March 15-29, 2011.

6.7.2 Difference of two proportions

6.23 Social experiment, Part I. A "social experiment" conducted by a TV program questioned what people do when they see a very obviously bruised woman getting picked on by her boyfriend. On two different occasions at the same restaurant, the same couple was depicted. In one scenario the woman was dressed "provocatively" and in the other scenario the woman was dressed "conservatively". The table below shows how many restaurant diners were present under each scenario, and whether or not they intervened.

		Scenario		
		Provocative	Conservative	Total
Intervene	Yes	5	15	20
	No	15	10	25
	Total	20	25	45

Explain why the sampling distribution of the difference between the proportions of interventions under provocative and conservative scenarios does not follow an approximately normal distribution.

6.24 Heart transplant success. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was officially designated a heart transplant candidate, meaning that he was gravely ill and might benefit from a new heart. Patients were randomly assigned into treatment and control groups. Patients in the treatment group received a transplant, and those in the control group did not. The table below displays how many patients survived and died in each group.[50]

	control	treatment
alive	4	24
dead	30	45

A hypothesis test would reject the conclusion that the survival rate is the same in each group, and so we might like to calculate a confidence interval. Explain why we cannot construct such an interval using the normal approximation. What might go wrong if we constructed the confidence interval despite this problem?

6.25 Gender and color preference. A 2001 study asked 1,924 male and 3,666 female undergraduate college students their favorite color. A 95% confidence interval for the difference between the proportions of males and females whose favorite color is black ($p_{male} - p_{female}$) was calculated to be (0.02, 0.06). Based on this information, determine if the following statements are true or false, and explain your reasoning for each statement you identify as false.[51]

(a) We are 95% confident that the true proportion of males whose favorite color is black is 2% lower to 6% higher than the true proportion of females whose favorite color is black.

(b) We are 95% confident that the true proportion of males whose favorite color is black is 2% to 6% higher than the true proportion of females whose favorite color is black.

(c) 95% of random samples will produce 95% confidence intervals that include the true difference between the population proportions of males and females whose favorite color is black.

(d) We can conclude that there is a significant difference between the proportions of males and females whose favorite color is black and that the difference between the two sample proportions is too large to plausibly be due to chance.

(e) The 95% confidence interval for ($p_{female} - p_{male}$) cannot be calculated with only the information given in this exercise.

[50] B. Turnbull et al. "Survivorship of Heart Transplant Data". In: *Journal of the American Statistical Association* 69 (1974), pp. 74–80.

[51] L Ellis and C Ficek. "Color preferences according to gender and sexual orientation". In: *Personality and Individual Differences* 31.8 (2001), pp. 1375–1379.

6.26 The Daily Show. A 2010 Pew Research foundation poll indicates that among 1,099 college graduates, 33% watch The Daily Show. Meanwhile, 22% of the 1,110 people with a high school degree but no college degree in the poll watch The Daily Show. A 95% confidence interval for $(p_{\text{college grad}} - p_{\text{HS or less}})$, where p is the proportion of those who watch The Daily Show, is (0.07, 0.15). Based on this information, determine if the following statements are true or false, and explain your reasoning if you identify the statement as false.[52]

(a) At the 5% significance level, the data provide convincing evidence of a difference between the proportions of college graduates and those with a high school degree or less who watch The Daily Show.

(b) We are 95% confident that 7% less to 15% more college graduates watch The Daily Show than those with a high school degree or less.

(c) 95% of random samples of 1,099 college graduates and 1,110 people with a high school degree or less will yield differences in sample proportions between 7% and 15%.

(d) A 90% confidence interval for $(p_{\text{college grad}} - p_{\text{HS or less}})$ would be wider.

(e) A 95% confidence interval for $(p_{\text{HS or less}} - p_{\text{college grad}})$ is (-0.15,-0.07).

6.27 Public Option, Part III. Exercise 6.13 presents the results of a poll evaluating support for the health care public option plan in 2009. 70% of 819 Democrats and 42% of 783 Independents support the public option.

(a) Calculate a 95% confidence interval for the difference between $(p_D - p_I)$ and interpret it in this context. We have already checked conditions for you.

(b) True or false: If we had picked a random Democrat and a random Independent at the time of this poll, it is more likely that the Democrat would support the public option than the Independent.

6.28 Sleep deprivation, CA vs. OR, Part I. According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents. Calculate a 95% confidence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived and interpret it in context of the data.[53]

6.29 Offshore drilling, Part I. A 2010 survey asked 827 randomly sampled registered voters in California "Do you support? Or do you oppose? Drilling for oil and natural gas off the Coast of California? Or do you not know enough to say?" Below is the distribution of responses, separated based on whether or not the respondent graduated from college.[54]

(a) What percent of college graduates and what percent of the non-college graduates in this sample do not know enough to have an opinion on drilling for oil and natural gas off the Coast of California?

(b) Conduct a hypothesis test to determine if the data provide strong evidence that the proportion of college graduates who do not have an opinion on this issue is different than that of non-college graduates.

	College	*Grad*
	Yes	No
Support	154	132
Oppose	180	126
Do not know	104	131
Total	438	389

[52] The Pew Research Center, Americans Spending More Time Following the News, data collected June 8-28, 2010.

[53] CDC, Perceived Insufficient Rest or Sleep Among Adults — United States, 2008.

[54] Survey USA, Election Poll #16804, data collected July 8-11, 2010.

6.30 Sleep deprivation, CA vs. OR, Part II. Exercise 6.28 provides data on sleep deprivation rates of Californians and Oregonians. The proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents.

(a) Conduct a hypothesis test to determine if these data provide strong evidence the rate of sleep deprivation is different for the two states. (Reminder: check conditions)

(b) It is possible the conclusion of the test in part (a) is incorrect. If this is the case, what type of error was made?

6.31 Offshore drilling, Part II. Results of a poll evaluating support for drilling for oil and natural gas off the coast of California were introduced in Exercise 6.29.

	College Grad	
	Yes	No
Support	154	132
Oppose	180	126
Do not know	104	131
Total	438	389

(a) What percent of college graduates and what percent of the non-college graduates in this sample support drilling for oil and natural gas off the Coast of California?

(b) Conduct a hypothesis test to determine if the data provide strong evidence that the proportion of college graduates who support off-shore drilling in California is different than that of non-college graduates.

6.32 Full body scan, Part I. A news article reports that "Americans have differing views on two potentially inconvenient and invasive practices that airports could implement to uncover potential terrorist attacks." This news piece was based on a survey conducted among a random sample of 1,137 adults nationwide, interviewed by telephone November 7-10, 2010, where one of the questions on the survey was "Some airports are now using 'full-body' digital x-ray machines to electronically screen passengers in airport security lines. Do you think these new x-ray machines should or should not be used at airports?" Below is a summary of responses based on party affiliation.[55]

		Party Affiliation		
		Republican	Democrat	Independent
	Should	264	299	351
Answer	Should not	38	55	77
	Don't know/No answer	16	15	22
	Total	318	369	450

(a) Conduct an appropriate hypothesis test evaluating whether there is a difference in the proportion of Republicans and Democrats who think the full-body scans should be applied in airports. Assume that all relevant conditions are met.

(b) The conclusion of the test in part (a) may be incorrect, meaning a testing error was made. If an error was made, was it a Type I or a Type II error? Explain.

[55] S. Condon. "Poll: 4 in 5 Support Full-Body Airport Scanners". In: *CBS News* (2010).

6.33 Sleep deprived transportation workers. The National Sleep Foundation conducted a survey on the sleep habits of randomly sampled transportation workers and a control sample of non-transportation workers. The results of the survey are shown below.[56]

		Transportation Professionals			
	Control	Pilots	Truck Drivers	Train Operators	Bux/Taxi/Limo Drivers
Less than 6 hours of sleep	35	19	35	29	21
6 to 8 hours of sleep	193	132	117	119	131
More than 8 hours	64	51	51	32	58
Total	292	202	203	180	210

Conduct a hypothesis test to evaluate if these data provide evidence of a difference between the proportions of truck drivers and non-transportation workers (the control group) who get less than 6 hours of sleep per day, i.e. are considered sleep deprived.

6.34 Prenatal vitamins and Autism. Researchers studying the link between prenatal vitamin use and autism surveyed the mothers of a random sample of children aged 24 - 60 months with autism and conducted another separate random sample for children with typical development. The table below shows the number of mothers in each group who did and did not use prenatal vitamins during the three months before pregnancy (periconceptional period).[57]

		Autism		
		Autism	Typical development	Total
Periconceptional	No vitamin	111	70	181
prenatal vitamin	Vitamin	143	159	302
	Total	254	229	483

(a) State appropriate hypotheses to test for independence of use of prenatal vitamins during the three months before pregnancy and autism.

(b) Complete the hypothesis test and state an appropriate conclusion. (Reminder: verify any necessary conditions for the test.)

(c) A New York Times article reporting on this study was titled “Prenatal Vitamins May Ward Off Autism”. Do you find the title of this article to be appropriate? Explain your answer. Additionally, propose an alternative title.[58]

6.35 HIV in sub-Saharan Africa. In July 2008 the US National Institutes of Health announced that it was stopping a clinical study early because of unexpected results. The study population consisted of HIV-infected women in sub-Saharan Africa who had been given single dose Nevaripine (a treatment for HIV) while giving birth, to prevent transmission of HIV to the infant. The study was a randomized comparison of continued treatment of a woman (after successful childbirth) with Nevaripine vs. Lopinavir, a second drug used to treat HIV. 240 women participated in the study; 120 were randomized to each of the two treatments. Twenty-four weeks after starting the study treatment, each woman was tested to determine if the HIV infection was becoming worse (an outcome called *virologic failure*). Twenty-six of the 120 women treated with Nevaripine experienced virologic failure, while 10 of the 120 women treated with the other drug experienced virologic failure.[59]

(a) Create a two-way table presenting the results of this study.

(b) State appropriate hypotheses to test for independence of treatment and virologic failure.

(c) Complete the hypothesis test and state an appropriate conclusion. (Reminder: verify any necessary conditions for the test.)

[56] National Sleep Foundation, 2012 Sleep in America Poll: Transportation Workers Sleep, 2012.

[57] R.J. Schmidt et al. “Prenatal vitamins, one-carbon metabolism gene variants, and risk for autism”. In: *Epidemiology* 22.4 (2011), p. 476.

[58] R.C. Rabin. “Patterns: Prenatal Vitamins May Ward Off Autism”. In: *New York Times* (2011).

[59] S. Lockman et al. “Response to antiretroviral therapy after a single, peripartum dose of nevirapine”. In: *Obstetrical & gynecological survey* 62.6 (2007), p. 361.

6.36 Diabetes and unemployment. A 2012 Gallup poll surveyed Americans about their employment status and whether or not they have diabetes. The survey results indicate that 1.5% of the 47,774 employed (full or part time) and 2.5% of the 5,855 unemployed 18-29 year olds have diabetes.[60]

(a) Create a two-way table presenting the results of this study.

(b) State appropriate hypotheses to test for independence of incidence of diabetes and employment status.

(c) The sample difference is about 1%. If we completed the hypothesis test, we would find that the p-value is very small (about 0), meaning the difference is statistically significant. Use this result to explain the difference between statistically significant and practically significant findings.

6.7.3 Testing for goodness of fit using chi-square

6.37 True or false, Part I. Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

(a) The chi-square distribution, just like the normal distribution, has two parameters, mean and standard deviation.

(b) The chi-square distribution is always right skewed, regardless of the value of the degrees of freedom parameter.

(c) The chi-square statistic is always positive.

(d) As the degrees of freedom increases, the shape of the chi-square distribution becomes more skewed.

6.38 True or false, Part II. Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

(a) As the degrees of freedom increases, the mean of the chi-square distribution increases.

(b) If you found $X^2 = 10$ with $df = 5$ you would fail to reject H_0 at the 5% significance level.

(c) When finding the p-value of a chi-square test, we always shade the tail areas in both tails.

(d) As the degrees of freedom increases, the variability of the chi-square distribution decreases.

6.39 Open source textbook. A professor using an open source introductory statistics book predicts that 60% of the students will purchase a hard copy of the book, 25% will print it out from the web, and 15% will read it online. At the end of the semester he asks his students to complete a survey where they indicate what format of the book they used. Of the 126 students, 71 said they bought a hard copy of the book, 30 said they printed it out from the web, and 25 said they read it online.

(a) State the hypotheses for testing if the professor's predictions were inaccurate.

(b) How many students did the professor expect to buy the book, print the book, and read the book exclusively online?

(c) This is an appropriate setting for a chi-square test. List the conditions required for a test and verify they are satisfied.

(d) Calculate the chi-squared statistic, the degrees of freedom associated with it, and the p-value.

(e) Based on the p-value calculated in part (d), what is the conclusion of the hypothesis test? Interpret your conclusion in this context.

[60] Gallup Wellbeing, Employed Americans in Better Health Than the Unemployed, data collected Jan. 2, 2011 - May 21, 2012.

6.40 Evolution vs. creationism. A Gallup Poll released in December 2010 asked 1019 adults living in the Continental U.S. about their belief in the origin of humans. These results, along with results from a more comprehensive poll from 2001 (that we will assume to be exactly accurate), are summarized in the table below:[61]

	Year	
Response	2010	2001
Humans evolved, with God guiding (1)	38%	37%
Humans evolved, but God had no part in process (2)	16%	12%
God created humans in present form (3)	40%	45%
Other / No opinion (4)	6%	6%

(a) Calculate the actual number of respondents in 2010 that fall in each response category.

(b) State hypotheses for the following research question: have beliefs on the origin of human life changed since 2001?

(c) Calculate the expected number of respondents in each category under the condition that the null hypothesis from part (b) is true.

(d) Conduct a chi-square test and state your conclusion. (Reminder: verify conditions.)

6.7.4 Testing for independence in two-way tables

6.41 Quitters. Does being part of a support group affect the ability of people to quit smoking? A county health department enrolled 300 smokers in a randomized experiment. 150 participants were assigned to a group that used a nicotine patch and met weekly with a support group; the other 150 received the patch and did not meet with a support group. At the end of the study, 40 of the participants in the patch plus support group had quit smoking while only 30 smokers had quit in the other group.

(a) Create a two-way table presenting the results of this study.

(b) Answer each of the following questions under the null hypothesis that being part of a support group does not affect the ability of people to quit smoking, and indicate whether the expected values are higher or lower than the observed values.

i. How many subjects in the "patch + support" group would you expect to quit?

ii. How many subjects in the "only patch" group would you expect to not quit?

6.42 Full body scan, Part II. The table below summarizes a data set we first encountered in Exercise 6.32 regarding views on full-body scans and political affiliation. The differences in each political group may be due to chance. Complete the following computations under the null hypothesis of independence between an individual's party affiliation and his support of full-body scans. It may be useful to first add on an extra column for row totals before proceeding with the computations.

		Party Affiliation		
		Republican	Democrat	Independent
Answer	Should	264	299	351
	Should not	38	55	77
	Don't know/No answer	16	15	22
	Total	318	369	450

(a) How many Republicans would you expect to not support the use of full-body scans?

(b) How many Democrats would you expect to support the use of full-body scans?

(c) How many Independents would you expect to not know or not answer?

[61] Four in 10 Americans Believe in Strict Creationism, December 17, 2010, http://www.gallup.com/poll/145286/Four-Americans-Believe-Strict-Creationism.aspx.

6.43 Offshore drilling, Part III. The table below summarizes a data set we first encountered in Exercise 6.29 that examines the responses of a random sample of college graduates and non-graduates on the topic of oil drilling. Complete a chi-square test for these data to check whether there is a statistically significant difference in responses from college graduates and non-graduates.

	College Grad	
	Yes	No
Support	154	132
Oppose	180	126
Do not know	104	131
Total	438	389

6.44 Coffee and Depression. Researchers conducted a study investigating the relationship between caffeinated coffee consumption and risk of depression in women. They collected data on 50,739 women free of depression symptoms at the start of the study in the year 1996, and these women were followed through 2006. The researchers used questionnaires to collect data on caffeinated coffee consumption, asked each individual about physician-diagnosed depression, and also asked about the use of antidepressants. The table below shows the distribution of incidences of depression by amount of caffeinated coffee consumption.[62]

		Caffeinated coffee consumption					
		≤ 1	2-6	1	2-3	≥ 4	
		cup/week	cups/week	cup/day	cups/day	cups/day	Total
Clinical	Yes	670	$\boxed{373}$	905	564	95	2,607
depression	No	11,545	6,244	16,329	11,726	2,288	48,132
	Total	12,215	6,617	17,234	12,290	2,383	50,739

(a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression?

(b) Write the hypotheses for the test you identified in part (a).

(c) Calculate the overall proportion of women who do and do not suffer from depression.

(d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. $(Observed - Expected)^2/Expected$.

(e) The test statistic is $X^2 = 20.93$. What is the p-value?

(f) What is the conclusion of the hypothesis test?

(g) One of the authors of this study was quoted on the NYTimes as saying it was "too early to recommend that women load up on extra coffee" based on just this study.[63] Do you agree with this statement? Explain your reasoning.

[62] M. Lucas et al. "Coffee, caffeine, and risk of depression among women". In: *Archives of internal medicine* 171.17 (2011), p. 1571.

[63] A. O'Connor. "Coffee Drinking Linked to Less Depression in Women". In: *New York Times* (2011).

6.45 Privacy on Facebook. A 2011 survey asked 806 randomly sampled adult Facebook users about their Facebook privacy settings. One of the questions on the survey was, "Do you know how to adjust your Facebook privacy settings to control what people can and cannot see?" The responses are cross-tabulated based on gender.[64]

		Gender		
		Male	Female	Total
Response	Yes	288	378	666
	No	61	62	123
	Not sure	10	7	17
	Total	359	447	806

(a) State appropriate hypotheses to test for independence of gender and whether or not Facebook users know how to adjust their privacy settings.

(b) Verify any necessary conditions for the test and determine whether or not a chi-square test can be completed.

6.46 Shipping holiday gifts. A December 2010 survey asked 500 randomly sampled Los Angeles residents which shipping carrier they prefer to use for shipping holiday gifts. The table below shows the distribution of responses by age group as well as the expected counts for each cell (shown in parentheses).

		Age						
		18-34		35-54		55+		Total
Shipping Method	USPS	72	(81)	97	(102)	76	(62)	245
	UPS	52	(53)	76	(68)	34	(41)	162
	FedEx	31	(21)	24	(27)	9	(16)	64
	Something else	7	(5)	6	(7)	3	(4)	16
	Not sure	3	(5)	6	(5)	4	(3)	13
	Total	165		209		126		500

(a) State the null and alternative hypotheses for testing for independence of age and preferred shipping method for holiday gifts among Los Angeles residents.

(b) Are the conditions for inference using a chi-square test satisfied?

6.7.5 Small sample hypothesis testing for a proportion

6.47 Bullying in schools. A 2012 Survey USA poll asked Florida residents how big of a problem they thought bullying was in local schools. 9 out of 191 18-34 year olds responded that bullying is no problem at all. Using these data, is it appropriate to construct a confidence interval using the formula $\hat{p} \pm z^{\star}\sqrt{\hat{p}(1-\hat{p})/n}$ for the true proportion of 18-34 year old Floridians who think bullying is no problem at all? If it is appropriate, construct the confidence interval. If it is not, explain why.

[64] Survey USA, News Poll #17960, data collected February 16-17, 2011.

6.48 Choose a test. We would like to test the following hypotheses:

$H_0 : p = 0.1$

$H_A : p \neq 0.1$

The sample size is 120 and the sample proportion is 8.5%. Determine which of the below test(s) is/are appropriate for this situation and explain your reasoning.

I. Z test for a proportion,
i.e. proportion test using normal model

II. Z test for comparing two proportions

III. χ^2 test of independence

IV. Simulation test for a proportion

V. t test for a mean

VI. ANOVA

6.49 The Egyptian Revolution. A popular uprising that started on January 25, 2011 in Egypt led to the 2011 Egyptian Revolution. Polls show that about 69% of American adults followed the news about the political crisis and demonstrations in Egypt closely during the first couple weeks following the start of the uprising. Among a random sample of 30 high school students, it was found that only 17 of them followed the news about Egypt closely during this time.[65]

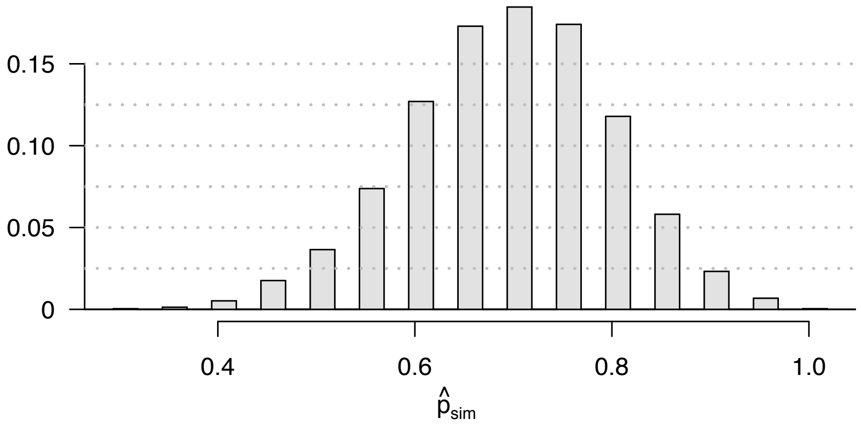
(a) Write the hypotheses for testing if the proportion of high school students who followed the news about Egypt is different than the proportion of American adults who did.

(b) Calculate the proportion of high schoolers in this sample who followed the news about Egypt closely during this time.

(c) Based on large sample theory, we modeled $\hat{p}$ using the normal distribution. Why should we be cautious about this approach for these data?

(d) The normal approximation will not be as reliable as a simulation, especially for a sample of this size. Describe how to perform such a simulation and, once you had results, how to estimate the p-value.

(e) Below is a histogram showing the distribution of $\hat{p}_{sim}$ in 10,000 simulations under the null hypothesis. Estimate the p-value using the plot and determine the conclusion of the hypothesis test.



[65] Gallup Politics, Americans' Views of Egypt Sharply More Negative, data collected February 2-5, 2011.

6.50 Assisted Reproduction. Assisted Reproductive Technology (ART) is a collection of techniques that help facilitate pregnancy (e.g. in vitro fertilization). A 2008 report by the Centers for Disease Control and Prevention estimated that ART has been successful in leading to a live birth in 31% of cases[66]. A new fertility clinic claims that their success rate is higher than average. A random sample of 30 of their patients yielded a success rate of 40%. A consumer watchdog group would like to determine if this provides strong evidence to support the company's claim.

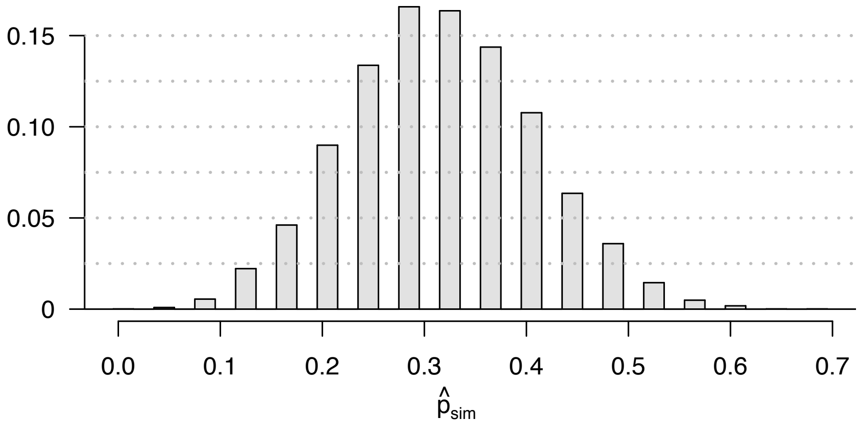
(a) Write the hypotheses to test if the success rate for ART at this clinic is significantly higher than the success rate reported by the CDC.

(b) Based on large sample theory, we modeled $\hat{p}$ using the normal distribution. Why is this not appropriate here?

(c) The normal approximation would be less reliable here, so we should use a simulation strategy. Describe a setup for a simulation that would be appropriate in this situation and how the p-value can be calculated using the simulation results.

(d) Below is a histogram showing the distribution of $\hat{p}_{sim}$ in 10,000 simulations under the null hypothesis. Estimate the p-value using the plot and use it to evaluate the hypotheses.

(e) After performing this analysis, the consumer group releases the following news headline: "Infertility clinic falsely advertises better success rates". Comment on the appropriateness of this statement.



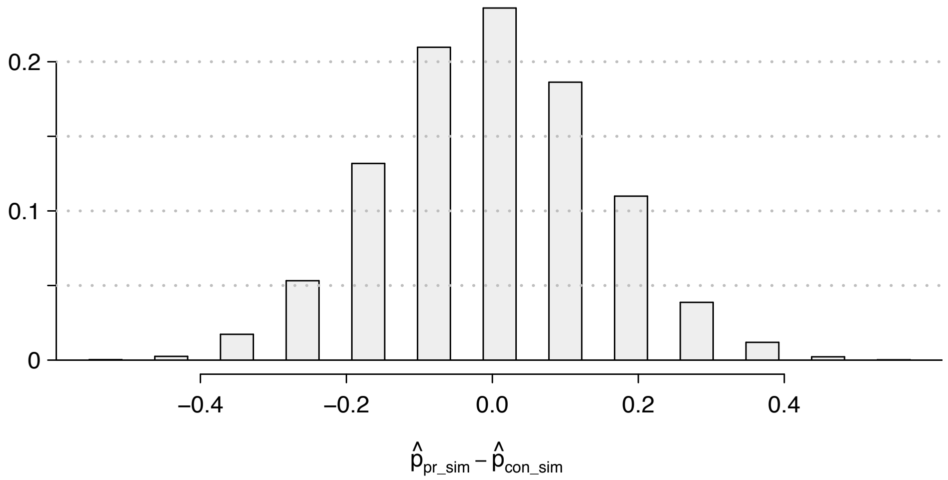
[66] CDC. *2008 Assisted Reproductive Technology Report.*

6.7.6 Hypothesis testing for two proportions

6.51 Social experiment, Part II. Exercise 6.23 introduces a "social experiment" conducted by a TV program that questioned what people do when they see a very obviously bruised woman getting picked on by her boyfriend. On two different occasions at the same restaurant, the same couple was depicted. In one scenario the woman was dressed "provocatively" and in the other scenario the woman was dressed "conservatively". The table below shows how many restaurant diners were present under each scenario, and whether or not they intervened.

		Scenario		
		Provocative	Conservative	Total
Intervene	Yes	5	15	20
	No	15	10	25
	Total	20	25	45

A simulation was conducted to test if people react differently under the two scenarios. 10,000 simulated differences were generated to construct the null distribution shown. The value $\hat{p}_{pr,sim}$ represents the proportion of diners who intervened in the simulation for the provocatively dressed woman, and $\hat{p}_{con,sim}$ is the proportion for the conservatively dressed woman.



(a) What are the hypotheses? For the purposes of this exercise, you may assume that each observed person at the restaurant behaved independently, though we would want to evaluate this assumption more rigorously if we were reporting these results.

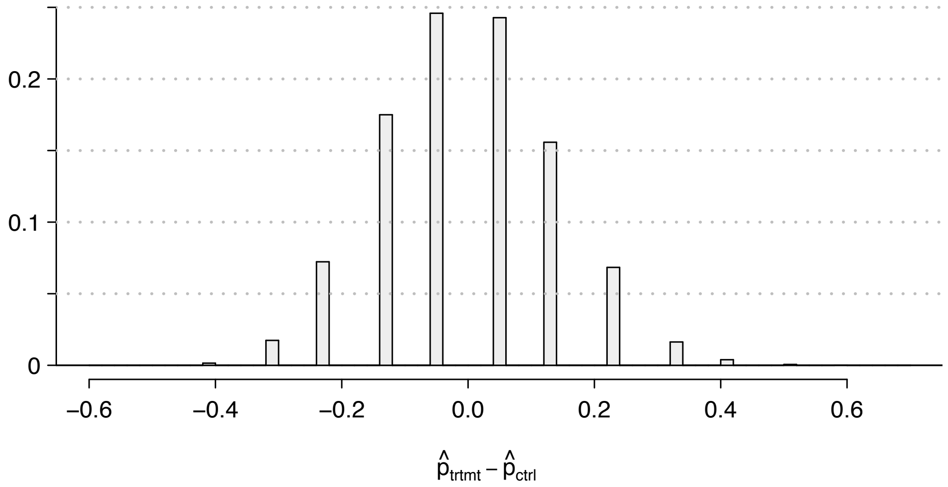
(b) Calculate the observed difference between the rates of intervention under the provocative and conservative scenarios: $\hat{p}_{pr} - \hat{p}_{con}$.

(c) Estimate the p-value using the figure above and determine the conclusion of the hypothesis test.

6.52 Is yawning contagious? An experiment conducted by the *MythBusters*, a science entertainment TV program on the Discovery Channel, tested if a person can be subconsciously influenced into yawning if another person near them yawns. 50 people were randomly assigned to two groups: 34 to a group where a person near them yawned (treatment) and 16 to a group where there wasn't a person yawning near them (control). The following table shows the results of this experiment.[67]

		Group		
		Treatment	Control	Total
Result	Yawn	10	4	14
	Not Yawn	24	12	36
	Total	34	16	50

A simulation was conducted to understand the distribution of the test statistic under the assumption of independence: having someone yawn near another person has no influence on if the other person will yawn. In order to conduct the simulation, a researcher wrote yawn on 14 index cards and not yawn on 36 index cards to indicate whether or not a person yawned. Then he shuffled the cards and dealt them into two groups of size 34 and 16 for treatment and control, respectively. He counted how many participants in each simulated group yawned in an apparent response to a nearby yawning person, and calculated the difference between the simulated proportions of yawning as $\hat{p}_{trtmt,sim} - \hat{p}_{ctrl,sim}$. This simulation was repeated 10,000 times using software to obtain 10,000 differences that are due to chance alone. The histogram shows the distribution of the simulated differences.



(a) What are the hypotheses?

(b) Calculate the observed difference between the yawning rates under the two scenarios.

(c) Estimate the p-value using the figure above and determine the conclusion of the hypothesis test.

[67] MythBusters, Season 3, Episode 28.