5.3 Difference of two means

In this section we consider a difference in two population means, $\mu_1 - \mu_2$, under the condition that the data are not paired. Just as with a single sample, we identify conditions to ensure we can use the t-distribution with a point estimate of the difference, $\bar{x}_1 - \bar{x}_2$.

We apply these methods in three contexts: determining whether stem cells can improve heart function, exploring the impact of pregnant womens' smoking habits on birth weights of newborns, and exploring whether there is statistically significant evidence that one variations of an exam is harder than another variation. This section is motivated by questions like "Is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?"

[13] Conditions have already verified and the standard error computed in Example 5.13. To find the interval, identify $t^\star_{72}$ (use $df = 70$ in the table, $t^\star_{70} = 1.99$) and plug it, the point estimate, and the standard error into the confidence interval formula:

$$\text{point estimate} \pm z^\star SE \quad \rightarrow \quad 12.76 \pm 1.99 \times 1.67 \quad \rightarrow \quad (9.44, 16.08)$$

We are 95% confident that Amazon is, on average, between \$9.44 and \$16.08 cheaper than the UCLA bookstore for UCLA course books.

5.3.1 Confidence interval for a difference of means

Does treatment using embryonic stem cells (ESCs) help improve heart function following a heart attack? Table 5.13 contains summary statistics for an experiment to test ESCs in sheep that had a heart attack. Each of these sheep was randomly assigned to the ESC or control group, and the change in their hearts' pumping capacity was measured in the study. A positive value corresponds to increased pumping capacity, which generally suggests a stronger recovery. Our goal will be to identify a 95% confidence interval for the effect of ESCs on the change in heart pumping capacity relative to the control group.

A point estimate of the difference in the heart pumping variable can be found using the difference in the sample means:

$$\bar{x}_{esc} - \bar{x}_{control} = 3.50 - (-4.33) = 7.83$$

	n	$\bar{x}$	s
ESCs	9	3.50	5.17
control	9	-4.33	2.76

Table 5.13: Summary statistics of the embryonic stem cell study.

Using the t-distribution for a difference in means

The t-distribution can be used for inference when working with the standardized difference of two means if (1) each sample meets the conditions for using the t-distribution and (2) the samples are independent.

Example 5.15 Can the t-distribution be used to make inference using the point estimate, $\bar{x}_{esc} - \bar{x}_{control} = 7.83$?

We check the two required conditions:

1. In this study, the sheep were independent of each other. Additionally, the distributions in Figure 5.14 don't show any clear deviations from normality, where we watch for prominent outliers in particular for such small samples. These findings imply each sample mean could itself be modeled using a t-distribution.
2. The sheep in each group were also independent of each other.

Because both conditions are met, we can use the t-distribution to model the difference of the two sample means.

We can quantify the variability in the point estimate, $\bar{x}_{esc} - \bar{x}_{control}$, using the following formula for its standard error:

$$SE_{\bar{x}_{esc} - \bar{x}_{control}} = \sqrt{\frac{\sigma_{esc}^2}{n_{esc}} + \frac{\sigma_{control}^2}{n_{control}}}$$

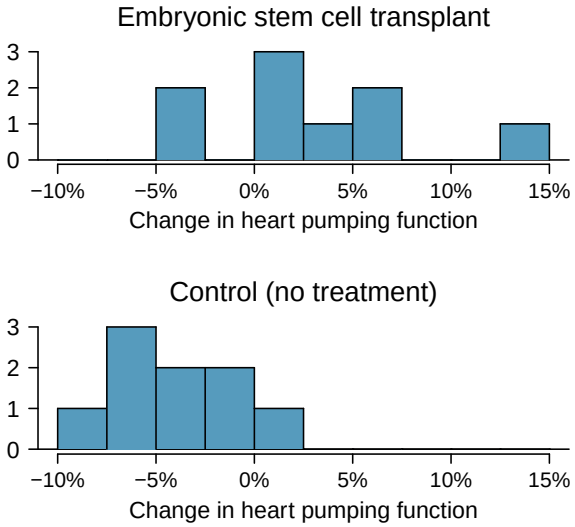


Figure 5.14: Histograms for both the embryonic stem cell group and the control group. Higher values are associated with greater improvement. We don't see any evidence of skew in these data; however, it is worth noting that skew would be difficult to detect with such a small sample.

We usually estimate this standard error using standard deviation estimates based on the samples:

$$SE_{\bar{x}_{esc}-\bar{x}_{control}} = \sqrt{\frac{\sigma_{esc}^2}{n_{esc}} + \frac{\sigma_{control}^2}{n_{control}}}$$
$$\approx \sqrt{\frac{s_{esc}^2}{n_{esc}} + \frac{s_{control}^2}{n_{control}}} = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95$$

Because we will use the t-distribution, we also must identify the appropriate degrees of freedom. This can be done using computer software. An alternative technique is to use the smaller of $n_1 - 1$ and $n_2 - 1$, which is the method we will typically apply in the examples and guided practice.[14]

Distribution of a difference of sample means

The sample difference of two means, $\bar{x}_1 - \bar{x}_2$, can be modeled using the t-distribution and the standard error

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{5.16}$$

when each sample mean can itself be modeled using a t-distribution and the samples are independent. To calculate the degrees of freedom, use statistical software or the smaller of $n_1 - 1$ and $n_2 - 1$.

[14]This technique for degrees of freedom is conservative with respect to a Type 1 Error; it is more difficult to reject the null hypothesis using this df method. In this example, computer software would have provided us a more precise degrees of freedom of $df = 12.225$.

Example 5.17 Calculate a 95% confidence interval for the effect of ESCs on the change in heart pumping capacity of sheep after they've suffered a heart attack.

We will use the sample difference and the standard error for that point estimate from our earlier calculations:

$$\bar{x}_{esc} - \bar{x}_{control} = 7.83$$

$$SE = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95$$

Using $df = 8$, we can identify the appropriate $t^\star_{df} = t^\star_8$ for a 95% confidence interval as 2.31. Finally, we can enter the values into the confidence interval formula:

$$\text{point estimate} \pm t^\star SE \quad \rightarrow \quad 7.83 \pm 2.31 \times 1.95 \quad \rightarrow \quad (3.32, 12.34)$$

We are 95% confident that embryonic stem cells improve the heart's pumping function in sheep that have suffered a heart attack by 3.32% to 12.34%.

5.3.2 Hypothesis tests based on a difference in means

A data set called `baby_smoke` represents a random sample of 150 cases of mothers and their newborns in North Carolina over a year. Four cases from this data set are represented in Table 5.15. We are particularly interested in two variables: `weight` and `smoke`. The `weight` variable represents the weights of the newborns and the `smoke` variable describes which mothers smoked during pregnancy. We would like to know, is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke? We will use the North Carolina sample to try to answer this question. The smoking group includes 50 cases and the nonsmoking group contains 100 cases, represented in Figure 5.16.

	fAge	mAge	weeks	weight	sexBaby	smoke
1	NA	13	37	5.00	female	nonsmoker
2	NA	14	36	5.88	female	nonsmoker
3	19	15	41	8.13	male	smoker
⋮	⋮	⋮	⋮	⋮	⋮	
150	45	50	36	9.25	female	nonsmoker

Table 5.15: Four cases from the `baby_smoke` data set. The value "NA", shown for the first two entries of the first variable, indicates that piece of data is missing.-2mm

Example 5.18 Set up appropriate hypotheses to evaluate whether there is a relationship between a mother smoking and average birth weight.

The null hypothesis represents the case of no difference between the groups.

H_0: There is no difference in average birth weight for newborns from mothers who did and did not smoke. In statistical notation: $\mu_n - \mu_s = 0$, where μ_n represents non-smoking mothers and μ_s represents mothers who smoked.

H_A: There is some difference in average newborn weights from mothers who did and did not smoke ($\mu_n - \mu_s \neq 0$).

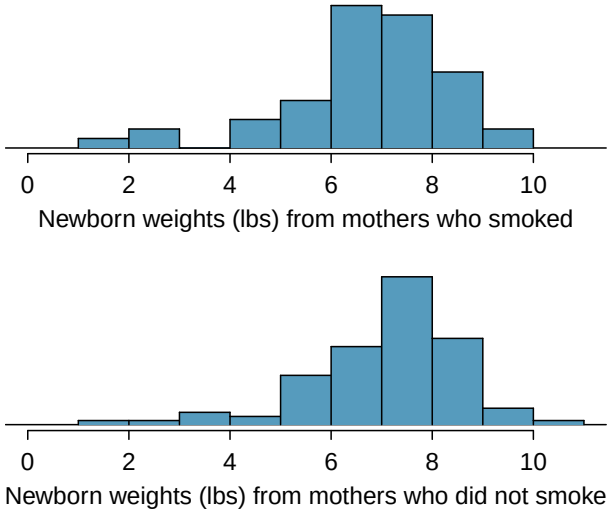


Figure 5.16: The top panel represents birth weights for infants whose mothers smoked. The bottom panel represents the birth weights for infants whose mothers who did not smoke. The distributions exhibit moderate-to-strong and strong skew, respectively.

We check the two conditions necessary to apply the t-distribution to the difference in sample means. (1) Because the data come from a simple random sample and consist of less than 10% of all such cases, the observations are independent. Additionally, while each distribution is strongly skewed, the sample sizes of 50 and 100 would make it reasonable to model each mean separately using a t-distribution. The skew is reasonable for these sample sizes of 50 and 100. (2) The independence reasoning applied in (1) also ensures the observations in each sample are independent. Since both conditions are satisfied, the difference in sample means may be modeled using a t-distribution.

	`smoker`	`nonsmoker`
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

Table 5.17: Summary statistics for the `baby_smoke` data set.

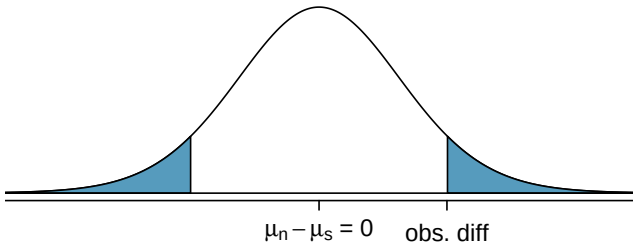
⊙ **Guided Practice 5.19** The summary statistics in Table 5.17 may be useful for this exercise. (a) What is the point estimate of the population difference, $\mu_n - \mu_s$? (b) Compute the standard error of the point estimate from part (a).[15]

[15] (a) The difference in sample means is an appropriate point estimate: $\bar{x}_n - \bar{x}_s = 0.40$. (b) The standard error of the estimate can be estimated using Equation (5.16):

$$SE = \sqrt{\frac{\sigma_n^2}{n_n} + \frac{\sigma_s^2}{n_s}} \approx \sqrt{\frac{s_n^2}{n_n} + \frac{s_s^2}{n_s}} = \sqrt{\frac{1.60^2}{100} + \frac{1.43^2}{50}} = 0.26$$

Example 5.20 Draw a picture to represent the p-value for the hypothesis test from Example 5.18.

To depict the p-value, we draw the distribution of the point estimate as though H_0 were true and shade areas representing at least as much evidence against H_0 as what was observed. Both tails are shaded because it is a two-sided test.



Example 5.21 Compute the p-value of the hypothesis test using the figure in Example 5.20, and evaluate the hypotheses using a significance level of $\alpha = 0.05$.

We start by computing the T-score:

$$T = \frac{0.40 - 0}{0.26} = 1.54$$

Next, we compare this value to values in the t-table in Appendix B.2 on page 430, where we use the smaller of $n_n - 1 = 99$ and $n_s - 1 = 49$ as the degrees of freedom: $df = 49$. The T-score falls between the first and second columns in the $df = 49$ row of the t-table, meaning the two-sided p-value falls between 0.10 and 0.20 (reminder, find tail areas along the top of the table). This p-value is larger than the significance value, 0.05, so we fail to reject the null hypothesis. There is insufficient evidence to say there is a difference in average birth weight of newborns from North Carolina mothers who did smoke during pregnancy and newborns from North Carolina mothers who did not smoke during pregnancy.

Guided Practice 5.22 Does the conclusion to Example 5.21 mean that smoking and average birth weight are unrelated?[16]

Guided Practice 5.23 If we made a Type 2 Error and there is a difference, what could we have done differently in data collection to be more likely to detect the difference?[17]

[16] Absolutely not. It is possible that there is some difference but we did not detect it. If there is a difference, we made a Type 2 Error. Notice: we also don't have enough information to, if there is an actual difference, confidently say which direction that difference would be in.

[17] We could have collected more data. If the sample sizes are larger, we tend to have a better shot at finding a difference if one exists.

Public service announcement: while we have used this relatively small data set as an example, larger data sets show that women who smoke tend to have smaller newborns. In fact, some in the tobacco industry actually had the audacity to tout that as a *benefit* of smoking:

> *It's true. The babies born from women who smoke are smaller, but they're just as healthy as the babies born from women who do not smoke. And some women would prefer having smaller babies.*
>
> - Joseph Cullman, Philip Morris' Chairman of the Board on CBS' *Face the Nation*, Jan 3, 1971

Fact check: the babies from women who smoke are not actually as healthy as the babies from women who do not smoke.[18]

5.3.3 Case study: two versions of a course exam

An instructor decided to run two slight variations of the same exam. Prior to passing out the exams, she shuffled the exams together to ensure each student received a random version. Summary statistics for how students performed on these two exams are shown in Table 5.18. Anticipating complaints from students who took Version B, she would like to evaluate whether the difference observed in the groups is so large that it provides convincing evidence that Version B was more difficult (on average) than Version A.

Version	n	$\bar{x}$	s	min	max
A	30	79.4	14	45	100
B	27	74.1	20	32	100

Table 5.18: Summary statistics of scores for each exam version.

⊙ **Guided Practice 5.24** Construct a hypotheses to evaluate whether the observed difference in sample means, $\bar{x}_A - \bar{x}_B = 5.3$, is due to chance.[19]

⊙ **Guided Practice 5.25** To evaluate the hypotheses in Guided Practice 5.24 using the t-distribution, we must first verify assumptions. (a) Does it seem reasonable that the scores are independent within each group? (b) What about the normality / skew condition for observations in each group? (c) Do you think scores from the two groups would be independent of each other, i.e. the two samples are independent?[20]

After verifying the conditions for each sample and confirming the samples are independent of each other, we are ready to conduct the test using the t-distribution. In this case,

[18] You can watch an episode of John Oliver on *This Week Tonight* to explore the present day offenses of the tobacco industry. Please be aware that there is some adult language: youtu.be/6UsHHOCH4q8.

[19] Because the teacher did not expect one exam to be more difficult prior to examining the test results, she should use a two-sided hypothesis test. H_0: the exams are equally difficult, on average. $\mu_A - \mu_B = 0$. H_A: one exam was more difficult than the other, on average. $\mu_A - \mu_B \neq 0$.

[20] (a) It is probably reasonable to conclude the scores are independent, provided there was no cheating. (b) The summary statistics suggest the data are roughly symmetric about the mean, and it doesn't seem unreasonable to suggest the data might be normal. Note that since these samples are each nearing 30, moderate skew in the data would be acceptable. (c) It seems reasonable to suppose that the samples are independent since the exams were handed out randomly.

we are estimating the true difference in average test scores using the sample data, so the point estimate is $\bar{x}_A - \bar{x}_B = 5.3$. The standard error of the estimate can be calculated as

$$SE = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \sqrt{\frac{14^2}{30} + \frac{20^2}{27}} = 4.62$$

Finally, we construct the test statistic:

$$T = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{(79.4 - 74.1) - 0}{4.62} = 1.15$$

If we have a computer handy, we can identify the degrees of freedom as 45.97. Otherwise we use the smaller of $n_1 - 1$ and $n_2 - 1$: $df = 26$.

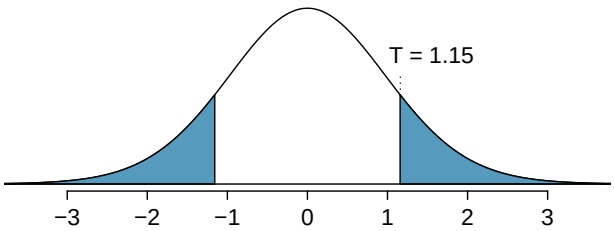


Figure 5.19: The t-distribution with 26 degrees of freedom. The shaded right tail represents values with $T \geq 1.15$. Because it is a two-sided test, we also shade the corresponding lower tail.

Example 5.26 Identify the p-value using $df = 26$ and provide a conclusion in the context of the case study.

We examine row $df = 26$ in the t-table. Because this value is smaller than the value in the left column, the p-value is larger than 0.200 (two tails!). Because the p-value is so large, we do not reject the null hypothesis. That is, the data do not convincingly show that one exam version is more difficult than the other, and the teacher should not be convinced that she should add points to the Version B exam scores.

5.3.4 Summary for inference using the t-distribution

Hypothesis tests. When applying the t-distribution for a hypothesis test, we proceed as follows:

- Write appropriate hypotheses.
- Verify conditions for using the t-distribution.
  - One-sample or differences from paired data: the observations (or differences) must be independent and nearly normal. For larger sample sizes, we can relax the nearly normal requirement, e.g. slight skew is okay for sample sizes of 15, moderate skew for sample sizes of 30, and strong skew for sample sizes of 60.
  - For a difference of means when the data are not paired: each sample mean must separately satisfy the one-sample conditions for the t-distribution, and the data in the groups must also be independent.

- Compute the point estimate of interest, the standard error, and the degrees of freedom. For df, use $n-1$ for one sample, and for two samples use either statistical software or the smaller of n_1-1 and n_2-1.
- Compute the T-score and p-value.
- Make a conclusion based on the p-value, and write a conclusion in context and in plain language so anyone can understand the result.

Confidence intervals. Similarly, the following is how we generally computed a confidence interval using a t-distribution:

- Verify conditions for using the t-distribution. (See above.)
- Compute the point estimate of interest, the standard error, the degrees of freedom, and $t^\star_{df}$.
- Calculate the confidence interval using the general formula, point estimate $\pm\, t^\star_{df} SE$.
- Put the conclusions in context and in plain language so even non-statisticians can understand the results.

Calculator videos
Videos covering confidence intervals and hypothesis tests for a difference of means using TI and Casio graphing calculators are available at openintro.org/videos.

5.3.5 Examining the standard error formula (special topic)

The formula for the standard error of the difference in two means is similar to the formula for other standard errors. Recall that the standard error of a single mean, $\bar{x}_1$, can be approximated by

$$SE_{\bar{x}_1} = \frac{s_1}{\sqrt{n_1}}$$

where s_1 and n_1 represent the sample standard deviation and sample size.

The standard error of the difference of two sample means can be constructed from the standard errors of the separate sample means:

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{SE^2_{\bar{x}_1} + SE^2_{\bar{x}_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{5.27}$$

This special relationship follows from probability theory.

Guided Practice 5.28 Prerequisite: Section 2.4. We can rewrite Equation (5.27) in a different way:

$$SE^2_{\bar{x}_1-\bar{x}_2} = SE^2_{\bar{x}_1} + SE^2_{\bar{x}_2}$$

Explain where this formula comes from using the ideas of probability theory.[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 $\bar{x}_1-\bar{x}_2$ is $\sigma^2_{\bar{x}_1}+\sigma^2_{\bar{x}_2}$. Because $\sigma^2_{\bar{x}_1}$ and $\sigma^2_{\bar{x}_2}$ are just another way of writing $SE^2_{\bar{x}_1}$ and $SE^2_{\bar{x}_2}$, the variance associated with $\bar{x}_1-\bar{x}_2$ may be written as $SE^2_{\bar{x}_1}+SE^2_{\bar{x}_2}$.

5.3.6 Pooled standard deviation estimate (special topic)

Occasionally, two populations will have standard deviations that are so similar that they can be treated as identical. For example, historical data or a well-understood biological mechanism may justify this strong assumption. In such cases, we can make the t-distribution approach slightly more precise by using a pooled standard deviation.

The **pooled standard deviation** of two groups is a way to use data from both samples to better estimate the standard deviation and standard error. If s_1 and s_2 are the standard deviations of groups 1 and 2 and there are good reasons to believe that the population standard deviations are equal, then we can obtain an improved estimate of the group variances by pooling their data:

$$s^2_{pooled} = \frac{s_1^2 \times (n_1 - 1) + s_2^2 \times (n_2 - 1)}{n_1 + n_2 - 2}$$

where n_1 and n_2 are the sample sizes, as before. To use this new statistic, we substitute s^2_{pooled} in place of s_1^2 and s_2^2 in the standard error formula, and we use an updated formula for the degrees of freedom:

$$df = n_1 + n_2 - 2$$

The benefits of pooling the standard deviation are realized through obtaining a better estimate of the standard deviation for each group and using a larger degrees of freedom parameter for the t-distribution. Both of these changes may permit a more accurate model of the sampling distribution of $\bar{x}_1 - \bar{x}_2$, if the standard deviations of the two groups are equal.

Caution: Pool standard deviations only after careful consideration

A pooled standard deviation is only appropriate when background research indicates the population standard deviations are nearly equal. When the sample size is large and the condition may be adequately checked with data, the benefits of pooling the standard deviations greatly diminishes.

Chapter 6

Inference for categorical data

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

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

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

6.1 Inference for a single proportion

In New York City on October 23rd, 2014, a doctor who had recently been treating Ebola patients in Guinea went to the hospital with a slight fever and was subsequently diagnosed with Ebola. Soon thereafter, an NBC 4 New York/The Wall Street Journal/Marist Poll found that 82% of New Yorkers favored a "mandatory 21-day quarantine for anyone who has come in contact with an Ebola patient".[1] This poll included responses of 1,042 New York adults between October 26th and 28th, 2014.

[1] Poll ID NY141026 on maristpoll.marist.edu.

6.1.1 Identifying when the sample proportion is nearly normal

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

$$\hat{p} = \frac{0 + 1 + 1 + \cdots + 0}{1042} = 0.82$$

The distribution of $\hat{p}$ is nearly normal when the distribution of 0's and 1's is not too strongly skewed for the sample size. The most common guideline for sample size and skew when working with proportions is to ensure that we expect to observe a minimum number of successes (1's) and failures (0's), typically at least 10 of each. The labels **success** and **failure** need not mean something positive or negative. These terms are just convenient words that are frequently used when discussing proportions.

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

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

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

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

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

$\hat{p}$
sample proportion

p
population proportion

Typically we don't know the true proportion, p, so we substitute some value to check conditions and to estimate the standard error. For confidence intervals, usually 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. Examples are presented for each of these cases in Sections 6.1.2 and 6.1.3.

TIP: Reminder on checking independence of observations

If data come from a simple random sample and consist of less than 10% of the population, then the independence assumption is reasonable. Alternatively, if the data come from a random process, we must evaluate the independence condition more carefully.

6.1.2 Confidence intervals for a proportion

We may want a confidence interval for the proportion of New York adults who favored a mandatory quarantine of anyone who had been in contact with an Ebola patient. Our point estimate, based on a sample of size $n = 1042$, is $\hat{p} = 0.82$. We would like to use the general confidence interval formula from Section 4.5. However, first we must verify that the sampling distribution of $\hat{p}$ is nearly normal and calculate the standard error of $\hat{p}$.

Observations are independent. The poll is based on a simple random sample and consists of fewer than 10% of the New York adult population, which verifies independence.

Success-failure condition. The sample size must also be sufficiently large, which is checked using the success-failure condition. There were $1042 \times \hat{p} \approx 854$ "successes" and $1042 \times (1 - \hat{p}) \approx 188$ "failures" in the sample, both easily greater than 10.

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

Guided Practice 6.2 Estimate the standard error of $\hat{p} = 0.82$ using Equation (6.1). Because p is unknown and the standard error is for a confidence interval, use $\hat{p}$ in place of p in the formula.[2]

Example 6.3 Construct a 95% confidence interval for p, the proportion of New York adults who supported a quarantine for anyone who has come into contact with an Ebola patient.

Using the standard error $SE = 0.012$ from Guided Practice 6.2, the point estimate 0.82, and $z^\star = 1.96$ for a 95% confidence interval, the confidence interval is

$$\text{point estimate} \pm z^\star SE \quad \rightarrow \quad 0.82 \pm 1.96 \times 0.012 \quad \rightarrow \quad (0.796, 0.844)$$

We are 95% confident that the true proportion of New York adults in October 2014 who supported a quarantine for anyone who had come into contact with an Ebola patient was between 0.796 and 0.844.

Notice that since the poll was around the time where a doctor in New York had come down with Ebola, the results may not be as applicable today as they were at the time the poll was taken. This highlights an important detail about polls: they provide data about public opinion at a single point in time.

Constructing a confidence interval for a proportion

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

[2] $SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{0.82(1-0.82)}{1042}} = 0.012$.

6.1.3 Hypothesis testing for a proportion

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

Guided Practice 6.4 Do a majority of American support nuclear arms reduction? Set up a one-sided hypothesis test to evaluate this question.[3]

Example 6.5 A simple random sample of 1,028 US adults in March 2013 found that 56% support nuclear arms reduction.[4] Does this provide convincing evidence that a majority of Americans supported nuclear arms reduction at the 5% significance level?

The poll was of a simple random sample that includes fewer than 10% of US adults, meaning the observations are independent. In a one-proportion hypothesis test, the success-failure condition is checked using the null proportion, which is $p_0 = 0.5$ in this context: $np_0 = n(1-p_0) = 1028 \times 0.5 = 514 > 10$. With these conditions verified, the normal model may be applied to $\hat{p}$.

Next the standard error can be computed. The null value p_0 is used again here, because this is a hypothesis test for a single proportion.

$$SE = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{1028}} = 0.016$$

A picture of the normal model is shown in Figure 6.1 with the p-value represented by the shaded region. Based on the normal model, the test statistic can be computed as the Z-score of the point estimate:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.56 - 0.50}{0.016} = 3.75$$

The upper tail area, representing the p-value, is about 0.0001. Because the p-value is smaller than 0.05, we reject H_0. The poll provides convincing evidence that a majority of Americans supported nuclear arms reduction efforts in March 2013.

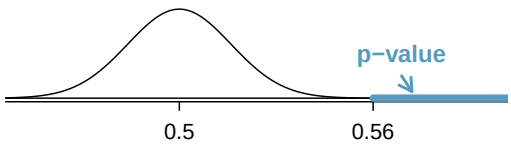


Figure 6.1: Sampling distribution for Example 6.5.

Hypothesis test for a proportion

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

[3] $H_0 : p = 0.50$. $H_A : p > 0.50$.
[4] www.gallup.com/poll/161198/favor-russian-nuclear-arms-reductions.aspx

Calculator videos
Videos covering confidence intervals and hypothesis tests for a single proportion using TI and Casio graphing calculators are available at openintro.org/videos.

6.1.4 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. More explicitly, our task is to find a sample size n so that the sample proportion is within some margin of error m of the actual proportion with a certain level of confidence.

Example 6.6 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 $m = 0.04$. For a 95% confidence level, the value $z^\star$ corresponds to 1.96:

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

There are two unknowns in the equation: p and n. If we have an estimate of p, perhaps from a similar 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.

Example 6.7 A manager is about to oversee the mass production of a new tire model in her factory, and she would like to estimate what proportion of these tires will be rejected through quality control. The quality control team has monitored the last three tire models produced by the factory, failing 1.7% of tires in the first model, 6.2% of the second model, and 1.3% of the third model. The manager would like to examine enough tires to estimate the failure rate of the new tire model to within about 2% with a 90% confidence level.

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

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

(a) For a 90% confidence interval, $z^\star = 1.65$, and since an estimate of the proportion 0.017 is available, we'll use it in the margin of error formula:

$$1.65 \times \sqrt{\frac{0.017(1-0.017)}{n}} < 0.02$$
$$113.7 < n$$

For sample size calculations, we always round up, so the first tire model suggests 114 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 396 and 88 tires, respectively.

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

It should also be noted that the success-failure condition is not met with $n = 114$ or $n = 88$. That is, we would need additional methods than what we've covered so far to analyze results based on those sample sizes.

Guided Practice 6.8 A recent estimate of Congress' approval rating was 19%.[5] What sample size does this estimate suggest we should use for a margin of error of 0.04 with 95% confidence?[6]

---

[5] www.gallup.com/poll/183128/five-months-gop-congress-approval-remains-low.aspx

[6] We complete the same computations as before, except now we use 0.19 instead of 0.5 for p:

$$1.96 \times \sqrt{\frac{p(1-p)}{n}} \approx 1.96 \times \sqrt{\frac{0.19(1-0.19)}{n}} \leq 0.04 \qquad \rightarrow \qquad n \geq 369.5$$

A sample size of 370 or more would be reasonable. (Reminder: always round up for sample size calculations!)

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

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

6.5.1 When the success-failure condition is not met

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

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

● **Example 6.43** Is it possible to assess the consultant's claim with the data provided?

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

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

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

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

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

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

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

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

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

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

6.5.2 Generating the null distribution and p-value by simulation

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

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

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

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

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

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

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

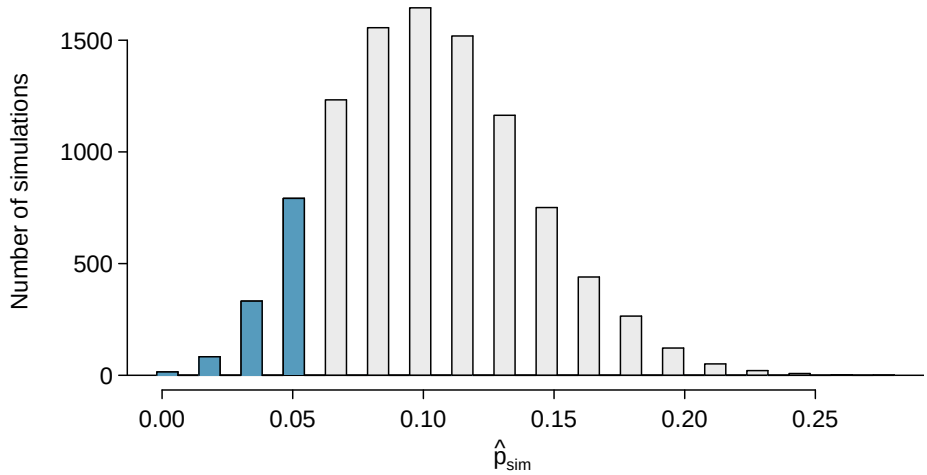


Figure 6.21: The null distribution for $\hat{p}$, created from 10,000 simulated studies. The left tail, representing the p-value for the hypothesis test, contains 12.22% of the simulations.

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

Figure 6.21 shows the results of 10,000 simulated studies. The proportions that are equal to or less than $\hat{p} = 0.048$ are shaded. The shaded areas represent sample proportions under the null distribution that provide at least as much evidence as $\hat{p}$ favoring the alternative hypothesis. There were 1222 simulated sample proportions with $\hat{p}_{sim} \leq 0.048$. We use these to construct the null distribution's left-tail area and find the p-value:

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

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

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

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

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

One-sided hypothesis test for p with a small sample

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

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

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

6.5.3 Generating the exact null distribution and p-value

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

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

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

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

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

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

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

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

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

If it were plotted, the exact null distribution would look almost identical to the simulated null distribution shown in Figure 6.21 on page 304.

6.5.4 Using simulation for goodness of fit tests

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

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

Since the minimum bin count condition was satisfied, the chi-square distribution is an excellent approximation of the null distribution, meaning the results should be very similar. Figure 6.22 shows the simulated null distribution using 100,000 simulated X^2_{sim} values with an overlaid curve of the chi-square distribution. The distributions are almost identical, and the p-values are essentially indistinguishable: 0.115 for the simulated null distribution and 0.117 for the theoretical null distribution.

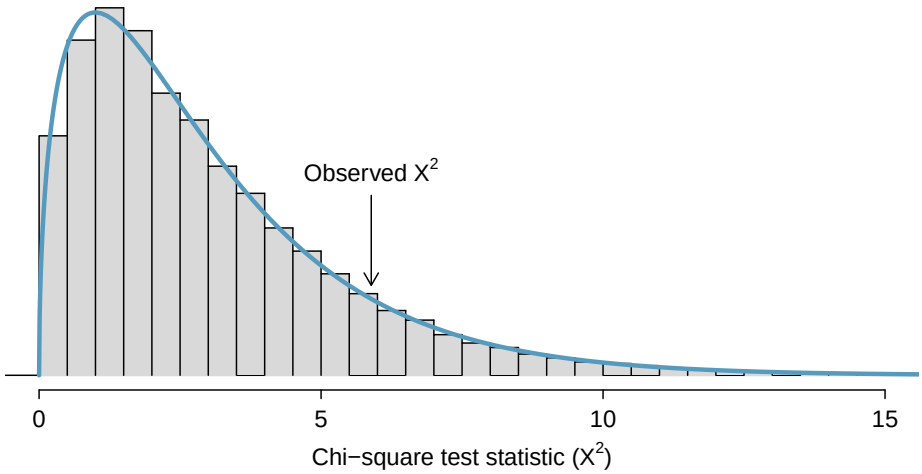


Figure 6.22: The precise null distribution for the juror example from Section 6.3 is shown as a histogram of simulated X^2_{sim} statistics, and the theoretical chi-square distribution is also shown.