Chapter 6

Inference for categorical data

In this chapter, we apply the methods and ideas from Chapter 5 in several contexts for categorical data. We'll start by revisiting what we learned for a single proportion, where the normal distribution can be used to model the uncertainty in the sample proportion. Next, we apply these same ideas to analyze the difference of two proportions using the normal model. Later in the chapter, we apply inference techniques to contingency tables; while we will use a different distribution in this context, the core ideas of hypothesis testing remain the same.

For videos, slides, and other resources, please visit
www.openintro.org/os

6.1 Inference for a single proportion

We encountered inference methods for a single proportion in Chapter 5, exploring point estimates, confidence intervals, and hypothesis tests. In this section, we'll do a review of these topics and also how to choose an appropriate sample size when collecting data for single proportion contexts.

6.1.1 Identifying when the sample proportion is nearly normal

A sample proportion $\hat{p}$ can be modeled using a normal distribution when the sample observations are independent and the sample size is sufficiently large.

SAMPLING DISTRIBUTION OF $\hat{p}$

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

1. The sample's observations are independent, e.g. are from a simple random sample.
2. We expected to see at least 10 successes and 10 failures in the sample, i.e. $np \geq 10$ and $n(1-p) \geq 10$. This is called the **success-failure condition**.

When these conditions are met, then the sampling distribution of $\hat{p}$ is nearly normal with mean p and standard error $SE = \sqrt{\frac{p(1-p)}{n}}$.

Typically we don't know the true proportion p, so we substitute some value to check conditions and estimate the standard error. For confidence intervals, the sample proportion $\hat{p}$ 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.

6.1.2 Confidence intervals for a proportion

A confidence interval provides a range of plausible values for the parameter p, and when $\hat{p}$ can be modeled using a normal distribution, the confidence interval for p takes the form

$$\hat{p} \pm z^\star \times SE$$

EXAMPLE 6.1

A simple random sample of 826 payday loan borrowers was surveyed to better understand their interests around regulation and costs. 70% of the responses supported new regulations on payday lenders. Is it reasonable to model $\hat{p} = 0.70$ using a normal distribution?

(E) The data are a random sample, so the observations are independent and representative of the population of interest.

We also must check the success-failure condition, which we do using $\hat{p}$ in place of p when computing a confidence interval:

Support: $np \approx 826 \times 0.70 = 578$ Not: $n(1-p) \approx 826 \times (1-0.70) = 248$

Since both values are at least 10, we can use the normal distribution to model $\hat{p}$.

GUIDED PRACTICE 6.2

G Estimate the standard error of $\hat{p} = 0.70$. Because p is unknown and the standard error is for a confidence interval, use $\hat{p}$ in place of p in the formula.[1]

EXAMPLE 6.3

Construct a 95% confidence interval for p, the proportion of payday borrowers who support increased regulation for payday lenders.

E Using the point estimate 0.70, $z^\star = 1.96$ for a 95% confidence interval, and the standard error $SE = 0.016$ from Guided Practice 6.2, the confidence interval is

$$\text{point estimate} \pm z^\star \times SE \quad \rightarrow \quad 0.70 \pm 1.96 \times 0.016 \quad \rightarrow \quad (0.669, 0.731)$$

We are 95% confident that the true proportion of payday borrowers who supported regulation at the time of the poll was between 0.669 and 0.731.

CONFIDENCE INTERVAL FOR A SINGLE PROPORTION

Once you've determined a one-proportion confidence interval would be helpful for an application, there are four steps to constructing the interval:

Prepare. Identify $\hat{p}$ and n, and determine what confidence level you wish to use.

Check. Verify the conditions to ensure $\hat{p}$ is nearly normal. For one-proportion confidence intervals, use $\hat{p}$ in place of p to check the success-failure condition.

Calculate. If the conditions hold, compute SE using $\hat{p}$, find $z^\star$, and construct the interval.

Conclude. Interpret the confidence interval in the context of the problem.

For additional one-proportion confidence interval examples, see Section 5.2.

6.1.3 Hypothesis testing for a proportion

One possible regulation for payday lenders is that they would be required to do a credit check and evaluate debt payments against the borrower's finances. We would like to know: would borrowers support this form of regulation?

GUIDED PRACTICE 6.4

G Set up hypotheses to evaluate whether borrowers have a majority support or majority opposition for this type of regulation.[2]

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.

[1] $SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{0.70(1-0.70)}{826}} = 0.016$.
[2] H_0: $p = 0.50$. H_A: $p \neq 0.50$.

GUIDED PRACTICE 6.5

G Do payday loan borrowers support a regulation that would require lenders to pull their credit report and evaluate their debt payments? From a random sample of 826 borrowers, 51% said they would support such a regulation. Is it reasonable to model $\hat{p} = 0.51$ using a normal distribution for a hypothesis test here?[3]

EXAMPLE 6.6

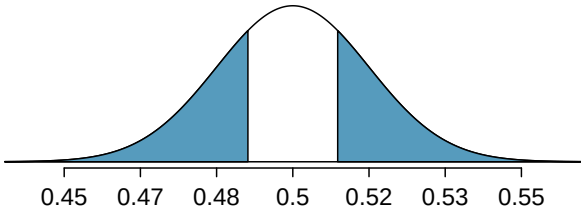
Using the hypotheses and data from Guided Practice 6.4 and 6.5, evaluate whether the poll provides convincing evidence that a majority of payday loan borrowers support a new regulation that would require lenders to pull credit reports and evaluate debt payments.

With hypotheses already set up and conditions checked, we can move onto calculations. The standard error in the context of a one-proportion hypothesis test is computed using the null value, p_0:

$$SE = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{826}} = 0.017$$

A picture of the normal model is shown below with the p-value represented by the shaded region.

E



Based on the normal model, the test statistic can be computed as the Z-score of the point estimate:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.51 - 0.50}{0.017} = 0.59$$

The single tail area is 0.2776, and the p-value, represented by both tail areas together, is 0.5552. Because the p-value is larger than 0.05, we do not reject H_0. The poll does not provide convincing evidence that a majority of payday loan borrowers support or oppose regulations around credit checks and evaluation of debt payments.

HYPOTHESIS TESTING FOR A SINGLE PROPORTION

Once you've determined a one-proportion hypothesis test is the correct procedure, there are four steps to completing the test:

Prepare. Identify the parameter of interest, list hypotheses, identify the significance level, and identify $\hat{p}$ and n.

Check. Verify conditions to ensure $\hat{p}$ is nearly normal under H_0. For one-proportion hypothesis tests, use the null value to check the success-failure condition.

Calculate. If the conditions hold, compute the standard error, again using p_0, compute the Z-score, and identify the p-value.

Conclude. Evaluate the hypothesis test by comparing the p-value to α, and provide a conclusion in the context of the problem.

For additional one-proportion hypothesis test examples, see Section 5.3.

[3] Independence holds since the poll is based on a random sample. The success-failure condition also holds, which is checked using the null value ($p_0 = 0.5$) from H_0: $np_0 = 826 \times 0.5 = 413$, $n(1-p_0) = 826 \times 0.5 = 413$.

6.1.4 When one or more conditions aren't met

We've spent a lot of time discussing conditions for when $\hat{p}$ can be reasonably modeled by a normal distribution. What happens when the success-failure condition fails? What about when the independence condition fails? In either case, the general ideas of confidence intervals and hypothesis tests remain the same, but the strategy or technique used to generate the interval or p-value change.

When the success-failure condition isn't met for a hypothesis test, we can simulate the null distribution of $\hat{p}$ using the null value, p_0. The simulation concept is similar to the ideas used in the malaria case study presented in Section 2.3, and an online section outlines this strategy:

www.openintro.org/r?go=stat_sim_prop_ht

For a confidence interval when the success-failure condition isn't met, we can use what's called the **Clopper-Pearson interval**. The details are beyond the scope of this book. However, there are many internet resources covering this topic.

The independence condition is a more nuanced requirement. When it isn't met, it is important to understand how and why it isn't met. For example, if we took a cluster sample (see Section 1.3), suitable statistical methods are available but would be beyond the scope of even most second or third courses in statistics. On the other hand, we'd be stretched to find any method that we could confidently apply to correct the inherent biases of data from a convenience sample.

While this book is scoped to well-constrained statistical problems, do remember that this is just the first book in what is a large library of statistical methods that are suitable for a very wide range of data and contexts.

6.1.5 Choosing a sample size when estimating a proportion

When collecting data, we choose a sample size suitable for the purpose of the study. Often times this means choosing a sample size large enough that the **margin of error** – which is the part we add and subtract from the point estimate in a confidence interval – is sufficiently small that the sample is useful. For example, our task might be to find a sample size n so that the sample proportion is within ± 0.04 of the actual proportion in a 95% confidence interval.

EXAMPLE 6.7

A university newspaper is conducting a survey to determine what fraction of students support a \$200 per year increase in fees to pay for a new football stadium. How big of a sample is required to ensure the margin of error is smaller than 0.04 using a 95% confidence level?

The margin of error for a sample proportion is

$$z^\star \sqrt{\frac{p(1-p)}{n}}$$

Our goal is to find the smallest sample size n so that this margin of error is smaller than 0.04. For a 95% confidence level, the value $z^\star$ corresponds to 1.96:

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} < 0.04$$

(E)

There are two unknowns in the equation: p and n. If we have an estimate of p, perhaps from a prior survey, we could enter in that value and solve for n. 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 value* if no estimate of the proportion is available:

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

We would need over 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.

When an estimate of the proportion is available, we use it in place of the worst case proportion value, 0.5.

GUIDED PRACTICE 6.8

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 1% with a 90% confidence level. There are three different failure rates to choose from. Perform the sample size computation for each separately, and identify three sample sizes to consider.[4]

EXAMPLE 6.9

The sample sizes vary widely in Guided Practice 6.8. Which of the three would you suggest using? What would influence your choice?

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 might consider the value corresponding to the larger sample. There are also other reasonable approaches.

Also observe that the success-failure condition would need to be checked in the final sample. For instance, if we sampled $n = 1584$ tires and found a failure rate of 0.5%, the normal approximation would not be reasonable, and we would require more advanced statistical methods for creating the confidence interval.

GUIDED PRACTICE 6.10

Suppose we want to continually track the support of payday borrowers for regulation on lenders, where we would conduct a new poll every month. Running such frequent polls is expensive, so we decide a wider margin of error of 5% for each individual survey would be acceptable. Based on the original sample of borrowers where 70% supported some form of regulation, how big should our monthly sample be for a margin of error of 0.05 with 95% confidence?[5]

---

[4] For a 90% confidence interval, $z^\star = 1.6449$, and since an estimate of the proportion 0.017 is available, we'll use it in the margin of error formula:

$$1.6449 \times \sqrt{\frac{0.017(1-0.017)}{n}} < 0.01 \quad \rightarrow \quad \frac{0.017(1-0.017)}{n} < \left(\frac{0.01}{1.6449}\right)^2 \quad \rightarrow \quad 452.15 < n$$

For sample size calculations, we always round up, so the first tire model suggests 453 tires would be sufficient.

A similar computation can be accomplished using 0.062 and 0.013 for p, and you should verify that using these proportions results in minimum sample sizes of 1574 and 348 tires, respectively.

[5] We complete the same computations as before, except now we use 0.70 instead of 0.5 for p:

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.96 \times \sqrt{\frac{0.70(1-0.70)}{n}} \leq 0.05 \quad \rightarrow \quad n \geq 322.7$$

A sample size of 323 or more would be reasonable. (Reminder: always round up for sample size calculations!) Given that we plan to track this poll over time, we also may want to periodically repeat these calculations to ensure that we're being thoughtful in our sample size recommendations in case the baseline rate fluctuates.

Exercises

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.[6]

(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.[7]

(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 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.

[6] A. Vaughn. "Poll finds young adults optimistic, but not about money". In: *Los Angeles Times* (2011).
[7] Demos.org. "The State of Young America: The Poll". In: (2011).

6.5 Gender equality. The General Social Survey asked a random sample of 1,390 Americans the following question: "On the whole, do you think it should or should not be the government's responsibility to promote equality between men and women?" 82% of the respondents said it "should be". At a 95% confidence level, this sample has 2% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.[8]

(a) We are 95% confident that between 80% and 84% of Americans in this sample think it's the government's responsibility to promote equality between men and women.

(b) We are 95% confident that between 80% and 84% of all Americans think it's the government's responsibility to promote equality between men and women.

(c) If we considered many random samples of 1,390 Americans, and we calculated 95% confidence intervals for each, 95% of these intervals would include the true population proportion of Americans who think it's the government's responsibility to promote equality between men and women.

(d) In order to decrease the margin of error to 1%, 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 Americans think it's the government's responsibility to promote equality between men and women.

6.6 Elderly drivers. 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.[9]

(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?

6.7 Fireworks on July 4th. A local news outlet reported that 56% of 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.[10]

6.8 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".[11]

(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.

6.9 Study abroad. A survey on 1,509 high school seniors who took the SAT and who completed an optional web survey shows that 55% of high school seniors are fairly certain that they will participate in a study abroad program in college.[12]

(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?

[8] National Opinion Research Center, General Social Survey, 2018.
[9] Marist Poll, Road Rules: Re-Testing Drivers at Age 65?, March 4, 2011.
[10] Survey USA, News Poll #19333, data collected on June 27, 2012.
[11] Gallup World, More Than One in 10 "Suffering" Worldwide, data collected throughout 2011.
[12] studentPOLL, College-Bound Students' Interests in Study Abroad and Other International Learning Activities, January 2008.

6.10 Legalization of marijuana, Part I. The General Social Survey asked 1,578 US residents: "Do you think the use of marijuana should be made legal, or not?" 61% of the respondents said it should be made legal.[13]

(a) Is 61% 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?

6.11 National Health Plan, Part I. A *Kaiser Family Foundation* poll for US adults in 2019 found that 79% of Democrats, 55% of Independents, and 24% of Republicans supported a generic "National Health Plan". There were 347 Democrats, 298 Republicans, and 617 Independents surveyed.[14]

(a) A political pundit on TV claims that a majority of Independents support a National Health Plan. Do these data provide strong evidence to support this type of statement?

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

6.12 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.[15]

(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.

6.13 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 any better or worse than random guessing at telling the difference between diet and regular soda?

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

6.14 Is college worth it? Part II. Exercise 6.12 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?

6.15 National Health Plan, Part II. Exercise 6.11 presents the results of a poll evaluating support for a generic "National Health Plan" in the US in 2019, reporting that 55% of Independents are supportive. If we wanted to estimate this number to within 1% with 90% confidence, what would be an appropriate sample size?

6.16 Legalize Marijuana, Part II. As discussed in Exercise 6.10, the General Social Survey reported a sample where about 61% 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?

[13] National Opinion Research Center, General Social Survey, 2018.

[14] Kaiser Family Foundation, The Public On Next Steps For The ACA And Proposals To Expand Coverage, data collected between Jan 9-14, 2019.

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

6.2 Difference of two proportions

We would like to extend the methods from Section 6.1 to apply confidence intervals and hypothesis tests to differences in population proportions: $p_1 - p_2$. In our investigations, we'll identify a reasonable point estimate of $p_1 - p_2$ based on the sample, and you may have already guessed its form: $\hat{p}_1 - \hat{p}_2$. Next, we'll apply the same processes we used in the single-proportion context: we verify that the point estimate can be modeled using a normal distribution, we compute the estimate's standard error, and we apply our inferential framework.

6.2.1 Sampling distribution of the difference of two proportions

Like with $\hat{p}$, the difference of two sample proportions $\hat{p}_1 - \hat{p}_2$ can be modeled using a normal distribution when certain conditions are met. First, we require a broader independence condition, and secondly, the success-failure condition must be met by both groups.

CONDITIONS FOR THE SAMPLING DISTRIBUTION OF $\hat{p}_1 - \hat{p}_2$ TO BE NORMAL

The difference $\hat{p}_1 - \hat{p}_2$ can be modeled using a normal distribution when

- *Independence, extended.* The data are independent within and between the two groups. Generally this is satisfied if the data come from two independent random samples or if the data come from a randomized experiment.
- *Success-failure condition.* The success-failure condition holds for both groups, where we check successes and failures in each group separately.

When these conditions are satisfied, the standard error of $\hat{p}_1 - \hat{p}_2$ is

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

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

6.2.2 Confidence intervals for $p_1 - p_2$

We can apply the generic confidence interval formula for a difference of two proportions, where we use $\hat{p}_1 - \hat{p}_2$ as the point estimate and substitute the SE formula:

$$\text{point estimate} \pm z^\star \times SE \qquad \rightarrow \qquad \hat{p}_1 - \hat{p}_2 \pm z^\star \times \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

We can also follow the same Prepare, Check, Calculate, Conclude steps for computing a confidence interval or completing a hypothesis test. The details change a little, but the general approach remain the same. Think about these steps when you apply statistical methods.

EXAMPLE 6.11

We consider an experiment for patients who underwent cardiopulmonary resuscitation (CPR) for a heart attack and were subsequently admitted to a hospital. 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. The results are shown in Figure 6.1. Check whether we can model the difference in sample proportions using the normal distribution.

(E)

We first check for independence: since this is a randomized experiment, this condition is satisfied.

Next, we check the success-failure condition for each group. We have at least 10 successes and 10 failures in each experiment arm (11, 14, 39, 26), so this condition is also satisfied.

With both conditions satisfied, the difference in sample proportions can be reasonably modeled using a normal distribution for these data.

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

Figure 6.1: Results for the CPR study. Patients in the treatment group were given a blood thinner, and patients in the control group were not.

EXAMPLE 6.12

Create and interpret a 90% confidence interval of the difference for the survival rates in the CPR study.

We'll use p_t for the survival rate in the treatment group and p_c for the control group:

$$\hat{p}_t - \hat{p}_c = \frac{14}{40} - \frac{11}{50} = 0.35 - 0.22 = 0.13$$

We use the standard error formula provided on page 217. As with the one-sample proportion case, we use the sample estimates of each proportion in the formula in the confidence interval context:

(E)

$$SE \approx \sqrt{\frac{0.35(1-0.35)}{40} + \frac{0.22(1-0.22)}{50}} = 0.095$$

For a 90% confidence interval, we use $z^\star = 1.6449$:

$$\text{point estimate} \ \pm \ z^\star \times SE \quad \rightarrow \quad 0.13 \ \pm \ 1.6449 \times 0.095 \quad \rightarrow \quad (-0.026, 0.286)$$

We are 90% confident that blood thinners have a difference of -2.6% to +28.6% percentage point impact on survival rate for patients who are like those in the study. Because 0% is contained in the interval, we do not have enough information to say whether blood thinners help or harm heart attack patients who have been admitted after they have undergone CPR.

GUIDED PRACTICE 6.13

A 5-year experiment was conducted to evaluate the effectiveness of fish oils on reducing cardiovascular events, where each subject was randomized into one of two treatment groups. We'll consider heart attack outcomes in these patients:

	heart attack	no event	Total
fish oil	145	12788	12933
placebo	200	12738	12938

Create a 95% confidence interval for the effect of fish oils on heart attacks for patients who are well-represented by those in the study. Also interpret the interval in the context of the study.[16]

6.2.3 Hypothesis tests for the difference of two proportions

A mammogram is an X-ray procedure used to check for breast cancer. Whether mammograms should be used is part of a controversial discussion, and it's the topic of our next example where we learn about 2-proportion hypothesis tests when H_0 is $p_1 - p_2 = 0$ (or equivalently, $p_1 = p_2$).

A 30-year study was conducted with nearly 90,000 female participants. During a 5-year screening period, each woman was randomized to one of two groups: in the first group, women received regular mammograms to screen for breast cancer, and in the second group, women received regular non-mammogram breast cancer exams. No intervention was made during the following 25 years of the study, and we'll consider death resulting from breast cancer over the full 30-year period. Results from the study are summarized in Figure 6.2.

If mammograms are much more effective than non-mammogram breast cancer exams, then we would expect to see additional deaths from breast cancer in the control group. On the other hand, if mammograms are not as effective as regular breast cancer exams, we would expect to see an increase in breast cancer deaths in the mammogram group.

	Death from breast cancer?	
	Yes	No
Mammogram	500	44,425
Control	505	44,405

Figure 6.2: Summary results for breast cancer study.

GUIDED PRACTICE 6.14

Is this study an experiment or an observational study?[17]

[16] Because the patients were randomized, the subjects are independent, both within and between the two groups. The success-failure condition is also met for both groups as all counts are at least 10. This satisfies the conditions necessary to model the difference in proportions using a normal distribution.

Compute the sample proportions ($\hat{p}_{\text{fish oil}} = 0.0112$, $\hat{p}_{\text{placebo}} = 0.0155$), point estimate of the difference ($0.0112 - 0.0155 = -0.0043$), and standard error ($SE = \sqrt{\frac{0.0112 \times 0.9888}{12933} + \frac{0.0155 \times 0.9845}{12938}} = 0.00145$). Next, plug the values into the general formula for a confidence interval, where we'll use a 95% confidence level with $z^\star = 1.96$:

$$-0.0043 \pm 1.96 \times 0.00145 \quad \rightarrow \quad (-0.0071, -0.0015)$$

We are 95% confident that fish oils decreases heart attacks by 0.15 to 0.71 percentage points (off of a baseline of about 1.55%) over a 5-year period for subjects who are similar to those in the study. Because the interval is entirely below 0, the data provide strong evidence that fish oil supplements reduce heart attacks in patients like those in the study.

[17] This is an experiment. Patients were randomized to receive mammograms or a standard breast cancer exam. We will be able to make causal conclusions based on this study.

GUIDED PRACTICE 6.15

G Set up hypotheses to test whether there was a difference in breast cancer deaths in the mammogram and control groups.[18]

In Example 6.16, we will check the conditions for using a normal distribution to analyze the results of the study. The details are very similar to that of confidence intervals. However, when the null hypothesis is that $p_1 - p_2 = 0$, we use a special proportion called the **pooled proportion** to check the success-failure condition:

$$\begin{aligned}\hat{p}_{pooled} &= \frac{\text{\# of patients who died from breast cancer in the entire study}}{\text{\# of patients in the entire study}} \\ &= \frac{500 + 505}{500 + 44{,}425 + 505 + 44{,}405} \\ &= 0.0112\end{aligned}$$

This proportion is an estimate of the breast cancer death rate across the entire study, and it's our best estimate of the proportions p_{mgm} and p_{ctrl} *if the null hypothesis is true that* $p_{mgm} = p_{ctrl}$. We will also use this pooled proportion when computing the standard error.

EXAMPLE 6.16

Is it reasonable to model the difference in proportions using a normal distribution in this study?

Because the patients are randomized, they can be treated as independent, both within and between groups. We also must check the success-failure condition for each group. Under the null hypothesis, the proportions p_{mgm} and p_{ctrl} are equal, so we check the success-failure condition with our best estimate of these values under H_0, the pooled proportion from the two samples, $\hat{p}_{pooled} = 0.0112$:

E

$$\begin{aligned}\hat{p}_{pooled} \times n_{mgm} &= 0.0112 \times 44{,}925 = 503 \qquad (1 - \hat{p}_{pooled}) \times n_{mgm} = 0.9888 \times 44{,}925 = 44{,}422 \\ \hat{p}_{pooled} \times n_{ctrl} &= 0.0112 \times 44{,}910 = 503 \qquad (1 - \hat{p}_{pooled}) \times n_{ctrl} = 0.9888 \times 44{,}910 = 44{,}407\end{aligned}$$

The success-failure condition is satisfied since all values are at least 10. With both conditions satisfied, we can safely model the difference in proportions using a normal distribution.

USE THE POOLED PROPORTION WHEN H_0 IS $p_1 - p_2 = 0$

When the null hypothesis is that the proportions are equal, use the pooled proportion ($\hat{p}_{pooled}$) to verify the success-failure condition and estimate the standard error:

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

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.

In Example 6.16, the pooled proportion was used to check the success-failure condition.[19] In the next example, we see the second place where the pooled proportion comes into play: the standard error calculation.

[18] H_0: the breast cancer death rate for patients screened using mammograms is the same as the breast cancer death rate for patients in the control, $p_{mgm} - p_{ctrl} = 0$.
H_A: the breast cancer death rate for patients screened using mammograms is different than the breast cancer death rate for patients in the control, $p_{mgm} - p_{ctrl} \neq 0$.

[19] For an example of a two-proportion hypothesis test that does not require the success-failure condition to be met, see Section 2.3.

EXAMPLE 6.17

Compute the point estimate of the difference in breast cancer death rates in the two groups, and use the pooled proportion $\hat{p}_{pooled} = 0.0112$ to calculate the standard error.

The point estimate of the difference in breast cancer death rates is

$$\begin{aligned}\hat{p}_{mgm} - \hat{p}_{ctrl} &= \frac{500}{500 + 44,425} - \frac{505}{505 + 44,405} \\ &= 0.01113 - 0.01125 \\ &= -0.00012\end{aligned}$$

The breast cancer death rate in the mammogram group was 0.012% less than in the control group. Next, the standard error is calculated *using the pooled proportion*, $\hat{p}_{pooled}$:

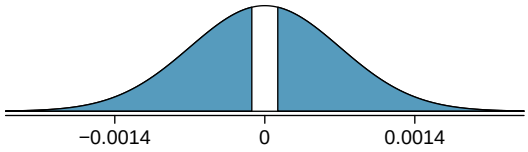
$$SE = \sqrt{\frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_{mgm}} + \frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_{ctrl}}} = 0.00070$$

EXAMPLE 6.18

Using the point estimate $\hat{p}_{mgm} - \hat{p}_{ctrl} = -0.00012$ and standard error $SE = 0.00070$, calculate a p-value for the hypothesis test and write a conclusion.

Just like in past tests, we first compute a test statistic and draw a picture:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{-0.00012 - 0}{0.00070} = -0.17$$



The lower tail area is 0.4325, which we double to get the p-value: 0.8650. Because this p-value is larger than 0.05, we do not reject the null hypothesis. That is, the difference in breast cancer death rates is reasonably explained by chance, and we do not observe benefits or harm from mammograms relative to a regular breast exam.

Can we conclude that mammograms have no benefits or harm? Here are a few considerations to keep in mind when reviewing the mammogram study as well as any other medical study:

- We do not reject the null hypothesis, which means we don't have sufficient evidence to conclude that mammograms reduce or increase breast cancer deaths.
- If mammograms are helpful or harmful, the data suggest the effect isn't very large.
- Are mammograms more or less expensive than a non-mammogram breast exam? If one option is much more expensive than the other and doesn't offer clear benefits, then we should lean towards the less expensive option.
- The study's authors also found that mammograms led to overdiagnosis of breast cancer, which means some breast cancers were found (or thought to be found) but that these cancers would not cause symptoms during patients' lifetimes. That is, something else would kill the patient before breast cancer symptoms appeared. This means some patients may have been treated for breast cancer unnecessarily, and this treatment is another cost to consider. It is also important to recognize that overdiagnosis can cause unnecessary physical or emotional harm to patients.

These considerations highlight the complexity around medical care and treatment recommendations. Experts and medical boards who study medical treatments use considerations like those above to provide their best recommendation based on the current evidence.

6.2.4 More on 2-proportion hypothesis tests (special topic)

When we conduct a 2-proportion hypothesis test, usually H_0 is $p_1 - p_2 = 0$. However, there are rare situations where we want to check for some difference in p_1 and p_2 that is some value other than 0. For example, maybe we care about checking a null hypothesis where $p_1 - p_2 = 0.1$. In contexts like these, we generally use $\hat{p}_1$ and $\hat{p}_2$ to check the success-failure condition and construct the standard error.

GUIDED PRACTICE 6.19

G A quadcopter company is considering a new manufacturer for rotor blades. The new manufacturer would be more expensive, but they claim their higher-quality blades are more reliable, with 3% more blades passing inspection than their competitor. Set up appropriate hypotheses for the test.[20]

Figure 6.3: A Phantom quadcopter.

Photo by David J (http://flic.kr/p/oiWLNu). CC-BY 2.0 license.
This photo has been cropped and a border has been added.

[20] H_0: The higher-quality blades will pass inspection 3% more frequently than the standard-quality blades. $p_{highQ} - p_{standard} = 0.03$. H_A: The higher-quality blades will pass inspection some amount different than 3% more often than the standard-quality blades. $p_{highQ} - p_{standard} \neq 0.03$.

EXAMPLE 6.20

The quality control engineer from Guided Practice 6.19 collects a sample of blades, examining 1000 blades from each company, and she finds that 899 blades pass inspection from the current supplier and 958 pass inspection from the prospective supplier. Using these data, evaluate the hypotheses from Guided Practice 6.19 with a significance level of 5%.

First, we check the conditions. The sample is not necessarily random, so to proceed we must assume the blades 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 is computed using the two sample proportions since we do not use a pooled proportion for this context:

(E)

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

In this hypothesis test, because the null is that $p_1 - p_2 = 0.03$, the sample proportions were used for the standard error calculation rather than a pooled proportion.

Next, we compute the test statistic and use it to find the p-value, which is depicted in Figure 6.4.

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

Using a standard normal distribution for this test statistic, we identify the right tail area as 0.006, and we double it to get the p-value: 0.012. We reject the null hypothesis because 0.012 is less than 0.05. Since we observed a larger-than-3% increase in blades that pass inspection, we have statistically significant evidence that the higher-quality blades pass inspection *more than* 3% as often as the currently used blades, exceeding the company's claims.

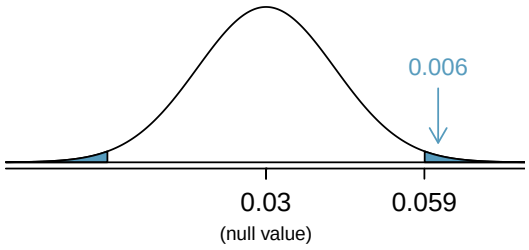


Figure 6.4: Distribution of the test statistic if the null hypothesis was true. The p-value is represented by the shaded areas.

6.2.5 Examining the standard error formula (special topic)

This subsection covers more theoretical topics that offer deeper insights into the origins of the standard error formula for the difference of two proportions. Ultimately, all of the standard error formulas we encounter in this chapter and in Chapter 7 can be derived from the probability principles of Section 3.4.

The formula for the standard error of the difference in two proportions can be deconstructed into the formulas for the standard errors of the individual sample proportions. Recall that the standard error of the individual sample proportions $\hat{p}_1$ and $\hat{p}_2$ are

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

The standard error of the difference of two sample proportions can be deconstructed from the standard errors of the separate sample proportions:

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

This special relationship follows from probability theory.

GUIDED PRACTICE 6.21

Prerequisite: Section 3.4. We can rewrite the equation above in a different way:

(G)

$$SE_{\hat{p}_1-\hat{p}_2}^2 = SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2$$

Explain where this formula comes from using the formula for the variability of the sum of two random variables.[21]

[21] The standard error squared represents the variance of the estimate. If X and Y are two random variables with variances σ_x^2 and σ_y^2, then the variance of $X - Y$ is $\sigma_x^2 + \sigma_y^2$. Likewise, the variance corresponding to $\hat{p}_1 - \hat{p}_2$ is $\sigma_{\hat{p}_1}^2 + \sigma_{\hat{p}_2}^2$. Because $\sigma_{\hat{p}_1}^2$ and $\sigma_{\hat{p}_2}^2$ are just another way of writing $SE_{\hat{p}_1}^2$ and $SE_{\hat{p}_2}^2$, the variance associated with $\hat{p}_1 - \hat{p}_2$ may be written as $SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2$.

Exercises

6.17 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.

6.18 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.[22]

	control	treatment
alive	4	24
dead	30	45

Suppose we are interested in estimating the difference in survival rate between the control and treatment groups using 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?

6.19 Gender and color preference. A 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 about undergraduate college students are true or false, and explain your reasoning for each statement you identify as false.[23]

(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.

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

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

6.20 Government shutdown. The United States federal government shutdown of 2018–2019 occurred from December 22, 2018 until January 25, 2019, a span of 35 days. A Survey USA poll of 614 randomly sampled Americans during this time period reported that 48% of those who make less than $40,000 per year and 55% of those who make $40,000 or more per year said the government shutdown has not at all affected them personally. A 95% confidence interval for $(p_{<40K} - p_{\geq 40K})$, where p is the proportion of those who said the government shutdown has not at all affected them personally, is (-0.16, 0.02). Based on this information, determine if the following statements are true or false, and explain your reasoning if you identify the statement as false.[24]

(a) At the 5% significance level, the data provide convincing evidence of a real difference in the proportion who are not affected personally between Americans who make less than $40,000 annually and Americans who make $40,000 annually.

(b) We are 95% confident that 16% more to 2% fewer Americans who make less than $40,000 per year are not at all personally affected by the government shutdown compared to those who make $40,000 or more per year.

(c) A 90% confidence interval for $(p_{<40K} - p_{\geq 40K})$ would be wider than the $(-0.16, 0.02)$ interval.

(d) A 95% confidence interval for $(p_{\geq 40K} - p_{<40K})$ is (-0.02, 0.16).

6.21 National Health Plan, Part III. Exercise 6.11 presents the results of a poll evaluating support for a generically branded "National Health Plan" in the United States. 79% of 347 Democrats and 55% of 617 Independents support a National Health Plan.

(a) Calculate a 95% confidence interval for the difference between the proportion of Democrats and Independents who support a National Health Plan $(p_D - p_I)$, 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 National Health Plan than the Independent.

6.22 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.[25]

6.23 Offshore drilling, Part I. A 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.[26]

(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

6.24 Sleep deprivation, CA vs. OR, Part II. Exercise 6.22 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?

[24] Survey USA, News Poll #24568, data collected on April 21, 2019.
[25] CDC, Perceived Insufficient Rest or Sleep Among Adults — United States, 2008.
[26] Survey USA, Election Poll #16804, data collected July 8-11, 2010.

6.25 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 6.23.

	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.

6.26 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, 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.[27]

		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) 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 1 or a Type 2 Error? Explain.

6.27 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.[28]

		Transportation Professionals			
	Control	Pilots	Truck Drivers	Train Operators	Bus/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.

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

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

6.28 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).[29]

		Autism		
		Autism	Typical development	Total
Periconceptional prenatal vitamin	No vitamin	111	70	181
	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.[30]

6.29 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.[31]

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

(b) State appropriate hypotheses to test for difference in virologic failure rates between treatment groups.

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

6.30 An apple a day keeps the doctor away. A physical education teacher at a high school wanting to increase awareness on issues of nutrition and health asked her students at the beginning of the semester whether they believed the expression "an apple a day keeps the doctor away", and 40% of the students responded yes. Throughout the semester she started each class with a brief discussion of a study highlighting positive effects of eating more fruits and vegetables. She conducted the same apple-a-day survey at the end of the semester, and this time 60% of the students responded yes. Can she used a two-proportion method from this section for this analysis? Explain your reasoning.

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

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

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

6.3 Testing for goodness of fit using chi-square

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 Figure 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

Figure 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.

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 25 years 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 6.22

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?

(E)

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.

GUIDED PRACTICE 6.23

(G) 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 Figure 6.6.

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

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

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

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.[32] 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:

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

[32] 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.

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:

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

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 a 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.

GUIDED PRACTICE 6.24

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?[33]

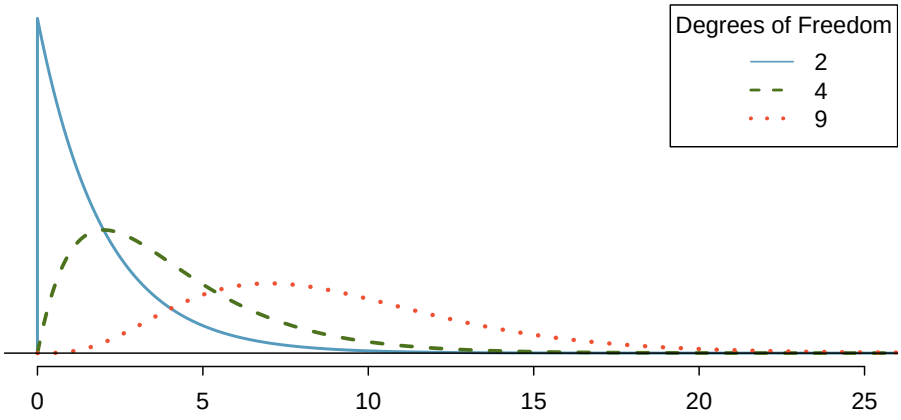


Figure 6.7: Three chi-square distributions with varying degrees of freedom.

[33] (a) The center becomes larger. If took a careful look, we could see that the mean 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.

Figure 6.7 and Guided Practice 6.24 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. The most common ways to do this are using computer software, using a graphing calculator, or using a table. For folks wanting to use the table option, we provide an outline of how to read the chi-square table in Appendix C.3, which is also where you may find the table. For the examples below, use your preferred approach to confirm you get the same answers.

EXAMPLE 6.25

Figure 6.8(a) shows a chi-square distribution with 3 degrees of freedom and an upper shaded tail starting at 6.25. Find the shaded area.

(E)

Using statistical software or a graphing calculator, we can find that the upper tail area for a chi-square distribution with 3 degrees of freedom (df) and a cutoff of 6.25 is 0.1001. That is, the shaded upper tail of Figure 6.8(a) has area 0.1.

EXAMPLE 6.26

Figure 6.8(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. Find the tail area.

(E)

Using software, we can find that the tail area shaded in Figure 6.8(b) to be 0.1165. If using a table, we would only be able to find a range of values for the tail area: between 0.1 and 0.2.

EXAMPLE 6.27

Figure 6.8(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.

(E)

Using software, we would obtain a tail area of 0.4038. If using the table in Appendix C.3, we would have identified that the tail area is larger than 0.3 but not be able to give the precise value.

GUIDED PRACTICE 6.28

(G) Figure 6.8(d) shows a cutoff of 11.7 on a chi-square distribution with 7 degrees of freedom. Find the area of the upper tail.[34]

GUIDED PRACTICE 6.29

(G) Figure 6.8(e) shows a cutoff of 10 on a chi-square distribution with 4 degrees of freedom. Find the area of the upper tail.[35]

GUIDED PRACTICE 6.30

(G) Figure 6.8(f) shows a cutoff of 9.21 with a chi-square distribution with 3 df. Find the area of the upper tail.[36]

[34] The area is 0.1109. If using a table, we would identify that it falls between 0.1 and 0.2.
[35] Precise value: 0.0404. If using the table: between 0.02 and 0.05.
[36] Precise value: 0.0266. If using the table: between 0.02 and 0.05.

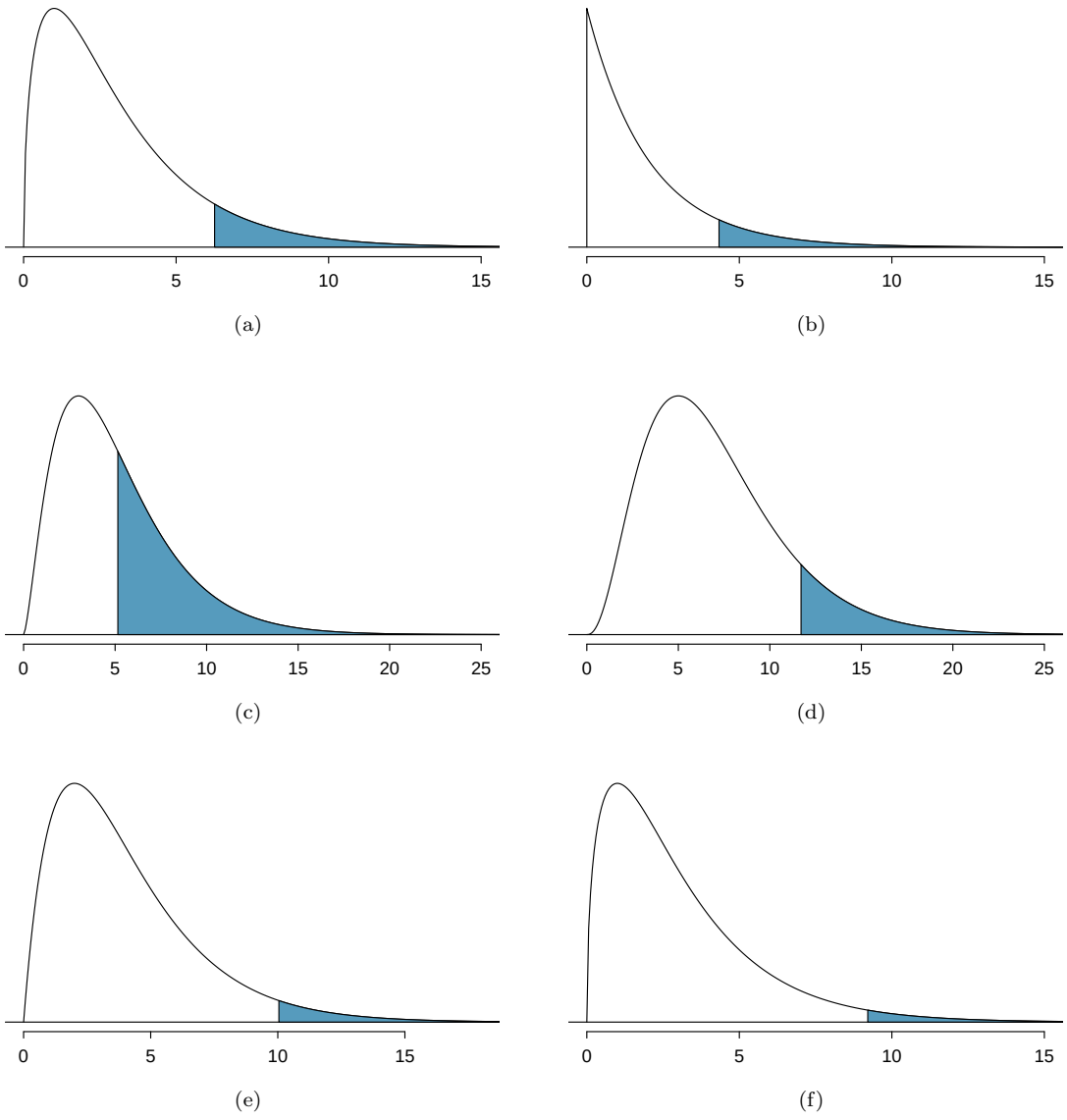


Figure 6.8: **(a)** Chi-square distribution with 3 degrees of freedom, area above 6.25 shaded. **(b)** 2 degrees of freedom, area above 4.3 shaded. **(c)** 5 degrees of freedom, area above 5.1 shaded. **(d)** 7 degrees of freedom, area above 11.7 shaded. **(e)** 4 degrees of freedom, area above 10 shaded. **(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 6.31

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?

(E)

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 a normal distribution 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 6.32

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.

(E)

The chi-square distribution and p-value are shown in Figure 6.9. Because larger chi-square values correspond to stronger evidence against the null hypothesis, we shade the upper tail to represent the p-value. Using statistical software (or the table in Appendix C.3), we can determine that the area is 0.1171. 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.

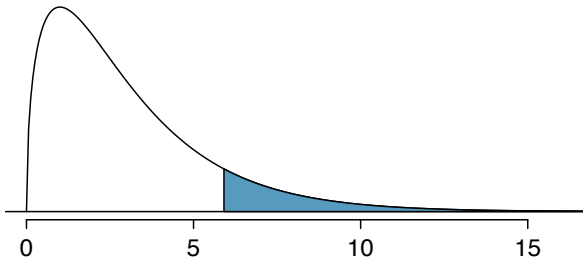


Figure 6.9: The p-value for the juror hypothesis test is shaded in the chi-square distribution with $df = 3$.

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.

CONDITIONS FOR THE CHI-SQUARE TEST

There are two 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.

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

We can apply the 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 10 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. Figure 6.10 shows the number of waiting days for a positive trading day during 10 years for the S&P500.

Days	1	2	3	4	5	6	7+	Total
Observed	717	369	155	69	28	14	10	1362

Figure 6.10: Observed distribution of the waiting time until a positive trading day for the S&P500.

We consider how many days one must wait until observing an `Up` day on the S&P500 stock index. 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 and summarize the waiting times in Figure 6.11 and Figure 6.12. The S&P500 was positive on 54.5% 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 Figure 6.11, can be compared to the expected counts from the *Geometric Model* row. The method for computing expected counts is discussed in Figure 6.11. In general, the expected counts are determined by (1) identifying the null proportion associated with each 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.

Days	1	2	3	4	5	6	7+	Total
Observed	717	369	155	69	28	14	10	1362
Geometric Model	743	338	154	70	32	14	12	1362

Figure 6.11: 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 ($P(D) = (1 - 0.545)^{D-1}(0.545)$) and multiply by the total number of streaks, 1362. For example, waiting for three days occurs under the geometric model about $0.455^2 \times 0.545 = 11.28\%$ of the time, which corresponds to $0.1128 \times 1362 = 154$ streaks.

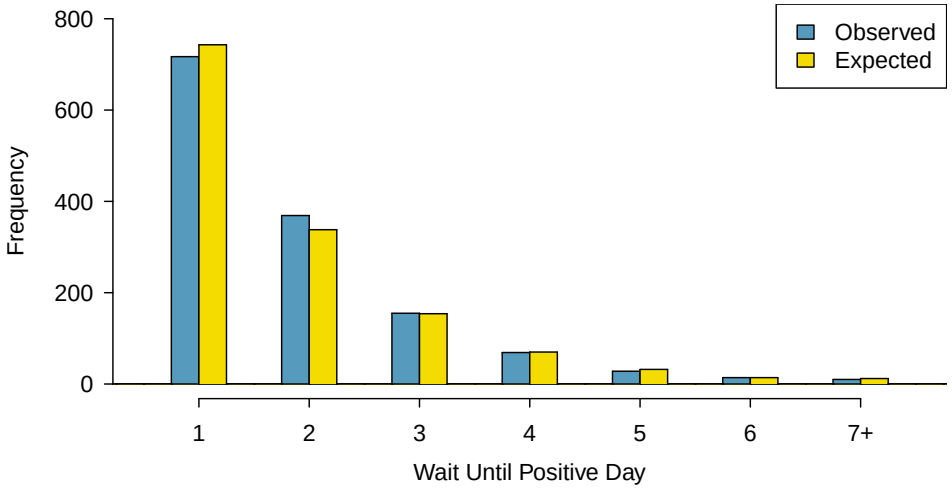


Figure 6.12: Side-by-side bar plot of the observed and expected counts for each waiting time.

EXAMPLE 6.33

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

(E)

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 Figure 6.11.

GUIDED PRACTICE 6.34

(G) Figure 6.11 provides a set of count data for waiting times ($O_1 = 717$, $O_2 = 369$, ...) and expected counts under the geometric distribution ($E_1 = 743$, $E_2 = 338$, ...). Compute the chi-square test statistic, X^2.[37]

GUIDED PRACTICE 6.35

(G) 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?[38]

EXAMPLE 6.36

If the observed counts follow the geometric model, then the chi-square test statistic $X^2 = 4.61$ would closely follow a chi-square distribution with $df = 6$. Using this information, compute a p-value.

(E)

Figure 6.13 shows the chi-square distribution, cutoff, and the shaded p-value. Using software, we can find the p-value: 0.5951. Ultimately, we do not have sufficient evidence to reject the notion that the wait times follow a geometric distribution for the last 10 years of data for the S&P500, i.e. we cannot reject the notion that trading days are independent.

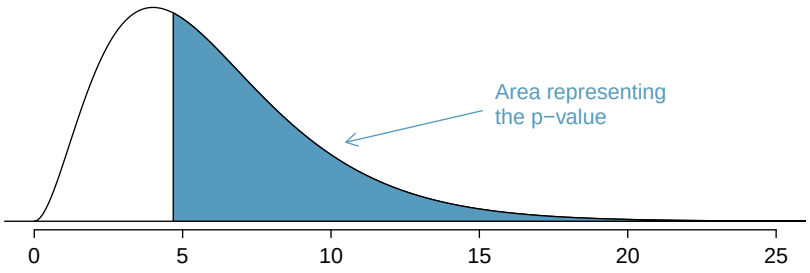


Figure 6.13: Chi-square distribution with 6 degrees of freedom. The p-value for the stock analysis is shaded.

EXAMPLE 6.37

In Example 6.36, we did not reject the null hypothesis that the trading days are independent during the last 10 of data. Why is this so important?

(E)

It may be tempting to think the market is "due" for an `Up` day if there have been several consecutive days where it has been down. However, we haven't found strong evidence that there's any such property where the market is "due" for a correction. At the very least, the analysis suggests any dependence between days is very weak.

[37] $X^2 = \frac{(717-743)^2}{743} + \frac{(369-338)^2}{338} + \cdots + \frac{(10-12)^2}{12} = 4.61$

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

Exercises

6.31 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.

6.32 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 $\chi^2 = 10$ 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.

6.33 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.

6.34 Barking deer. Microhabitat factors associated with forage and bed sites of barking deer in Hainan Island, China were examined. In this region woods make up 4.8% of the land, cultivated grass plot makes up 14.7%, and deciduous forests make up 39.6%. Of the 426 sites where the deer forage, 4 were categorized as woods, 16 as cultivated grassplot, and 61 as deciduous forests. The table below summarizes these data.[39]

Woods	Cultivated grassplot	Deciduous forests	Other	Total
4	16	61	345	426

(a) Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others.

(b) What type of test can we use to answer this research question?

(c) Check if the assumptions and conditions required for this test are satisfied.

(d) Do these data provide convincing evidence that barking deer prefer to forage in certain habitats over others? Conduct an appropriate hypothesis test to answer this research question.

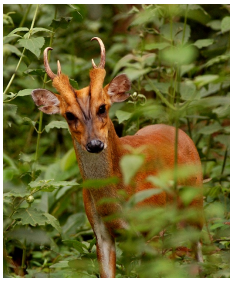


Photo by Shrikant Rao (http://flic.kr/p/4Xjdkk) CC BY 2.0 license

[39] Liwei Teng et al. "Forage and bed sites characteristics of Indian muntjac (Muntiacus muntjak) in Hainan Island, China". In: *Ecological Research* 19.6 (2004), pp. 675–681.

6.4 Testing for independence in two-way tables

We all buy used products – cars, computers, textbooks, and so on – and we sometimes assume the sellers of those products will be forthright about any underlying problems with what they're selling. This is not something we should take for granted. Researchers recruited 219 participants in a study where they would sell a used iPod[40] that was known to have frozen twice in the past. The participants were incentivized to get as much money as they could for the iPod since they would receive a 5% cut of the sale on top of $10 for participating. The researchers wanted to understand what types of questions would elicit the seller to disclose the freezing issue.

Unbeknownst to the participants who were the sellers in the study, the buyers were collaborating with the researchers to evaluate the influence of different questions on the likelihood of getting the sellers to disclose the past issues with the iPod. The scripted buyers started with "Okay, I guess I'm supposed to go first. So you've had the iPod for 2 years ..." and ended with one of three questions:

- General: What can you tell me about it?
- Positive Assumption: It doesn't have any problems, does it?
- Negative Assumption: What problems does it have?

The question is the treatment given to the sellers, and the response is whether the question prompted them to disclose the freezing issue with the iPod. The results are shown in Figure 6.14, and the data suggest that asking the, *What problems does it have?*, was the most effective at getting the seller to disclose the past freezing issues. However, you should also be asking yourself: could we see these results due to chance alone, or is this in fact evidence that some questions are more effective for getting at the truth?

	General	Positive Assumption	Negative Assumption	Total
Disclose Problem	2	23	36	61
Hide Problem	71	50	37	158
Total	73	73	73	219

Figure 6.14: Summary of the iPod study, where a question was posed to the study participant who acted

DIFFERENCES OF ONE-WAY TABLES VS 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 the iPod experiment is really about assessing whether there is statistically significant evidence that the success each question had on getting the participant to disclose the problem with the iPod. In other words, the goal is to check whether the buyer's question was independent of whether the seller disclosed a problem.

[40] For readers not as old as the authors, an iPod is basically an iPhone without any cellular service, assuming it was one of the later generations. Earlier generations were more basic.

6.4.1 Expected counts in two-way tables

Like with one-way tables, we will need to compute estimated counts for each cell in a two-way table.

EXAMPLE 6.38

From the experiment, we can compute the proportion of all sellers who disclosed the freezing problem as $61/219 = 0.2785$. If there really is no difference among the questions and 27.85% of sellers were going to disclose the freezing problem no matter the question that was put to them, how many of the 73 people in the `General` group would we have expected to disclose the freezing problem?

We would predict that $0.2785 \times 73 = 20.33$ sellers would disclose the problem. Obviously we observed fewer than this, though it is not yet clear if that is due to chance variation or whether that is because the questions vary in how effective they are at getting to the truth.

GUIDED PRACTICE 6.39

If the questions were actually equally effective, meaning about 27.85% of respondents would disclose the freezing issue regardless of what question they were asked, about how many sellers would we expect to *hide* the freezing problem from the Positive Assumption group?[41]

We can compute the expected number of sellers who we would expect to disclose or hide the freezing issue for all groups, if the questions had no impact on what they disclosed, using the same strategy employed in Example 6.38 and Guided Practice 6.39. These expected counts were used to construct Figure 6.15, which is the same as Figure 6.14, except now the expected counts have been added in parentheses.

	General	Positive Assumption	Negative Assumption	Total
Disclose Problem	2 **(20.33)**	23 **(20.33)**	36 **(20.33)**	61
Hide Problem	71 **(52.67)**	50 **(52.67)**	37 **(52.67)**	158
Total	73	73	73	219

Figure 6.15: 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 27.85% of each column should be in the first row:

$$0.2785 \times (\text{column 1 total}) = 20.33$$
$$0.2785 \times (\text{column 2 total}) = 20.33$$
$$0.2785 \times (\text{column 3 total}) = 20.33$$

Looking back to how 0.2785 was computed – as the fraction of sellers who disclosed the freezing issue (158/219) – these three expected counts could have been computed as

$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 1 total}) = 20.33$$
$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 2 total}) = 20.33$$
$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 3 total}) = 20.33$$

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.

[41] We would expect $(1 - 0.2785) \times 73 = 52.67$. It is okay that this result, like the result from Example 6.38, 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

General formula	$\frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$
Row 1, Col 1	$\frac{(2-20.33)^2}{20.33} = 16.53$
Row 1, Col 2	$\frac{(23-20.33)^2}{20.33} = 0.35$
⋮	⋮
Row 2, Col 3	$\frac{(37-52.67)^2}{52.67} = 4.66$

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

$$X^2 = 16.53 + 0.35 + \cdots + 4.66 = 40.13$$

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.[42] 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 questions had no impact on the sellers in the experiment), then the test statistic $X^2 = 40.13$ 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 6.16.

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.

When analyzing 2-by-2 contingency tables, one guideline is to use the two-proportion methods introduced in Section 6.2.

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

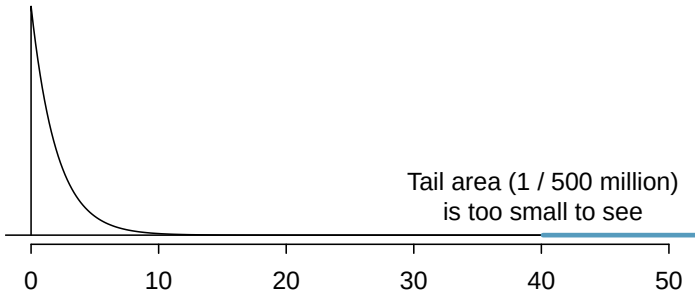


Figure 6.16: Visualization of the p-value for $X^2 = 40.13$ when $df = 2$.

EXAMPLE 6.40

Find the p-value and draw a conclusion about whether the question affects the sellers likelihood of reporting the freezing problem.

(E)

Using a computer, we can compute a very precise value for the tail area above $X^2 = 40.13$ for a chi-square distribution with 2 degrees of freedom: 0.000000002. (If using the table in Appendix C.3, we would identify the p-value is smaller than 0.001.) Using a significance level of $\alpha = 0.05$, the null hypothesis is rejected since the p-value is smaller. That is, the data provide convincing evidence that the question asked did affect a seller's likelihood to tell the truth about problems with the iPod.

EXAMPLE 6.41

Figure 6.17 summarizes the results of an experiment evaluating three treatments for Type 2 Diabetes in patients aged 10-17 who were being treated with metformin. The three treatments considered were continued treatment with metformin (`met`), treatment with metformin combined with rosiglitazone (`rosi`), or a lifestyle intervention program. Each patient had a primary outcome, which was either lacked glycemic control (failure) or did not lack that control (success). What are appropriate hypotheses for this test?

(E)

H_0: There is no difference in the effectiveness of the three treatments.

H_A: There is some difference in effectiveness between the three treatments, e.g. perhaps the `rosi` treatment performed better than `lifestyle`.

	Failure	Success	Total
`lifestyle`	109	125	234
`met`	120	112	232
`rosi`	90	143	233
Total	319	380	699

Figure 6.17: Results for the Type 2 Diabetes study.

GUIDED PRACTICE 6.42

A chi-square test for a two-way table may be used to test the hypotheses in Example 6.41. As a first step, compute the expected values for each of the six table cells.[43]

GUIDED PRACTICE 6.43

Compute the chi-square test statistic for the data in Figure 6.17.[44]

GUIDED PRACTICE 6.44

Because there are 3 rows and 2 columns, the degrees of freedom for the test is $df = (3-1)\times(2-1) = 2$. Use $X^2 = 8.16$, $df = 2$, evaluate whether to reject the null hypothesis using a significance level of 0.05.[45]

[43] The expected count for row one / column one is found by multiplying the row one total (234) and column one total (319), then dividing by the table total (699): $\frac{234\times319}{699} = 106.8$. Similarly for the second column and the first row: $\frac{234\times380}{699} = 127.2$. Row 2: 105.9 and 126.1. Row 3: 106.3 and 126.7.

[44] For each cell, compute $\frac{(\text{obs}-\text{exp})^2}{exp}$. For instance, the first row and first column: $\frac{(109-106.8)^2}{106.8} = 0.05$. Adding the results of each cell gives the chi-square test statistic: $X^2 = 0.05 + \cdots + 2.11 = 8.16$.

[45] If using a computer, we can identify the p-value as 0.017. That is, we reject the null hypothesis because the p-value is less than 0.05, and we conclude that at least one of the treatments is more or less effective than the others at treating Type 2 Diabetes for glycemic control.

Exercises

6.35 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 "patch only" group would you expect to not quit?

6.36 Full body scan, Part II. The table below summarizes a data set we first encountered in Exercise 6.26 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?

6.37 Offshore drilling, Part III. The table below summarizes a data set we first encountered in Exercise 6.23 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

6.38 Parasitic worm. Lymphatic filariasis is a disease caused by a parasitic worm. Complications of the disease can lead to extreme swelling and other complications. Here we consider results from a randomized experiment that compared three different drug treatment options to clear people of the this parasite, which people are working to eliminate entirely. The results for the second year of the study are given below:[46]

	Clear at Year 2	Not Clear at Year 2
Three drugs	52	2
Two drugs	31	24
Two drugs annually	42	14

(a) Set up hypotheses for evaluating whether there is any difference in the performance of the treatments, and also check conditions.

(b) Statistical software was used to run a chi-square test, which output:

$$X^2 = 23.7 \qquad df = 2 \qquad \text{p-value} = 7.2\text{e-}6$$

Use these results to evaluate the hypotheses from part (a), and provide a conclusion in the context of the problem.

[46] Christopher King et al. "A Trial of a Triple-Drug Treatment for Lymphatic Filariasis". In: *New England Journal of Medicine* 379 (2018), pp. 1801–1810.

Chapter exercises

6.39 Active learning. A teacher wanting to increase the active learning component of her course is concerned about student reactions to changes she is planning to make. She conducts a survey in her class, asking students whether they believe more active learning in the classroom (hands on exercises) instead of traditional lecture will helps improve their learning. She does this at the beginning and end of the semester and wants to evaluate whether students' opinions have changed over the semester. Can she used the methods we learned in this chapter for this analysis? Explain your reasoning.

6.40 Website experiment. The OpenIntro website occasionally experiments with design and link placement. We conducted one experiment testing three different placements of a download link for this textbook on the book's main page to see which location, if any, led to the most downloads. The number of site visitors included in the experiment was 701 and is captured in one of the response combinations in the following table:

	Download	No Download
Position 1	13.8%	18.3%
Position 2	14.6%	18.5%
Position 3	12.1%	22.7%

(a) Calculate the actual number of site visitors in each of the six response categories.

(b) Each individual in the experiment had an equal chance of being in any of the three experiment groups. However, we see that there are slightly different totals for the groups. Is there any evidence that the groups were actually imbalanced? Make sure to clearly state hypotheses, check conditions, calculate the appropriate test statistic and the p-value, and make your conclusion in context of the data.

(c) Complete an appropriate hypothesis test to check whether there is evidence that there is a higher rate of site visitors clicking on the textbook link in any of the three groups.

6.41 Shipping holiday gifts. A local news 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?

6.42 The Civil War. A national survey conducted 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.[47]

(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.

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

6.43 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?

6.44 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?

6.45 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.

6.46 Diabetes and unemployment. A 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.[48]

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

(b) State appropriate hypotheses to test for difference in proportions of diabetes between employed and unemployed Americans.

(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.

6.47 Rock-paper-scissors. Rock-paper-scissors is a hand game played by two or more people where players choose to sign either rock, paper, or scissors with their hands. For your statistics class project, you want to evaluate whether players choose between these three options randomly, or if certain options are favored above others. You ask two friends to play rock-paper-scissors and count the times each option is played. The following table summarizes the data:

Rock	Paper	Scissors
43	21	35

Use these data to evaluate whether players choose between these three options randomly, or if certain options are favored above others. Make sure to clearly outline each step of your analysis, and interpret your results in context of the data and the research question.

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

6.48 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.[49]

(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%.

6.49 Browsing on the mobile device. A survey of 2,254 American adults indicates that 17% of cell phone owners browse the internet exclusively on their phone rather than a computer or other device.[50]

(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.[51] 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.

6.50 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.[52]

		Caffeinated coffee consumption					
		≤ 1 cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	≥ 4 cups/day	Total
Clinical depression	Yes	670	373	905	564	95	2,607
	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 $\chi^2 = 20.93$. 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.[53] Do you agree with this statement? Explain your reasoning.

[49] Gallup, Americans Issue Split Decision on Healthcare Ruling, data collected June 28, 2012.
[50] Pew Internet, Cell Internet Use 2012, data collected between March 15 - April 13, 2012.
[51] S. Chang. "The Chinese Love to Use Feature Phone to Access the Internet". In: *M.I.C Gadget* (2012).
[52] M. Lucas et al. "Coffee, caffeine, and risk of depression among women". In: *Archives of internal medicine* 171.17 (2011), p. 1571.
[53] A. O'Connor. "Coffee Drinking Linked to Less Depression in Women". In: *New York Times* (2011).