Chapter 5

Inference for numerical data

Chapter 4 introduced a framework for statistical inference based on confidence intervals and hypotheses. In this chapter, we encounter several new point estimates and scenarios. In each case, the inference ideas remain the same:

1. Determine which point estimate or test statistic is useful.
2. Identify an appropriate distribution for the point estimate or test statistic.
3. Apply the ideas from Chapter 4 using the distribution from step 2.

5.1 One-sample means with the t-distribution

We required a large sample in Chapter 4 for two reasons:

1. The sampling distribution of $\bar{x}$ tends to be more normal when the sample is large.
2. The calculated standard error is typically very accurate when using a large sample.

So what should we do when the sample size is small? As we'll discuss in Section 5.1.1, if the population data are nearly normal, then $\bar{x}$ will also follow a normal distribution, which addresses the first problem. The accuracy of the standard error is trickier, and for this challenge we'll introduce a new distribution called the t-distribution.

While we emphasize the use of the t-distribution for small samples, this distribution is also generally used for large samples, where it produces similar results to those from the normal distribution.

5.1.1 The normality condition

A special case of the Central Limit Theorem ensures the distribution of sample means will be nearly normal, regardless of sample size, when the data come from a nearly normal distribution.

Central Limit Theorem for normal data

The sampling distribution of the mean is nearly normal when the sample observations are independent and come from a nearly normal distribution. This is true for any sample size.

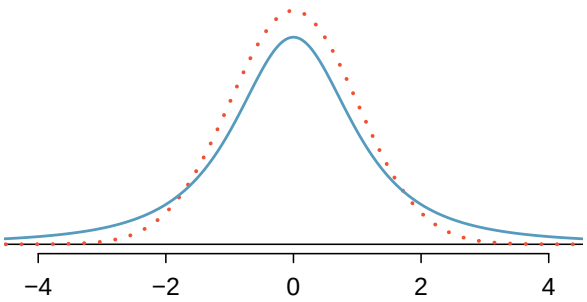


Figure 5.1: Comparison of a t-distribution (solid line) and a normal distribution (dotted line).

While this seems like a very helpful special case, there is one small problem. It is inherently difficult to verify normality in small data sets.

Caution: Checking the normality condition

We should exercise caution when verifying the normality condition for small samples. It is important to not only examine the data but also think about where the data come from. For example, ask: would I expect this distribution to be symmetric, and am I confident that outliers are rare?

You may relax the normality condition as the sample size goes up. If the sample size is 10 or more, slight skew is not problematic. Once the sample size hits about 30, then moderate skew is reasonable. Data with strong skew or outliers require a more cautious analysis.

5.1.2 Introducing the t-distribution

In the cases where we will use a small sample to calculate the standard error, it will be useful to rely on a new distribution for inference calculations: the t-distribution. A t-distribution, shown as a solid line in Figure 5.1, has a bell shape. However, its tails are thicker than the normal model's. This means observations are more likely to fall beyond two standard deviations from the mean than under the normal distribution.[1] While our estimate of the standard error will be a little less accurate when we are analyzing a small data set, these extra thick tails of the t-distribution are exactly the correction we need to resolve the problem of a poorly estimated standard error.

The t-distribution, always centered at zero, has a single parameter: degrees of freedom. The **degrees of freedom (df)** describe the precise form of the bell-shaped t-distribution. Several t-distributions are shown in Figure 5.2. When there are more degrees of freedom, the t-distribution looks very much like the standard normal distribution.

[1] The standard deviation of the t-distribution is actually a little more than 1. However, it is useful to always think of the t-distribution as having a standard deviation of 1 in all of our applications.

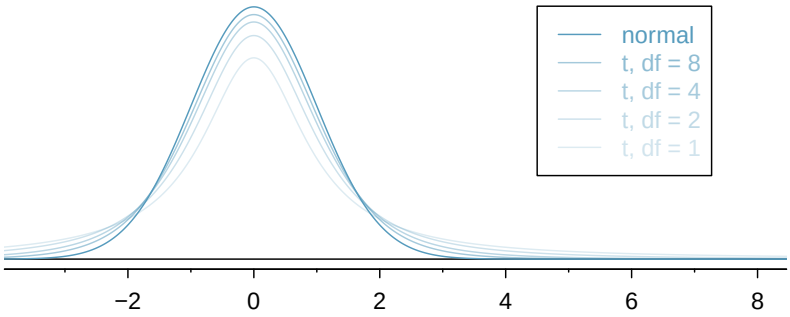


Figure 5.2: The larger the degrees of freedom, the more closely the t-distribution resembles the standard normal model.

Degrees of freedom (df)

The degrees of freedom describe the shape of the t-distribution. The larger the degrees of freedom, the more closely the distribution approximates the normal model.

When the degrees of freedom is about 30 or more, the t-distribution is nearly indistinguishable from the normal distribution. In Section 5.1.3, we relate degrees of freedom to sample size.

It's very useful to become familiar with the t-distribution, because it allows us greater flexibility than the normal distribution when analyzing numerical data. We use a **t-table**, partially shown in Table 5.3, in place of the normal probability table. A larger t-table is in Appendix B.2 on page 430. In practice, it's more common to use statistical software instead of a table, and you can see some of these options at

www.openintro.org/stat/prob-tables

Each row in the t-table represents a t-distribution with different degrees of freedom. The columns correspond to tail probabilities. For instance, if we know we are working with the t-distribution with $df = 18$, we can examine row 18, which is highlighted in Table 5.3. If we want the value in this row that identifies the cutoff for an upper tail of 10%, we can look in the column where *one tail* is 0.100. This cutoff is 1.33. If we had wanted the cutoff for the lower 10%, we would use -1.33. Just like the normal distribution, all t-distributions are symmetric.

Example 5.1 What proportion of the t-distribution with 18 degrees of freedom falls below -2.10?

Just like a normal probability problem, we first draw the picture in Figure 5.4 and shade the area below -2.10. To find this area, we identify the appropriate row: $df = 18$. Then we identify the column containing the absolute value of -2.10; it is the third column. Because we are looking for just one tail, we examine the top line of the table, which shows that a one tail area for a value in the third row corresponds to 0.025. About 2.5% of the distribution falls below -2.10. In the next example we encounter a case where the exact t value is not listed in the table.

	one tail	0.100	0.050	0.025	0.010	0.005
	two tails	0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	⋮	⋮	⋮	⋮	⋮	
	17	1.33	1.74	2.11	2.57	2.90
	18	**1.33**	**1.73**	**2.10**	**2.55**	**2.88**
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	⋮	⋮	⋮	⋮	⋮	
	400	1.28	1.65	1.97	2.34	2.59
	500	1.28	1.65	1.96	2.33	2.59
	∞	1.28	1.64	1.96	2.33	2.58

Table 5.3: An abbreviated look at the t-table. Each row represents a different t-distribution. The columns describe the cutoffs for specific tail areas. The row with $df = 18$ has been **highlighted**.

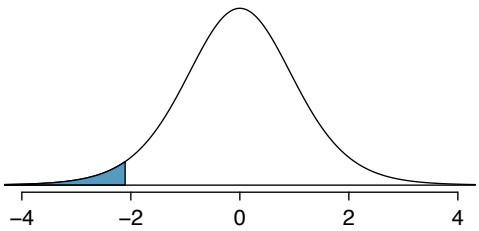


Figure 5.4: The t-distribution with 18 degrees of freedom. The area below -2.10 has been shaded.

Example 5.2 A t-distribution with 20 degrees of freedom is shown in the left panel of Figure 5.5. Estimate the proportion of the distribution falling above 1.65.

We identify the row in the t-table using the degrees of freedom: $df = 20$. Then we look for 1.65; it is not listed. It falls between the first and second columns. Since these values bound 1.65, their tail areas will bound the tail area corresponding to 1.65. We identify the one tail area of the first and second columns, 0.050 and 0.10, and we conclude that between 5% and 10% of the distribution is more than 1.65 standard deviations above the mean. If we like, we can identify the precise area using statistical software: 0.0573.

Example 5.3 A t-distribution with 2 degrees of freedom is shown in the right panel of Figure 5.5. Estimate the proportion of the distribution falling more than 3 units from the mean (above or below).

As before, first identify the appropriate row: $df = 2$. Next, find the columns that capture 3; because $2.92 < 3 < 4.30$, we use the second and third columns. Finally, we find bounds for the tail areas by looking at the two tail values: 0.05 and 0.10. We use the two tail values because we are looking for two (symmetric) tails.

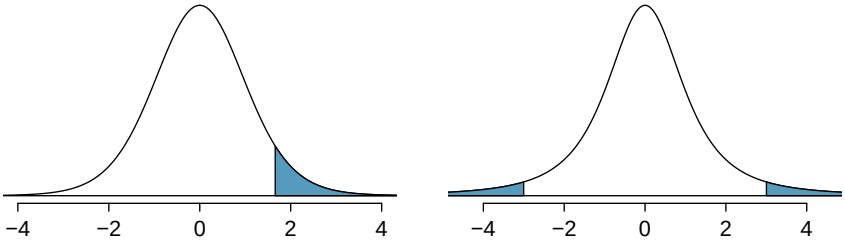


Figure 5.5: Left: The t-distribution with 20 degrees of freedom, with the area above 1.65 shaded. Right: The t-distribution with 2 degrees of freedom, with the area further than 3 units from 0 shaded.

⊙ **Guided Practice 5.4** What proportion of the t-distribution with 19 degrees of freedom falls above -1.79 units?[2]

5.1.3 Conditions for using the t-distribution for inference on a sample mean

To proceed with the t-distribution for inference about a single mean, we first check two conditions.

Independence of observations. We verify this condition just as we did before. We collect a simple random sample from less than 10% of the population, or if the data are from an experiment or random process, we check to the best of our abilities that the observations were independent.

Observations come from a nearly normal distribution. This second condition is difficult to verify with small data sets. We often (i) take a look at a plot of the data for obvious departures from the normal model, and (ii) consider whether any previous experiences alert us that the data may not be nearly normal.

When examining a sample mean and estimated standard error from a sample of n independent and nearly normal observations, we use a t-distribution with $n - 1$ degrees of freedom (df). For example, if the sample size was 19, then we would use the t-distribution with $df = 19 - 1 = 18$ degrees of freedom and proceed exactly as we did in Chapter 4, except that *now we use the t-distribution.*

TIP: When to use the t-distribution
Use the t-distribution for inference of the sample mean when observations are independent and nearly normal. You may relax the nearly normal condition as the sample size increases. For example, the data distribution may be moderately skewed when the sample size is at least 30.

[2] We find the shaded area *above* -1.79 (we leave the picture to you). The small left tail is between 0.025 and 0.05, so the larger upper region must have an area between 0.95 and 0.975.

5.1.4 One sample t-confidence intervals

Dolphins are at the top of the oceanic food chain, which causes dangerous substances such as mercury to concentrate in their organs and muscles. This is an important problem for both dolphins and other animals, like humans, who occasionally eat them. For instance, this is particularly relevant in Japan where school meals have included dolphin at times.

Figure 5.6: A Risso's dolphin.

Photo by Mike Baird (www.bairdphotos.com). CC BY 2.0 license.

Here we identify a confidence interval for the average mercury content in dolphin muscle using a sample of 19 Risso's dolphins from the Taiji area in Japan.[3] The data are summarized in Table 5.7. The minimum and maximum observed values can be used to evaluate whether or not there are obvious outliers or skew.

n	$\bar{x}$	s	minimum	maximum
19	4.4	2.3	1.7	9.2

Table 5.7: Summary of mercury content in the muscle of 19 Risso's dolphins from the Taiji area. Measurements are in μg/wet g (micrograms of mercury per wet gram of muscle).

Example 5.5 Are the independence and normality conditions satisfied for this data set?

The observations are a simple random sample and consist of less than 10% of the population, therefore independence is reasonable. The summary statistics in Table 5.7 do not suggest any skew or outliers; all observations are within 2.5 standard deviations of the mean. Based on this evidence, the normality assumption seems reasonable.

[3] Taiji was featured in the movie *The Cove*, and it is a significant source of dolphin and whale meat in Japan. Thousands of dolphins pass through the Taiji area annually, and we will assume these 19 dolphins represent a simple random sample from those dolphins. Data reference: Endo T and Haraguchi K. 2009. High mercury levels in hair samples from residents of Taiji, a Japanese whaling town. Marine Pollution Bulletin 60(5):743-747.

In the normal model, we used $z^\star$ and the standard error to determine the width of a confidence interval. We revise the confidence interval formula slightly when using the t-distribution:

$$\bar{x} \pm t^\star_{df} SE$$

$t^\star_{df}$
Multiplication factor for t conf. interval

The sample mean and estimated standard error are computed just as before ($\bar{x} = 4.4$ and $SE = s/\sqrt{n} = 0.528$). The value $t^\star_{df}$ is a cutoff we obtain based on the confidence level and the t-distribution with df degrees of freedom. Before determining this cutoff, we will first need the degrees of freedom.

Degrees of freedom for a single sample

If the sample has n observations and we are examining a single mean, then we use the t-distribution with $df = n - 1$ degrees of freedom.

In our current example, we should use the t-distribution with $df = 19 - 1 = 18$ degrees of freedom. Then identifying $t^\star_{18}$ is similar to how we found $z^\star$.

- For a 95% confidence interval, we want to find the cutoff $t^\star_{18}$ such that 95% of the t-distribution is between -$t^\star_{18}$ and $t^\star_{18}$.
- We look in the t-table on page 222, find the column with area totaling 0.05 in the two tails (third column), and then the row with 18 degrees of freedom: $t^\star_{18} = 2.10$.

Generally the value of $t^\star_{df}$ is slightly larger than what we would get under the normal model with $z^\star$.

Finally, we can substitute all our values into the confidence interval equation to create the 95% confidence interval for the average mercury content in muscles from Risso's dolphins that pass through the Taiji area:

$$\bar{x} \pm t^\star_{18} SE \quad \rightarrow \quad 4.4 \pm 2.10 \times 0.528 \quad \rightarrow \quad (3.29, 5.51)$$

We are 95% confident the average mercury content of muscles in Risso's dolphins is between 3.29 and 5.51 μg/wet gram, which is considered extremely high.

Finding a t-confidence interval for the mean

Based on a sample of n independent and nearly normal observations, a confidence interval for the population mean is

$$\bar{x} \pm t^\star_{df} SE$$

where $\bar{x}$ is the sample mean, $t^\star_{df}$ corresponds to the confidence level and degrees of freedom, and SE is the standard error as estimated by the sample.

Guided Practice 5.6 The FDA's webpage provides some data on mercury content of fish.[4] Based on a sample of 15 croaker white fish (Pacific), a sample mean and standard deviation were computed as 0.287 and 0.069 ppm (parts per million), respectively. The 15 observations ranged from 0.18 to 0.41 ppm. We will assume these observations are independent. Based on the summary statistics of the data, do you have any objections to the normality condition of the individual observations?[5]

Example 5.7 Estimate the standard error of $\bar{x} = 0.287$ ppm using the data summaries in Guided Practice 5.6. If we are to use the t-distribution to create a 90% confidence interval for the actual mean of the mercury content, identify the degrees of freedom we should use and also find $t^\star_{df}$.

The standard error: $SE = \frac{0.069}{\sqrt{15}} = 0.0178$. Degrees of freedom: $df = n - 1 = 14$.

Looking in the column where two tails is 0.100 (for a 90% confidence interval) and row $df = 14$, we identify $t^\star_{14} = 1.76$.

Guided Practice 5.8 Using the results of Guided Practice 5.6 and Example 5.7, compute a 90% confidence interval for the average mercury content of croaker white fish (Pacific).[6]

5.1.5 One sample t-tests

Is the typical US runner getting faster or slower over time? We consider this question in the context of the Cherry Blossom Race, which is a 10-mile race in Washington, DC each spring.[7]

The average time for all runners who finished the Cherry Blossom Race in 2006 was 93.29 minutes (93 minutes and about 17 seconds). We want to determine using data from 100 participants in the 2012 Cherry Blossom Race whether runners in this race are getting faster or slower, versus the other possibility that there has been no change.

Guided Practice 5.9 What are appropriate hypotheses for this context?[8]

Guided Practice 5.10 The data come from a simple random sample from less than 10% of all participants, so the observations are independent. However, should we be worried about skew in the data? See Figure 5.8 for a histogram of the differences.[9]

With independence satisfied and slight skew not a concern for this large of a sample, we can proceed with performing a hypothesis test using the t-distribution.

---

[4] www.fda.gov/food/foodborneillnesscontaminants/metals/ucm115644.htm

[5] There are no obvious outliers; all observations are within 2 standard deviations of the mean. If there is skew, it is not evident. There are no red flags for the normal model based on this (limited) information, and we do not have reason to believe the mercury content is not nearly normal in this type of fish.

[6] $\bar{x} \pm t^\star_{14} SE \rightarrow 0.287 \pm 1.76 \times 0.0178 \rightarrow (0.256, 0.318)$. We are 90% confident that the average mercury content of croaker white fish (Pacific) is between 0.256 and 0.318 ppm.

[7] www.cherryblossom.org

[8] H_0: The average 10 mile run time was the same for 2006 and 2012. $\mu = 93.29$ minutes. H_A: The average 10 mile run time for 2012 was *different* than that of 2006. $\mu \neq 93.29$ minutes.

[9] With a sample of 100, we should only be concerned if there is very strong skew. The histogram of the data suggests, at worst, slight skew.

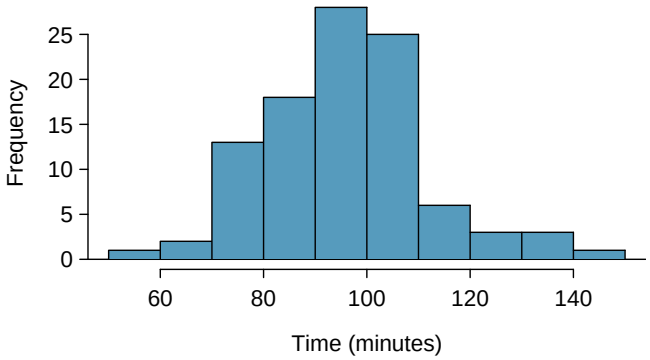


Figure 5.8: A histogram of `time` for the sample Cherry Blossom Race data.

⊙ **Guided Practice 5.11** The sample mean and sample standard deviation of the sample of 100 runners from the 2012 Cherry Blossom Race are 95.61 and 15.78 minutes, respectively. Recall that the sample size is 100. What is the p-value for the test, and what is your conclusion?[10]

When using a t-distribution, we use a T-score (same as Z-score)

To help us remember to use the t-distribution, we use a T to represent the test statistic, and we often call this a **T-score**. The Z-score and T-score are computed in the exact same way and are conceptually identical: each represents how many standard errors the observed value is from the null value.

Calculator videos

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

[10] With the conditions satisfied for the t-distribution, we can compute the standard error ($SE = 15.78/\sqrt{100} = 1.58$ and the *T-score*: $T = \frac{95.61-93.29}{1.58} = 1.47$. (There is more on this after the guided practice, but a T-score and Z-score are calculated in the same way.) For $df = 100 - 1 = 99$, we would find $T = 1.47$ to fall between the first and second column, which means the p-value is between 0.10 and 0.20 (use $df = 90$ and consider two tails since the test is two-sided). The p-value could also have been calculated more precisely with statistical software: 0.1447. Because the p-value is greater than 0.05, we do not reject the null hypothesis. That is, the data do not provide strong evidence that the average run time for the Cherry Blossom Run in 2012 is any different than the 2006 average.

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:

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

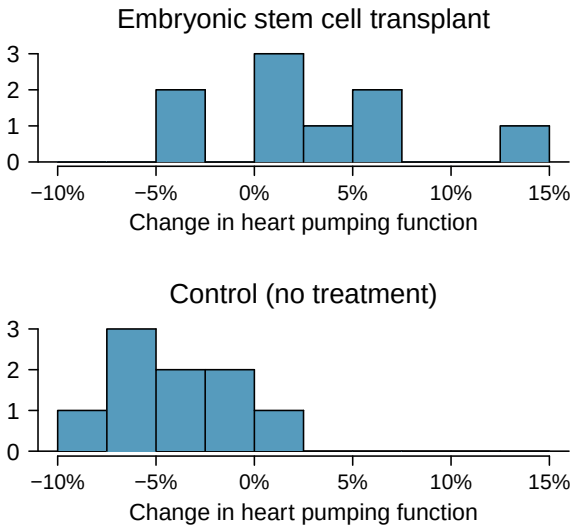


Figure 5.14: Histograms for both the embryonic stem cell group and the control group. Higher values are associated with greater improvement. We don't see any evidence of skew in these data; however, it is worth noting that skew would be difficult to detect with such a small sample.

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 ($\mu_n - \mu_s \neq 0$).

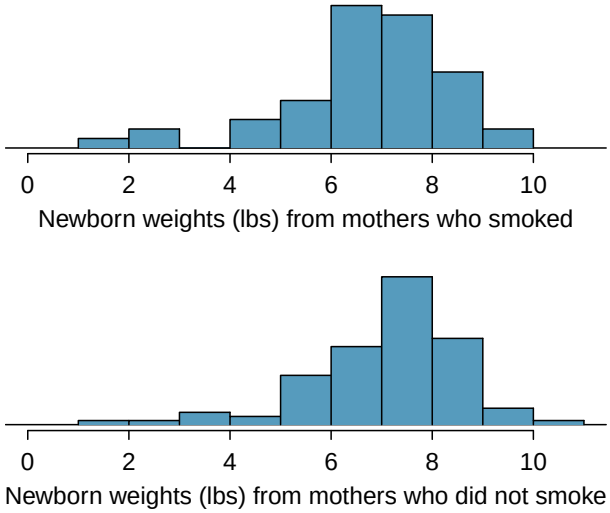


Figure 5.16: The top panel represents birth weights for infants whose mothers smoked. The bottom panel represents the birth weights for infants whose mothers who did not smoke. The distributions exhibit moderate-to-strong and strong skew, respectively.

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.

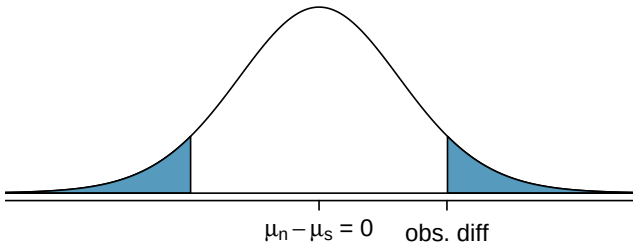
⊙ **Guided Practice 5.19** The summary statistics in Table 5.17 may be useful for this exercise. (a) What is the point estimate of the population difference, $\mu_n - \mu_s$? (b) Compute the standard error of the point estimate from part (a).[15]

[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 $n_2 - 1$: $df = 26$.

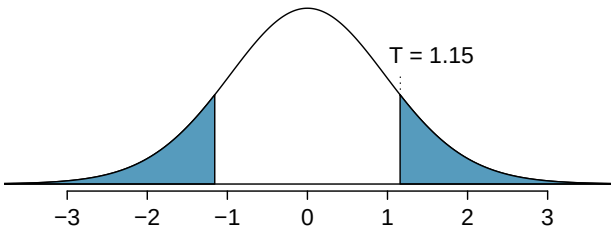


Figure 5.19: The t-distribution with 26 degrees of freedom. The shaded right tail represents values with $T \geq 1.15$. Because it is a two-sided test, we also shade the corresponding lower tail.

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.

6.2 Difference of two proportions

We would like to make conclusions about the difference in two population proportions: $p_1 - p_2$. We consider three examples. In the first, we compare the approval of the 2010 healthcare law under two different question phrasings. In the second application, we examine the efficacy of mammograms in reducing deaths from breast cancer. In the last example, a quadcopter company weighs whether to switch to a higher quality manufacturer of rotor blades.

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

6.2.1 Sample distribution of the difference of two proportions

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

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

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

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

The standard error of the difference in sample proportions is

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

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

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

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

The standard error for the difference in two proportions takes a similar form. The reasons behind this similarity are rooted in the probability theory of Section 2.4, which is described for this context in Guided Practice 5.28 on page 238.

6.2.2 Confidence intervals for $p_1 - p_2$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6.2.3 Hypothesis tests for $p_1 - p_2$

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 examine 2-proportion hypothesis test 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.[8] 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 Table 6.3.

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

Table 6.3: Summary results for breast cancer study.

⊙ **Guided Practice 6.11** Is this study an experiment or an observational study?[9]

⊙ **Guided Practice 6.12** Set up hypotheses to test whether there was a difference in breast cancer deaths in the mammogram and control groups.[10]

In Example 6.13, we will check the conditions for using the normal model to analyze the results of the study. The details are very similar to that of confidence intervals. However, this time we use a special proportion called the **pooled proportion** to check the success-failure condition:

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

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.

[8] Miller AB. 2014. *Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial.* BMJ 2014;348:g366.

[9] 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.

[10] 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, $p_{mgm} - p_{ctrl} \neq 0$.

Example 6.13 Can we use the normal model to analyze this study?

Because the patients are randomized, they can be treated as independent.

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, $\hat{p} = 0.0112$:

$$\hat{p} \times n_{mgm} = 0.0112 \times 44{,}925 = 503 \quad (1-\hat{p}) \times n_{mgm} = 0.9888 \times 44{,}925 = 44{,}422$$
$$\hat{p} \times n_{ctrl} = 0.0112 \times 44{,}910 = 503 \quad (1-\hat{p}) \times n_{ctrl} = 0.9888 \times 44{,}910 = 44{,}407$$

The success-failure condition is satisfied since all values are at least 10, and we can safely apply the normal model.

Use the pooled proportion estimate when H_0 is $p_1 - p_2 = 0$

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

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

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

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

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

In Example 6.13, the pooled proportion was used to check the success-failure condition. In the next example, we see the second place where the pooled proportion comes into play: the standard error calculation.

Example 6.14 Compute the point estimate of the difference in breast cancer death rates in the two groups, and use the pooled proportion $\hat{p} = 0.0112$ to calculate the standard error.

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

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

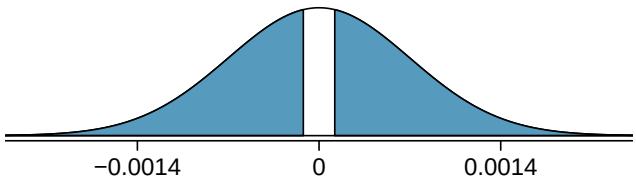
The breast cancer death rate in the mammogram group was 0.012% less than in the control group. Next, the standard error is calculated *using the pooled proportion*, $\hat{p}$:

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

Example 6.15 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 important considerations to keep in mind when reviewing the mammogram study as well as any other medical study:

- If mammograms are helpful or harmful, the data suggest the effect isn't very large. So while we do not accept the null hypothesis, we also don't have sufficient evidence to conclude that mammograms reduce or increase breast cancer deaths.
- 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.

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

Figure 6.4: 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.

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 $p_1 - p_2 = 0.1$.[11] 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 6.16** A quadcopter company is considering a new manufacturer for rotor blades. The new manufacturer would be more expensive but their higher-quality blades are more reliable, resulting in happier customers and fewer warranty claims. However, management must be convinced that the more expensive blades are worth the conversion before they approve the switch. If there is strong evidence of a more than 3% improvement in the percent of blades that pass inspection, management says they will switch suppliers, otherwise they will maintain the current supplier. Set up appropriate hypotheses for the test.[12]

● **Example 6.17** The quality control engineer from Guided Practice 6.16 collects a sample of blades, examining 1000 blades from each company and finds that 899 blades pass inspection from the current supplier and 958 pass inspection from the prospective supplier. Using these data, evaluate the hypothesis setup of Guided Practice 6.16 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

[11] We can also encounter a similar situation with a difference of two means, though no such example was given in Chapter 5 since the methods remain exactly the same in the context of sample means. On the other hand, the success-failure condition and the calculation of the standard error vary slightly in different proportion contexts.

[12] H_0: The higher-quality blades will pass inspection just 3% more frequently than the standard-quality blades. $p_{highQ} - p_{standard} = 0.03$. H_A: The higher-quality blades will pass inspection >3% more often than the standard-quality blades. $p_{highQ} - p_{standard} > 0.03$.

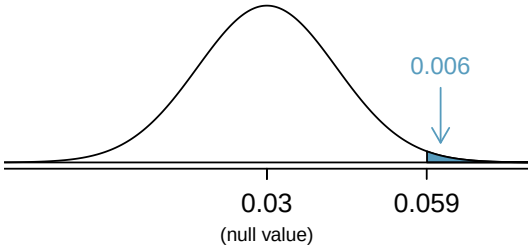


Figure 6.5: Distribution of the test statistic if the null hypothesis was true. The p-value is represented by the shaded area.

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:

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

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 6.5.

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

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