Chapter 7

Inference for numerical data

Chapter 5 introduced a framework for statistical inference based on confidence intervals and hypotheses using the normal distribution for sample proportions. In this chapter, we encounter several new point estimates and a couple new distributions. In each case, the inference ideas remain the same: determine which point estimate or test statistic is useful, identify an appropriate distribution for the point estimate or test statistic, and apply the ideas of inference.

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

7.1 One-sample means with the t-distribution

Similar to how we can model the behavior of the sample proportion $\hat{p}$ using a normal distribution, the sample mean $\bar{x}$ can also be modeled using a normal distribution when certain conditions are met. However, we'll soon learn that a new distribution, called the t-distribution, tends to be more useful when working with the sample mean. We'll first learn about this new distribution, then we'll use it to construct confidence intervals and conduct hypothesis tests for the mean.

7.1.1 The sampling distribution of $\bar{x}$

The sample mean tends to follow a normal distribution centered at the population mean, μ, when certain conditions are met. Additionally, we can compute a standard error for the sample mean using the population standard deviation σ and the sample size n.

CENTRAL LIMIT THEOREM FOR THE SAMPLE MEAN

When we collect a sufficiently large sample of n independent observations from a population with mean μ and standard deviation σ, the sampling distribution of $\bar{x}$ will be nearly normal with

$$\text{Mean} = \mu \qquad\qquad \text{Standard Error } (SE) = \frac{\sigma}{\sqrt{n}}$$

Before diving into confidence intervals and hypothesis tests using $\bar{x}$, we first need to cover two topics:

- When we modeled $\hat{p}$ using the normal distribution, certain conditions had to be satisfied. The conditions for working with $\bar{x}$ are a little more complex, and we'll spend Section 7.1.2 discussing how to check conditions for inference.
- The standard error is dependent on the population standard deviation, σ. However, we rarely know σ, and instead we must estimate it. Because this estimation is itself imperfect, we use a new distribution called the t-distribution to fix this problem, which we discuss in Section 7.1.3.

7.1.2 Evaluating the two conditions required for modeling $\bar{x}$

Two conditions are required to apply the Central Limit Theorem for a sample mean $\bar{x}$:

Independence. The sample observations must be independent, The most common way to satisfy this condition is when the sample is a simple random sample from the population. If the data come from a random process, analogous to rolling a die, this would also satisfy the independence condition.

Normality. When a sample is small, we also require that the sample observations come from a normally distributed population. We can relax this condition more and more for larger and larger sample sizes. This condition is obviously vague, making it difficult to evaluate, so next we introduce a couple rules of thumb to make checking this condition easier.

RULES OF THUMB: HOW TO PERFORM THE NORMALITY CHECK

There is no perfect way to check the normality condition, so instead we use two rules of thumb:

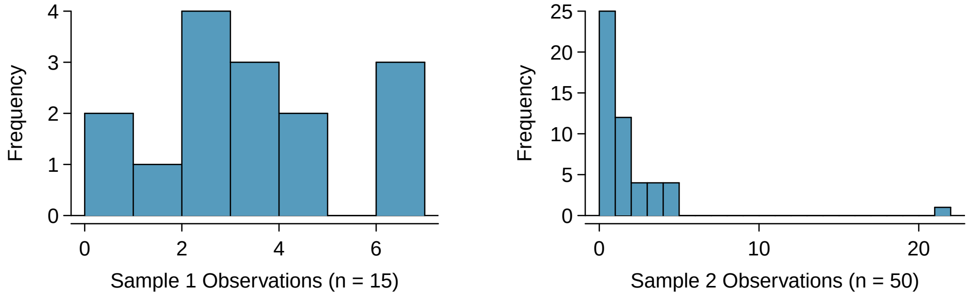
$\mathbf{n < 30}$: If the sample size n is less than 30 and there are no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.

$\mathbf{n \geq 30}$: If the sample size n is at least 30 and there are no *particularly extreme* outliers, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

In this first course in statistics, you aren't expected to develop perfect judgement on the normality condition. However, you are expected to be able to handle clear cut cases based on the rules of thumb.[1]

EXAMPLE 7.1

Consider the following two plots that come from simple random samples from different populations. Their sample sizes are $n_1 = 15$ and $n_2 = 50$.



Are the independence and normality conditions met in each case?

Each samples is from a simple random sample of its respective population, so the independence condition is satisfied. Let's next check the normality condition for each using the rule of thumb.

The first sample has fewer than 30 observations, so we are watching for any clear outliers. None are present; while there is a small gap in the histogram between 5 and 6, this gap is small and 20% of the observations in this small sample are represented in that far right bar of the histogram, so we can hardly call these clear outliers. With no clear outliers, the normality condition is reasonably met.

The second sample has a sample size greater than 30 and includes an outlier that appears to be roughly 5 times further from the center of the distribution than the next furthest observation. This is an example of a particularly extreme outlier, so the normality condition would not be satisfied.

In practice, it's typical to also do a mental check to evaluate whether we have reason to believe the underlying population would have moderate skew (if $n < 30$) or have particularly extreme outliers ($n \geq 30$) beyond what we observe in the data. For example, consider the number of followers for each individual account on Twitter, and then imagine this distribution. The large majority of accounts have built up a couple thousand followers or fewer, while a relatively tiny fraction have amassed tens of millions of followers, meaning the distribution is extremely skewed. When we know the data come from such an extremely skewed distribution, it takes some effort to understand what sample size is large enough for the normality condition to be satisfied.

7.1.3 Introducing the t-distribution

In practice, we cannot directly calculate the standard error for $\bar{x}$ since we do not know the population standard deviation, σ. We encountered a similar issue when computing the standard error for a sample proportion, which relied on the population proportion, p. Our solution in the proportion context was to use sample value in place of the population value when computing the standard error. We'll employ a similar strategy for computing the standard error of $\bar{x}$, using the sample standard deviation s in place of σ:

$$SE = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

This strategy tends to work well when we have a lot of data and can estimate σ using s accurately. However, the estimate is less precise with smaller samples, and this leads to problems when using the normal distribution to model $\bar{x}$.

[1] More nuanced guidelines would consider further relaxing the *particularly extreme outlier* check when the sample size is very large. However, we'll leave further discussion here to a future course.

We'll find it useful to use a new distribution for inference calculations called the ***t*-distribution**. A t-distribution, shown as a solid line in Figure 7.1, has a bell shape. However, its tails are thicker than the normal distribution's, meaning observations are more likely to fall beyond two standard deviations from the mean than under the normal distribution. The extra thick tails of the t-distribution are exactly the correction needed to resolve the problem of using s in place of σ in the SE calculation.

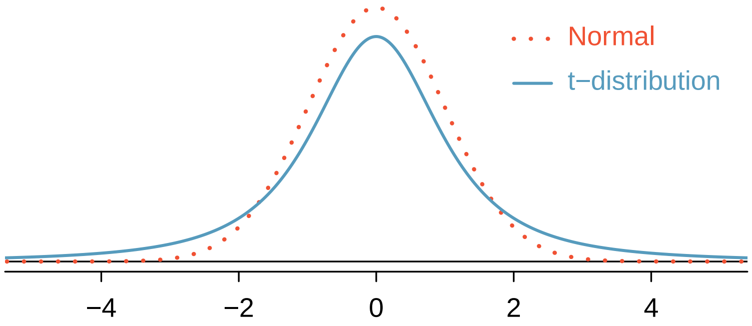


Figure 7.1: Comparison of a t-distribution and a normal distribution.

The t-distribution is always centered at zero and has a single parameter: degrees of freedom. The **degrees of freedom (*df*)** describes the precise form of the bell-shaped t-distribution. Several t-distributions are shown in Figure 7.2 in comparison to the normal distribution.

In general, we'll use a t-distribution with $df = n-1$ to model the sample mean when the sample size is n. That is, when we have more observations, the degrees of freedom will be larger and the t-distribution will look more like the standard normal distribution; when the degrees of freedom is about 30 or more, the t-distribution is nearly indistinguishable from the normal distribution.

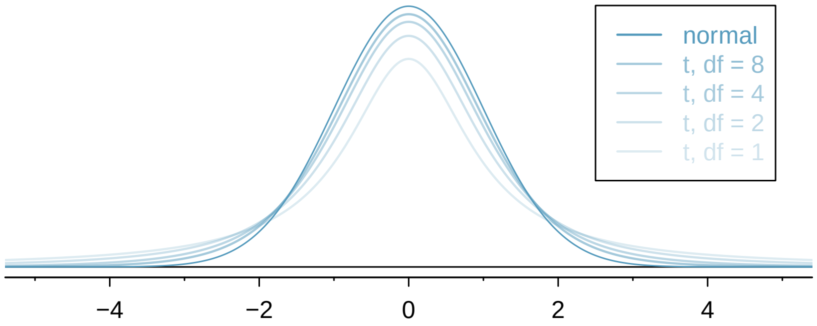


Figure 7.2: The larger the degrees of freedom, the more closely the t-distribution resembles the standard normal distribution.

DEGREES OF FREEDOM (*df*)

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

When modeling $\bar{x}$ using the t-distribution, use $df = n - 1$.

The t-distribution allows us greater flexibility than the normal distribution when analyzing numerical data. In practice, it's common to use statistical software, such as R, Python, or SAS for these analyses. Alternatively, a graphing calculator or a ***t*-table** may be used; the t-table is similar to the normal distribution table, and it may be found in Appendix C.2, which includes usage instructions and examples for those who wish to use this option. No matter the approach you choose, apply your method using the examples below to confirm your working understanding of the t-distribution.

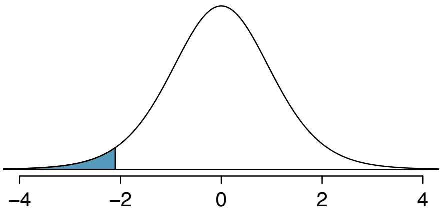


Figure 7.3: The t-distribution with 18 degrees of freedom. The area below -2.10 has been shaded.

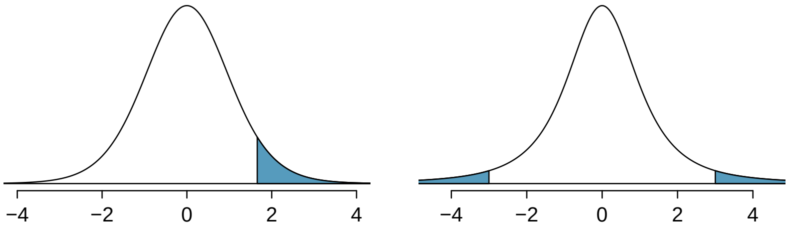


Figure 7.4: Left: The t-distribution with 20 degrees of freedom, with the area above 1.65 shaded. Right: The t-distribution with 2 degrees of freedom, with the area further than 3 units from 0 shaded.

EXAMPLE 7.2

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 7.3 and shade the area below -2.10. Using statistical software, we can obtain a precise value: 0.0250.

EXAMPLE 7.3

A t-distribution with 20 degrees of freedom is shown in the left panel of Figure 7.4. Estimate the proportion of the distribution falling above 1.65.

With a normal distribution, this would correspond to about 0.05, so we should expect the t-distribution to give us a value in this neighborhood. Using statistical software: 0.0573.

EXAMPLE 7.4

A t-distribution with 2 degrees of freedom is shown in the right panel of Figure 7.4. Estimate the proportion of the distribution falling more than 3 units from the mean (above or below).

With so few degrees of freedom, the t-distribution will give a more notably different value than the normal distribution. Under a normal distribution, the area would be about 0.003 using the 68-95-99.7 rule. For a t-distribution with $df = 2$, the area in both tails beyond 3 units totals 0.0955. This area is dramatically different than what we obtain from the normal distribution.

GUIDED PRACTICE 7.5

What proportion of the t-distribution with 19 degrees of freedom falls above -1.79 units? Use your preferred method for finding tail areas.[2]

[2] We want to find the shaded area *above* -1.79 (we leave the picture to you). The lower tail area has an area of 0.0447, so the upper area would have an area of $1 - 0.0447 = 0.9553$.

7.1.4 One sample t-confidence intervals

Let's get our first taste of applying the t-distribution in the context of an example about the mercury content of dolphin muscle. Elevated mercury concentrations are an important problem for both dolphins and other animals, like humans, who occasionally eat them.

Figure 7.5: A Risso's dolphin.

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

We will 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. The data are summarized in Figure 7.6. The minimum and maximum observed values can be used to evaluate whether or not there are clear outliers.

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

Figure 7.6: Summary of mercury content in the muscle of 19 Risso's dolphins from the Taiji area. Measurements are in micrograms of mercury per wet gram of muscle (μg/wet g).

EXAMPLE 7.6

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

The observations are a simple random sample, therefore independence is reasonable. The summary statistics in Figure 7.6 do not suggest any clear outliers, since all observations are within 2.5 standard deviations of the mean. Based on this evidence, the normality condition seems reasonable.

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:

$$\text{point estimate} \ \pm\ t^\star_{df} \times SE \qquad \rightarrow \qquad \bar{x} \ \pm\ t^\star_{df} \times \frac{s}{\sqrt{n}}$$

EXAMPLE 7.7

Using the summary statistics in Figure 7.6, compute the standard error for the average mercury content in the $n = 19$ dolphins.

We plug in s and n into the formula: $SE = s/\sqrt{n} = 2.3/\sqrt{19} = 0.528$.

The value $t^\star_{df}$ is a cutoff we obtain based on the confidence level and the t-distribution with df degrees of freedom. That cutoff is found in the same way as with a normal distribution: we find $t^\star_{df}$ such that the fraction of the t-distribution with df degrees of freedom within a distance $t^\star_{df}$ of 0 matches the confidence level of interest.

EXAMPLE 7.8

When $n = 19$, what is the appropriate degrees of freedom? Find $t^\star_{df}$ for this degrees of freedom and the confidence level of 95%

The degrees of freedom is easy to calculate: $df = n - 1 = 18$.

Using statistical software, we find the cutoff where the upper tail is equal to 2.5%: $t^\star_{18} = 2.10$. The area below -2.10 will also be equal to 2.5%. That is, 95% of the t-distribution with $df = 18$ lies within 2.10 units of 0.

EXAMPLE 7.9

Compute and interpret the 95% confidence interval for the average mercury content in Risso's dolphins.

We can construct the confidence interval as

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

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

$$\text{point estimate} \ \pm \ t^\star_{df} \times SE \qquad \rightarrow \qquad \bar{x} \ \pm \ t^\star_{df} \times \frac{s}{\sqrt{n}}$$

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

GUIDED PRACTICE 7.10

The FDA's webpage provides some data on mercury content of fish. 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?[3]

EXAMPLE 7.11

Estimate the standard error of $\bar{x} = 0.287$ ppm using the data summaries in Guided Practice 7.10. 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 and $t^\star_{df}$.

The standard error: $SE = \frac{0.069}{\sqrt{15}} = 0.0178$.

Degrees of freedom: $df = n - 1 = 14$.

Since the goal is a 90% confidence interval, we choose $t^\star_{14}$ so that the two-tail area is 0.1: $t^\star_{14} = 1.76$.

[3] The sample size is under 30, so we check for obvious outliers: since all observations are within 2 standard deviations of the mean, there are no such clear outliers.

CONFIDENCE INTERVAL FOR A SINGLE MEAN

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

Prepare. Identify $\bar{x}$, s, n, and determine what confidence level you wish to use.

Check. Verify the conditions to ensure $\bar{x}$ is nearly normal.

Calculate. If the conditions hold, compute SE, find $t^{\star}_{df}$, and construct the interval.

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

GUIDED PRACTICE 7.12

G

Using the information and results of Guided Practice 7.10 and Example 7.11, compute a 90% confidence interval for the average mercury content of croaker white fish (Pacific).[4]

GUIDED PRACTICE 7.13

G

The 90% confidence interval from Guided Practice 7.12 is 0.256 ppm to 0.318 ppm. Can we say that 90% of croaker white fish (Pacific) have mercury levels between 0.256 and 0.318 ppm?[5]

7.1.5 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.

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 2017 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 7.14

G

What are appropriate hypotheses for this context?[6]

GUIDED PRACTICE 7.15

G

The data come from a simple random sample of all participants, so the observations are independent. However, should we be worried about the normality condition? See Figure 7.7 for a histogram of the differences and evaluate if we can move forward.[7]

When completing a hypothesis test for the one-sample mean, the process is nearly identical to completing a hypothesis test for a single proportion. First, we find the Z-score using the observed value, null value, and standard error; however, we call it a **T-score** since we use a t-distribution for calculating the tail area. Then we find the p-value using the same ideas we used previously: find the one-tail area under the sampling distribution, and double it.

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

[5] No, a confidence interval only provides a range of plausible values for a population parameter, in this case the population mean. It does not describe what we might observe for individual observations.

[6] H_0: The average 10-mile run time was the same for 2006 and 2017. $\mu = 93.29$ minutes. H_A: The average 10-mile run time for 2017 was *different* than that of 2006. $\mu \neq 93.29$ minutes.

[7] With a sample of 100, we should only be concerned if there is are particularly extreme outliers. The histogram of the data doesn't show any outliers of concern (and arguably, no outliers at all).

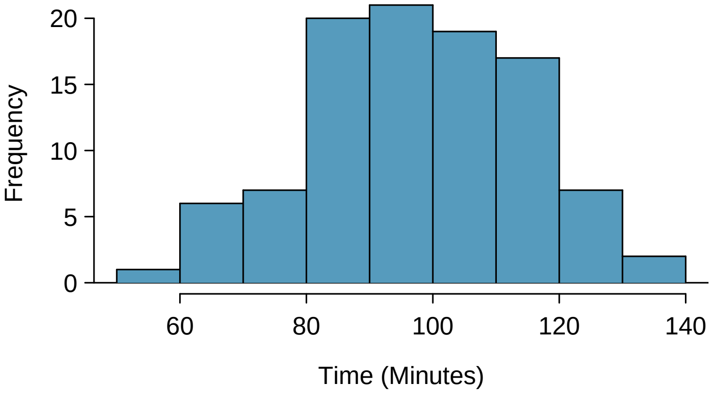


Figure 7.7: A histogram of `time` for the sample Cherry Blossom Race data.

EXAMPLE 7.16

With both the independence and normality conditions satisfied, we can proceed with a hypothesis test using the *t*-distribution. The sample mean and sample standard deviation of the sample of 100 runners from the 2017 Cherry Blossom Race are 97.32 and 16.98 minutes, respectively. Recall that the sample size is 100 and the average run time in 2006 was 93.29 minutes. Find the test statistic and p-value. What is your conclusion?

To find the test statistic (T-score), we first must determine the standard error:

$$SE = 16.98/\sqrt{100} = 1.70$$

Now we can compute the *T-score* using the sample mean (97.32), null value (93.29), and SE:

$$T = \frac{97.32 - 93.29}{1.70} = 2.37$$

For $df = 100 - 1 = 99$, we can determine using statistical software (or a *t*-table) that the one-tail area is 0.01, which we double to get the p-value: 0.02.

Because the p-value is smaller than 0.05, we reject the null hypothesis. That is, the data provide strong evidence that the average run time for the Cherry Blossom Run in 2017 is different than the 2006 average. Since the observed value is above the null value and we have rejected the null hypothesis, we would conclude that runners in the race were slower on average in 2017 than in 2006.

HYPOTHESIS TESTING FOR A SINGLE MEAN

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

Prepare. Identify the parameter of interest, list out hypotheses, identify the significance level, and identify $\bar{x}$, s, and n.

Check. Verify conditions to ensure $\bar{x}$ is nearly normal.

Calculate. If the conditions hold, compute SE, compute the T-score, and identify the p-value.

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

Exercises

7.1 Identify the critical t. An independent random sample is selected from an approximately normal population with unknown standard deviation. Find the degrees of freedom and the critical t-value ($t^\star$) for the given sample size and confidence level.

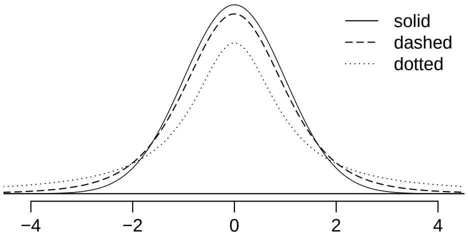
(a) $n = 6$, CL $= 90\%$

(b) $n = 21$, CL $= 98\%$

(c) $n = 29$, CL $= 95\%$

(d) $n = 12$, CL $= 99\%$

7.2 t-distribution. The figure on the right shows three unimodal and symmetric curves: the standard normal (z) distribution, the t-distribution with 5 degrees of freedom, and the t-distribution with 1 degree of freedom. Determine which is which, and explain your reasoning.



7.3 Find the p-value, Part I. An independent random sample is selected from an approximately normal population with an unknown standard deviation. Find the p-value for the given sample size and test statistic. Also determine if the null hypothesis would be rejected at $\alpha = 0.05$.

(a) $n = 11$, $T = 1.91$

(b) $n = 17$, $T = -3.45$

(c) $n = 7$, $T = 0.83$

(d) $n = 28$, $T = 2.13$

7.4 Find the p-value, Part II. An independent random sample is selected from an approximately normal population with an unknown standard deviation. Find the p-value for the given sample size and test statistic. Also determine if the null hypothesis would be rejected at $\alpha = 0.01$.

(a) $n = 26$, $T = 2.485$

(b) $n = 18$, $T = 0.5$

7.5 Working backwards, Part I. A 95% confidence interval for a population mean, μ, is given as (18.985, 21.015). This confidence interval is based on a simple random sample of 36 observations. Calculate the sample mean and standard deviation. Assume that all conditions necessary for inference are satisfied. Use the t-distribution in any calculations.

7.6 Working backwards, Part II. A 90% confidence interval for a population mean is (65, 77). The population distribution is approximately normal and the population standard deviation is unknown. This confidence interval is based on a simple random sample of 25 observations. Calculate the sample mean, the margin of error, and the sample standard deviation.

7.7 Sleep habits of New Yorkers. New York is known as "the city that never sleeps". A random sample of 25 New Yorkers were asked how much sleep they get per night. Statistical summaries of these data are shown below. The point estimate suggests New Yorkers sleep less than 8 hours a night on average. Is the result statistically significant?

n	$\bar{x}$	s	min	max
25	7.73	0.77	6.17	9.78

(a) Write the hypotheses in symbols and in words.

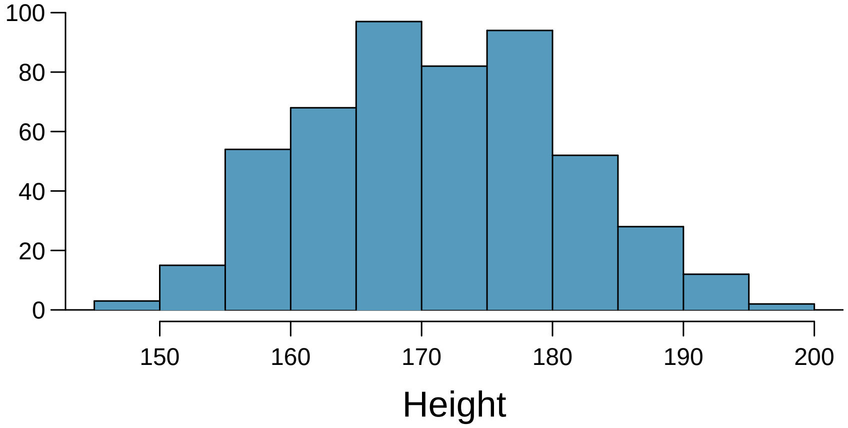
(b) Check conditions, then calculate the test statistic, T, and the associated degrees of freedom.

(c) Find and interpret the p-value in this context. Drawing a picture may be helpful.

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

(e) If you were to construct a 90% confidence interval that corresponded to this hypothesis test, would you expect 8 hours to be in the interval?

7.8 Heights of adults. Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender, for 507 physically active individuals. The histogram below shows the sample distribution of heights in centimeters.[8]



Min	147.2
Q1	163.8
Median	170.3
Mean	171.1
SD	9.4
Q3	177.8
Max	198.1

(a) What is the point estimate for the average height of active individuals? What about the median?

(b) What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?

(c) Is a person who is 1m 80cm (180 cm) tall considered unusually tall? And is a person who is 1m 55cm (155cm) considered unusually short? Explain your reasoning.

(d) The researchers take another random sample of physically active individuals. Would you expect the mean and the standard deviation of this new sample to be the ones given above? Explain your reasoning.

(e) The sample means obtained are point estimates for the mean height of all active individuals, if the sample of individuals is equivalent to a simple random sample. What measure do we use to quantify the variability of such an estimate? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

7.9 Find the mean. You are given the following hypotheses:

$$H_0 : \mu = 60$$
$$H_A : \mu \neq 60$$

We know that the sample standard deviation is 8 and the sample size is 20. For what sample mean would the p-value be equal to 0.05? Assume that all conditions necessary for inference are satisfied.

[8] G. Heinz et al. "Exploring relationships in body dimensions". In: *Journal of Statistics Education* 11.2 (2003).

7.10 $t^\star$ vs. $z^\star$. For a given confidence level, $t^\star_{df}$ is larger than $z^\star$. Explain how t^*_{df} being slightly larger than z^* affects the width of the confidence interval.

7.11 Play the piano. Georgianna claims that in a small city renowned for its music school, the average child takes less than 5 years of piano lessons. We have a random sample of 20 children from the city, with a mean of 4.6 years of piano lessons and a standard deviation of 2.2 years.

(a) Evaluate Georgianna's claim (or that the opposite might be true) using a hypothesis test.

(b) Construct a 95% confidence interval for the number of years students in this city take piano lessons, and interpret it in context of the data.

(c) Do your results from the hypothesis test and the confidence interval agree? Explain your reasoning.

7.12 Auto exhaust and lead exposure. Researchers interested in lead exposure due to car exhaust sampled the blood of 52 police officers subjected to constant inhalation of automobile exhaust fumes while working traffic enforcement in a primarily urban environment. The blood samples of these officers had an average lead concentration of 124.32 μg/l and a SD of 37.74 μg/l; a previous study of individuals from a nearby suburb, with no history of exposure, found an average blood level concentration of 35 μg/l.[9]

(a) Write down the hypotheses that would be appropriate for testing if the police officers appear to have been exposed to a different concentration of lead.

(b) Explicitly state and check all conditions necessary for inference on these data.

(c) Regardless of your answers in part (b), test the hypothesis that the downtown police officers have a higher lead exposure than the group in the previous study. Interpret your results in context.

7.13 Car insurance savings. A market researcher wants to evaluate car insurance savings at a competing company. Based on past studies he is assuming that the standard deviation of savings is \$100. He wants to collect data such that he can get a margin of error of no more than \$10 at a 95% confidence level. How large of a sample should he collect?

7.14 SAT scores. The standard deviation of SAT scores for students at a particular Ivy League college is 250 points. Two statistics students, Raina and Luke, want to estimate the average SAT score of students at this college as part of a class project. They want their margin of error to be no more than 25 points.

(a) Raina wants to use a 90% confidence interval. How large a sample should she collect?

(b) Luke wants to use a 99% confidence interval. Without calculating the actual sample size, determine whether his sample should be larger or smaller than Raina's, and explain your reasoning.

(c) Calculate the minimum required sample size for Luke.

[9] WI Mortada et al. "Study of lead exposure from automobile exhaust as a risk for nephrotoxicity among traffic policemen." In: *American journal of nephrology* 21.4 (2000), pp. 274–279.

7.2 Paired data

In an earlier edition of this textbook, we found that Amazon prices were, on average, lower than those of the UCLA Bookstore for UCLA courses in 2010. It's been several years, and many stores have adapted to the online market, so we wondered, how is the UCLA Bookstore doing today?

We sampled 201 UCLA courses. Of those, 68 required books could be found on Amazon. A portion of the data set from these courses is shown in Figure 7.8, where prices are in US dollars.

	subject	course_number	bookstore	amazon	price_difference
1	American Indian Studies	M10	47.97	47.45	0.52
2	Anthropology	2	14.26	13.55	0.71
3	Arts and Architecture	10	13.50	12.53	0.97
⋮	⋮	⋮	⋮	⋮	⋮
68	Jewish Studies	M10	35.96	32.40	3.56

Figure 7.8: Four cases of the `textbooks` data set.

7.2.1 Paired observations

Each textbook has two corresponding prices in the data set: one for the UCLA Bookstore and one for Amazon. When two sets of observations have this special correspondence, they are said to be **paired**.

PAIRED DATA

Two sets of observations are *paired* if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations. In the textbook data, we look at the differences in prices, which is represented as the `price_difference` variable in the data set. Here the differences are taken as

$$\text{UCLA Bookstore price} - \text{Amazon price}$$

It is important that we always subtract using a consistent order; here Amazon prices are always subtracted from UCLA prices. The first difference shown in Figure 7.8 is computed as $47.97 - 47.45 = 0.52$. Similarly, the second difference is computed as $14.26 - 13.55 = 0.71$, and the third is $13.50 - 12.53 = 0.97$. A histogram of the differences is shown in Figure 7.9. Using differences between paired observations is a common and useful way to analyze paired data.

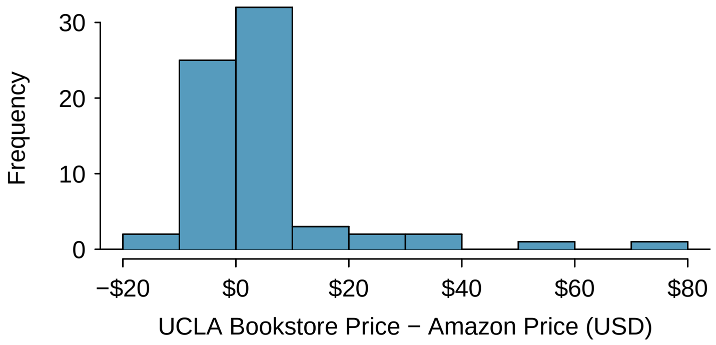


Figure 7.9: Histogram of the difference in price for each book sampled.

7.2.2 Inference for paired data

To analyze a paired data set, we simply analyze the differences. We can use the same t-distribution techniques we applied in Section 7.1.

n_{diff}	$\bar{x}_{diff}$	s_{diff}
68	3.58	13.42

Figure 7.10: Summary statistics for the 68 price differences.

EXAMPLE 7.17

Set up a hypothesis test to determine whether, on average, there is a difference between Amazon's price for a book and the UCLA bookstore's price. Also, check the conditions for whether we can move forward with the test using the t-distribution.

We are considering two scenarios: there is no difference or there is some difference in average prices.

H_0: $\mu_{diff} = 0$. There is no difference in the average textbook price.

H_A: $\mu_{diff} \neq 0$. There is a difference in average prices.

Next, we check the independence and normality conditions. The observations are based on a simple random sample, so independence is reasonable. While there are some outliers, $n = 68$ and none of the outliers are particularly extreme, so the normality of $\bar{x}$ is satisfied. With these conditions satisfied, we can move forward with the t-distribution.

EXAMPLE 7.18

Complete the hypothesis test started in Example 7.17.

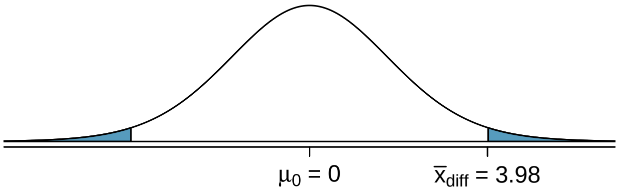
To compute the test compute the standard error associated with $\bar{x}_{diff}$ using the standard deviation of the differences ($s_{diff} = 13.42$) and the number of differences ($n_{diff} = 68$):

$$SE_{\bar{x}_{diff}} = \frac{s_{diff}}{\sqrt{n_{diff}}} = \frac{13.42}{\sqrt{68}} = 1.63$$

The test statistic is the T-score of $\bar{x}_{diff}$ under the null condition that the actual mean difference is 0:

$$T = \frac{\bar{x}_{diff} - 0}{SE_{\bar{x}_{diff}}} = \frac{3.58 - 0}{1.63} = 2.20$$

To visualize the p-value, the sampling distribution of $\bar{x}_{diff}$ is drawn as though H_0 is true, and the p-value is represented by the two shaded tails:



The degrees of freedom is $df = 68 - 1 = 67$. Using statistical software, we find the one-tail area of 0.0156. Doubling this area gives the p-value: 0.0312.

Because the p-value is less than 0.05, we reject the null hypothesis. Amazon prices are, on average, lower than the UCLA Bookstore prices for UCLA courses.

G

GUIDED PRACTICE 7.19

Create a 95% confidence interval for the average price difference between books at the UCLA bookstore and books on Amazon.[10]

G

GUIDED PRACTICE 7.20

We have strong evidence that Amazon is, on average, less expensive. How should this conclusion affect UCLA student buying habits? Should UCLA students always buy their books on Amazon?[11]

[10] Conditions have already verified and the standard error computed in Example 7.17. To find the interval, identify $t^{\star}_{67}$ using statistical software or the t-table ($t^{\star}_{67} = 2.00$), and plug it, the point estimate, and the standard error into the confidence interval formula:

$$\text{point estimate} \ \pm\ z^{\star} \times SE \quad\to\quad 3.58 \ \pm\ 2.00 \times 1.63 \quad\to\quad (0.32, 6.84)$$

We are 95% confident that Amazon is, on average, between \$0.32 and \$6.84 less expensive than the UCLA Bookstore for UCLA course books.

[11] The average price difference is only mildly useful for this question. Examine the distribution shown in Figure 7.9. There are certainly a handful of cases where Amazon prices are far below the UCLA Bookstore's, which suggests it is worth checking Amazon (and probably other online sites) before purchasing. However, in many cases the Amazon price is above what the UCLA Bookstore charges, and most of the time the price isn't that different. Ultimately, if getting a book immediately from the bookstore is notably more convenient, e.g. to get started on reading or homework, it's likely a good idea to go with the UCLA Bookstore unless the price difference on a specific book happens to be quite large.

For reference, this is a very different result from what we (the authors) had seen in a similar data set from 2010. At that time, Amazon prices were almost uniformly lower than those of the UCLA Bookstore's and by a large margin, making the case to use Amazon over the UCLA Bookstore quite compelling at that time. Now we frequently check multiple websites to find the best price.

Exercises

7.15 Air quality. Air quality measurements were collected in a random sample of 25 country capitals in 2013, and then again in the same cities in 2014. We would like to use these data to compare average air quality between the two years. Should we use a paired or non-paired test? Explain your reasoning.

7.16 True / False: paired. Determine if the following statements are true or false. If false, explain.

(a) In a paired analysis we first take the difference of each pair of observations, and then we do inference on these differences.

(b) Two data sets of different sizes cannot be analyzed as paired data.

(c) Consider two sets of data that are paired with each other. Each observation in one data set has a natural correspondence with exactly one observation from the other data set.

(d) Consider two sets of data that are paired with each other. Each observation in one data set is subtracted from the average of the other data set's observations.

7.17 Paired or not? Part I. In each of the following scenarios, determine if the data are paired.

(a) Compare pre- (beginning of semester) and post-test (end of semester) scores of students.

(b) Assess gender-related salary gap by comparing salaries of randomly sampled men and women.

(c) Compare artery thicknesses at the beginning of a study and after 2 years of taking Vitamin E for the same group of patients.

(d) Assess effectiveness of a diet regimen by comparing the before and after weights of subjects.

7.18 Paired or not? Part II. In each of the following scenarios, determine if the data are paired.

(a) We would like to know if Intel's stock and Southwest Airlines' stock have similar rates of return. To find out, we take a random sample of 50 days, and record Intel's and Southwest's stock on those same days.

(b) We randomly sample 50 items from Target stores and note the price for each. Then we visit Walmart and collect the price for each of those same 50 items.

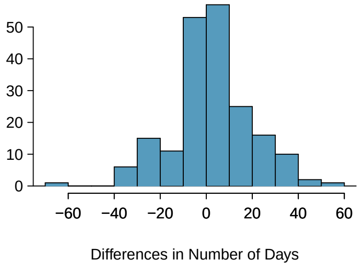
(c) A school board would like to determine whether there is a difference in average SAT scores for students at one high school versus another high school in the district. To check, they take a simple random sample of 100 students from each high school.

7.19 Global warming, Part I. Let's consider a limited set of climate data, examining temperature differences in 1948 vs 2018. We sampled 197 locations from the National Oceanic and Atmospheric Administration's (NOAA) historical data, where the data was available for both years of interest. We want to know: were there more days with temperatures exceeding 90°F in 2018 or in 1948?[12] The difference in number of days exceeding 90°F (number of days in 2018 - number of days in 1948) was calculated for each of the 197 locations. The average of these differences was 2.9 days with a standard deviation of 17.2 days. We are interested in determining whether these data provide strong evidence that there were more days in 2018 that exceeded 90°F from NOAA's weather stations.

(a) Is there a relationship between the observations collected in 1948 and 2018? Or are the observations in the two groups independent? Explain.

(b) Write hypotheses for this research in symbols and in words.

(c) Check the conditions required to complete this test. A histogram of the differences is given to the right.



(d) Calculate the test statistic and find the p-value.

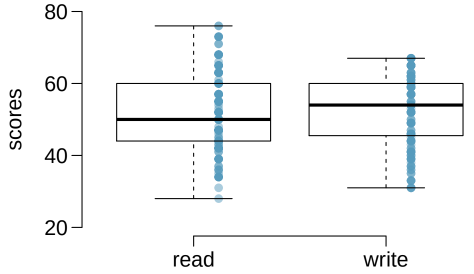
(e) Use $\alpha = 0.05$ to evaluate the test, and interpret your conclusion in context.

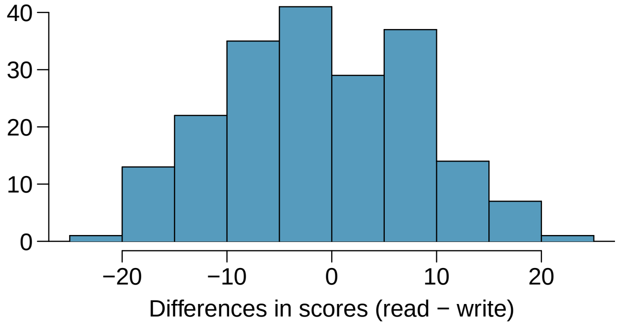
(f) What type of error might we have made? Explain in context what the error means.

(g) Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the number of days exceeding 90°F from 1948 and 2018 to include 0? Explain your reasoning.

[12] NOAA, www.ncdc.noaa.gov/cdo-web/datasets, April 24, 2019.

7.20 High School and Beyond, Part I. The National Center of Education Statistics conducted a survey of high school seniors, collecting test data on reading, writing, and several other subjects. Here we examine a simple random sample of 200 students from this survey. Side-by-side box plots of reading and writing scores as well as a histogram of the differences in scores are shown below.





(a) Is there a clear difference in the average reading and writing scores?

(b) Are the reading and writing scores of each student independent of each other?

(c) Create hypotheses appropriate for the following research question: is there an evident difference in the average scores of students in the reading and writing exam?

(d) Check the conditions required to complete this test.

(e) The average observed difference in scores is $\bar{x}_{read-write} = -0.545$, and the standard deviation of the differences is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams?

(f) What type of error might we have made? Explain what the error means in the context of the application.

(g) Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the reading and writing scores to include 0? Explain your reasoning.

7.21 Global warming, Part II. We considered the change in the number of days exceeding 90°F from 1948 and 2018 at 197 randomly sampled locations from the NOAA database in Exercise 7.19. The mean and standard deviation of the reported differences are 2.9 days and 17.2 days.

(a) Calculate a 90% confidence interval for the average difference between number of days exceeding 90°F between 1948 and 2018. We've already checked the conditions for you.

(b) Interpret the interval in context.

(c) Does the confidence interval provide convincing evidence that there were more days exceeding 90°F in 2018 than in 1948 at NOAA stations? Explain.

7.22 High school and beyond, Part II. We considered the differences between the reading and writing scores of a random sample of 200 students who took the High School and Beyond Survey in Exercise 7.20. The mean and standard deviation of the differences are $\bar{x}_{read-write} = -0.545$ and 8.887 points.

(a) Calculate a 95% confidence interval for the average difference between the reading and writing scores of all students.

(b) Interpret this interval in context.

(c) Does the confidence interval provide convincing evidence that there is a real difference in the average scores? Explain.

7.3 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, $\bar{x}_1 - \bar{x}_2$, and a new standard error formula. Other than these two differences, the details are almost identical to the one-mean procedures.

We apply these methods in three contexts: determining whether stem cells can improve heart function, exploring the relationship between pregnant womens' smoking habits and birth weights of newborns, and exploring whether there is statistically significant evidence that one variation 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?"

7.3.1 Confidence interval for a difference of means

Does treatment using embryonic stem cells (ESCs) help improve heart function following a heart attack? Figure 7.11 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. Figure 7.12 provides histograms of the two data sets. 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.

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

Figure 7.11: Summary statistics of the embryonic stem cell study.

The point estimate of the difference in the heart pumping variable is straightforward to find: it is the difference in the sample means.

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

For the question of whether we can model this difference using a t-distribution, we'll need to check new conditions. Like the 2-proportion cases, we will require a more robust version of independence so we are confident the two groups are also independent. Secondly, we also check for normality in each group separately, which in practice is a check for outliers.

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

- *Independence, extended.* The data are independent within and between the two groups, e.g. the data come from independent random samples or from a randomized experiment.
- *Normality.* We check the outliers rules of thumb for each group separately.

The standard error may be computed as

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The official formula for the degrees of freedom is quite complex and is generally computed using software, so instead you may use the smaller of $n_1 - 1$ and $n_2 - 1$ for the degrees of freedom if software isn't readily available.

EXAMPLE 7.21

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

First, we check for independence. Because the sheep were randomized into the groups, independence within and between groups is satisfied.

Figure 7.12 does not reveal any clear outliers in either group. (The ESC group does look a bit more variability, but this is not the same as having clear outliers.)

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

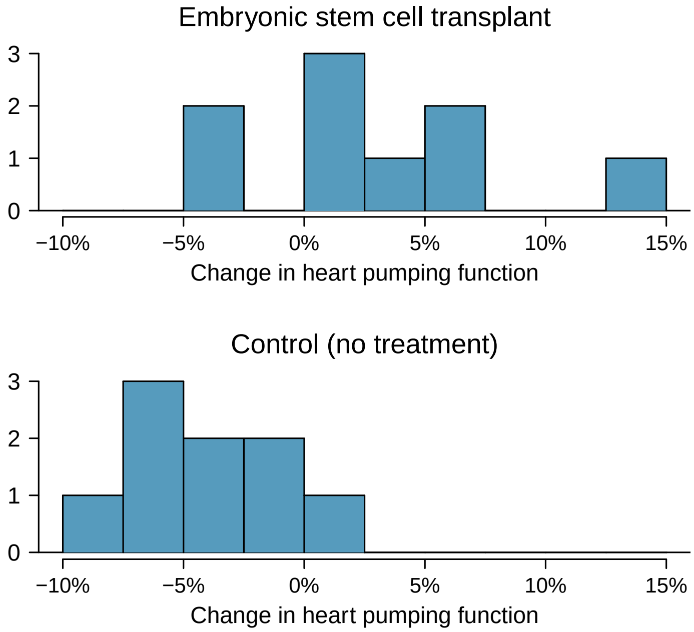


Figure 7.12: Histograms for both the embryonic stem cell and control group.

As with the one-sample case, we always compute the standard error using sample standard deviations rather than population standard deviations:

$$SE = \sqrt{\frac{s_{esc}^2}{n_{esc}} + \frac{s_{control}^2}{n_{control}}} = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95$$

Generally, we use statistical software to find the appropriate degrees of freedom, or if software isn't available, we can use the smaller of $n_1 - 1$ and $n_2 - 1$ for the degrees of freedom, e.g. if using a t-table to find tail areas. For transparency in the Examples and Guided Practice, we'll use the latter approach for finding df; in the case of the ESC example, this means we'll use $df = 8$.

EXAMPLE 7.22

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 that we computed earlier calculations:

$$\bar{x}_{esc} - \bar{x}_{control} = 7.83 \qquad\qquad SE = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95$$

Using $df = 8$, we can identify the critical value of $t^\star_8 = 2.31$ for a 95% confidence interval. Finally, we can enter the values into the confidence interval formula:

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

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%.

As with past statistical inference applications, there is a well-trodden procedure.

Prepare. Retrieve critical contextual information, and if appropriate, set up hypotheses.

Check. Ensure the required conditions are reasonably satisfied.

Calculate. Find the standard error, and then construct a confidence interval, or if conducting a hypothesis test, find a test statistic and p-value.

Conclude. Interpret the results in the context of the application.

The details change a little from one setting to the next, but this general approach remain the same.

7.3.2 Hypothesis tests for the difference of two means

A data set called `ncbirths` 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 Figure 7.13. 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.

	fage	mage	weeks	weight	sex	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

Figure 7.13: Four cases from the `ncbirths` data set. The value "NA", shown for the first two entries of the first variable, indicates that piece of data is missing.

EXAMPLE 7.23

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

E

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

We check the two conditions necessary to model the difference in sample means using the t-distribution.

- Because the data come from a simple random sample, the observations are independent, both within and between samples.
- With both data sets over 30 observations, we inspect the data in Figure 7.14 for any particularly extreme outliers and find none.

Since both conditions are satisfied, the difference in sample means may be modeled using a t-distribution.

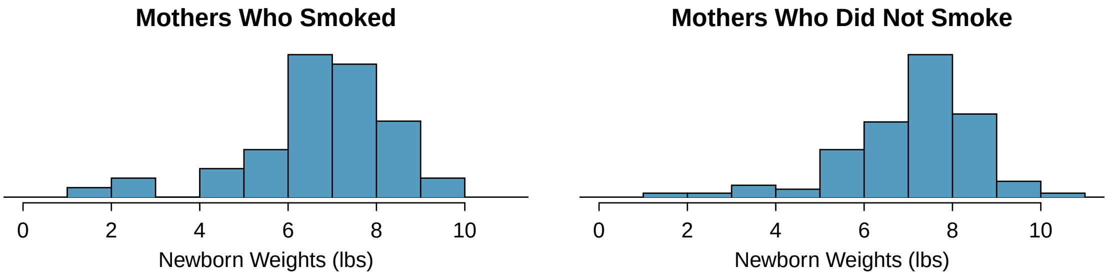


Figure 7.14: The left panel represents birth weights for infants whose mothers smoked. The right panel represents the birth weights for infants whose mothers who did not smoke.

GUIDED PRACTICE 7.24

G

The summary statistics in Figure 7.15 may be useful for this Guided Practice.[13]

(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).

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

Figure 7.15: Summary statistics for the `ncbirths` data set.

[13] (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 calculated using the standard error formula:

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

E

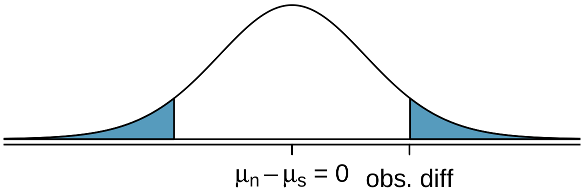
EXAMPLE 7.25

Complete the hypothesis test started in Example 7.23 and Guided Practice 7.24. Use a significance level of $\alpha = 0.05$. For reference, $\bar{x}_n - \bar{x}_s = 0.40$, $SE = 0.26$, and the sample sizes were $n_n = 100$ and $n_s = 50$.

We can find the test statistic for this test using the values from Guided Practice 7.24:

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

The p-value is represented by the two shaded tails in the following plot:



We find the single tail area using software (or the t-table in Appendix C.2). We'll use the smaller of $n_n - 1 = 99$ and $n_s - 1 = 49$ as the degrees of freedom: $df = 49$. The one tail area is 0.065; doubling this value gives the two-tail area and p-value, 0.135.

The p-value is larger than the significance value, 0.05, so we do not 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.

G

GUIDED PRACTICE 7.26

We've seen much research suggesting smoking is harmful during pregnancy, so how could we fail to reject the null hypothesis in Example 7.25? [14]

G

GUIDED PRACTICE 7.27

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?[15]

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.[16]

[14] It is possible that there is a difference but we did not detect it. If there is a difference, we made a Type 2 Error.

[15] 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. In fact, this is exactly what we would find if we examined a larger data set!

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

7.3.3 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 Figure 7.16. 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

Figure 7.16: Summary statistics of scores for each exam version.

GUIDED PRACTICE 7.28

Construct hypotheses to evaluate whether the observed difference in sample means, $\bar{x}_A - \bar{x}_B = 5.3$, is due to chance. We will later evaluate these hypotheses using $\alpha = 0.01$.[17]

GUIDED PRACTICE 7.29

To evaluate the hypotheses in Guided Practice 7.28 using the t-distribution, we must first verify conditions.[18]

(a) Does it seem reasonable that the scores are independent?

(b) Any concerns about outliers?

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

[17] 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. $\mu_A - \mu_B \neq 0$.

[18] (a) Since the exams were shuffled, the "treatment" in this case was randomly assigned, so independence within and between groups is satisfied. (b) The summary statistics suggest the data are roughly symmetric about the mean, and the min/max values don't suggest any outliers of concern.

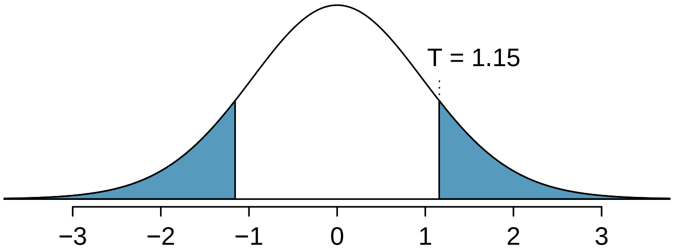


Figure 7.17: The t-distribution with 26 degrees of freedom and the p-value from exam example represented as the shaded areas.

E

EXAMPLE 7.30

Identify the p-value depicted in Figure 7.17 using $df = 26$, and provide a conclusion in the context of the case study.

Using software, we can find the one-tail area (0.13) and then double this value to get the two-tail area, which is the p-value: 0.26. (Alternatively, we could use the t-table in Appendix C.2.)

In Guided Practice 7.28, we specified that we would use $\alpha = 0.01$. Since the p-value is larger than α, 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.

7.3.4 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 very 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:

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

where n_1 and n_2 are the sample sizes, as before. To use this new statistic, we substitute s_{pooled}^2 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:

$$df = n_1 + n_2 - 2$$

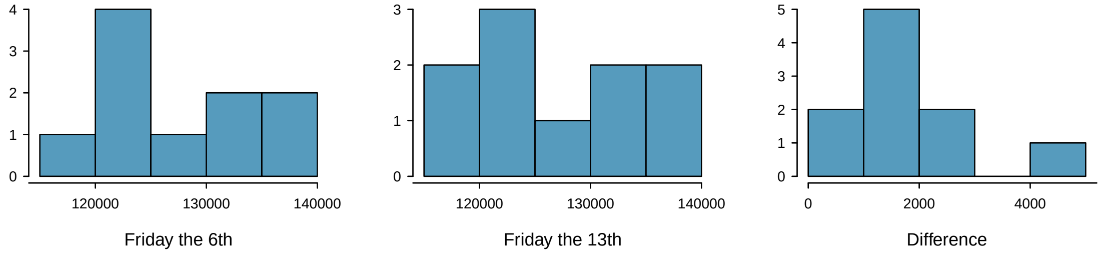
The benefits of pooling the standard deviation are realized through obtaining a better estimate of the standard deviation for each group and using a larger degrees of freedom parameter for the t-distribution. Both of these changes may permit a more accurate model of the sampling distribution of $\bar{x}_1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

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.

Exercises

7.23 Friday the 13th, Part I. In the early 1990's, researchers in the UK collected data on traffic flow, number of shoppers, and traffic accident related emergency room admissions on Friday the 13th and the previous Friday, Friday the 6th. The histograms below show the distribution of number of cars passing by a specific intersection on Friday the 6th and Friday the 13th for many such date pairs. Also given are some sample statistics, where the difference is the number of cars on the 6th minus the number of cars on the 13th.[19]



	6th	13th	Diff.
$\bar{x}$	128,385	126,550	1,835
s	7,259	7,664	1,176
n	10	10	10

(a) Are there any underlying structures in these data that should be considered in an analysis? Explain.

(b) What are the hypotheses for evaluating whether the number of people out on Friday the 6th is different than the number out on Friday the 13th?

(c) Check conditions to carry out the hypothesis test from part (b).

(d) Calculate the test statistic and the p-value.

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

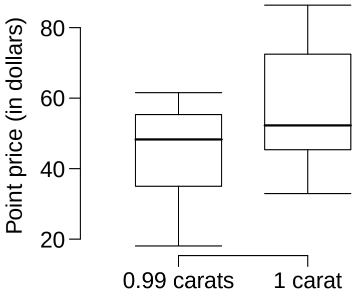
(f) Interpret the p-value in this context.

(g) What type of error might have been made in the conclusion of your test? Explain.

7.24 Diamonds, Part I. Prices of diamonds are determined by what is known as the 4 Cs: cut, clarity, color, and carat weight. The prices of diamonds go up as the carat weight increases, but the increase is not smooth. For example, the difference between the size of a 0.99 carat diamond and a 1 carat diamond is undetectable to the naked human eye, but the price of a 1 carat diamond tends to be much higher than the price of a 0.99 diamond. In this question we use two random samples of diamonds, 0.99 carats and 1 carat, each sample of size 23, and compare the average prices of the diamonds. In order to be able to compare equivalent units, we first divide the price for each diamond by 100 times its weight in carats. That is, for a 0.99 carat diamond, we divide the price by 99. For a 1 carat diamond, we divide the price by 100. The distributions and some sample statistics are shown below.[20]

Conduct a hypothesis test to evaluate if there is a difference between the average standardized prices of 0.99 and 1 carat diamonds. Make sure to state your hypotheses clearly, check relevant conditions, and interpret your results in context of the data.

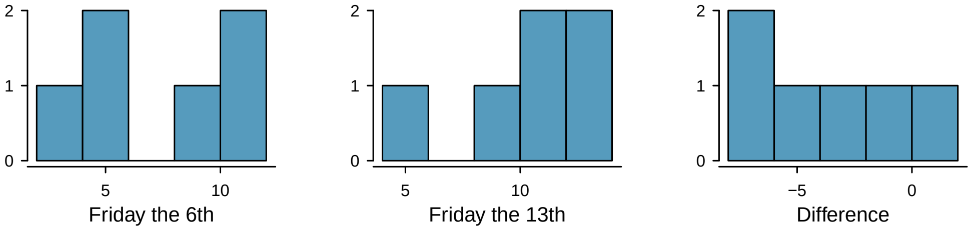
	0.99 carats	1 carat
Mean	\$44.51	\$56.81
SD	\$13.32	\$16.13
n	23	23



[19] T.J. Scanlon et al. "Is Friday the 13th Bad For Your Health?" In: *BMJ* 307 (1993), pp. 1584–1586.

[20] H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

7.25 Friday the 13th, Part II. The Friday the 13^{th} study reported in Exercise 7.23 also provides data on traffic accident related emergency room admissions. The distributions of these counts from Friday the 6^{th} and Friday the 13^{th} are shown below for six such paired dates along with summary statistics. You may assume that conditions for inference are met.



	6^{th}	13^{th}	diff
Mean	7.5	10.83	-3.33
SD	3.33	3.6	3.01
n	6	6	6

(a) Conduct a hypothesis test to evaluate if there is a difference between the average numbers of traffic accident related emergency room admissions between Friday the 6^{th} and Friday the 13^{th}.

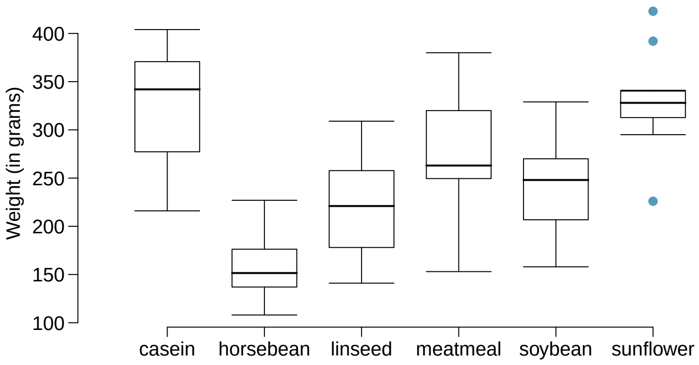
(b) Calculate a 95% confidence interval for the difference between the average numbers of traffic accident related emergency room admissions between Friday the 6^{th} and Friday the 13^{th}.

(c) The conclusion of the original study states, "Friday 13th is unlucky for some. The risk of hospital admission as a result of a transport accident may be increased by as much as 52%. Staying at home is recommended." Do you agree with this statement? Explain your reasoning.

7.26 Diamonds, Part II. In Exercise 7.24, we discussed diamond prices (standardized by weight) for diamonds with weights 0. 99 carats and 1 carat. See the table for summary statistics, and then construct a 95% confidence interval for the average difference between the standardized prices of 0.99 and 1 carat diamonds. You may assume the conditions for inference are met.

	0.99 carats	1 carat
Mean	\$44.51	\$56.81
SD	\$13.32	\$16.13
n	23	23

7.27 Chicken diet and weight, Part I. Chicken farming is a multi-billion dollar industry, and any methods that increase the growth rate of young chicks can reduce consumer costs while increasing company profits, possibly by millions of dollars. An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. Below are some summary statistics from this data set along with box plots showing the distribution of weights by feed type.[21]



	Mean	SD	n
casein	323.58	64.43	12
horsebean	160.20	38.63	10
linseed	218.75	52.24	12
meatmeal	276.91	64.90	11
soybean	246.43	54.13	14
sunflower	328.92	48.84	12

(a) Describe the distributions of weights of chickens that were fed linseed and horsebean.

(b) Do these data provide strong evidence that the average weights of chickens that were fed linseed and horsebean are different? Use a 5% significance level.

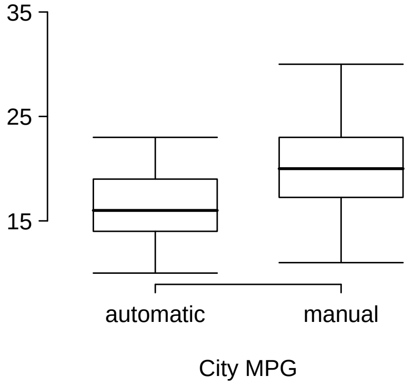
(c) What type of error might we have committed? Explain.

(d) Would your conclusion change if we used $\alpha = 0.01$?

[21] Chicken Weights by Feed Type, from the `datasets` package in R..

7.28 Fuel efficiency of manual and automatic cars, Part I. Each year the US Environmental Protection Agency (EPA) releases fuel economy data on cars manufactured in that year. Below are summary statistics on fuel efficiency (in miles/gallon) from random samples of cars with manual and automatic transmissions. Do these data provide strong evidence of a difference between the average fuel efficiency of cars with manual and automatic transmissions in terms of their average city mileage? Assume that conditions for inference are satisfied.[22]

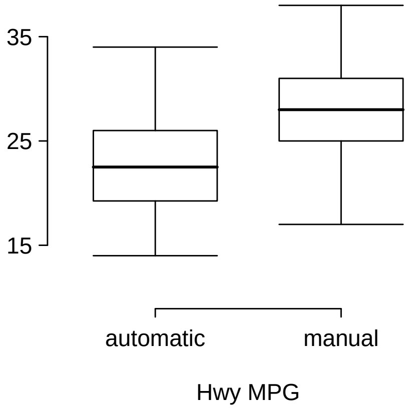
	City MPG	
	Automatic	Manual
Mean	16.12	19.85
SD	3.58	4.51
n	26	26



7.29 Chicken diet and weight, Part II. Casein is a common weight gain supplement for humans. Does it have an effect on chickens? Using data provided in Exercise 7.27, test the hypothesis that the average weight of chickens that were fed casein is different than the average weight of chickens that were fed soybean. If your hypothesis test yields a statistically significant result, discuss whether or not the higher average weight of chickens can be attributed to the casein diet. Assume that conditions for inference are satisfied.

7.30 Fuel efficiency of manual and automatic cars, Part II. The table provides summary statistics on highway fuel economy of the same 52 cars from Exercise 7.28. Use these statistics to calculate a 98% confidence interval for the difference between average highway mileage of manual and automatic cars, and interpret this interval in the context of the data.[23]

	Hwy MPG	
	Automatic	Manual
Mean	22.92	27.88
SD	5.29	5.01
n	26	26



[22] U.S. Department of Energy, Fuel Economy Data, 2012 Datafile.
[23] U.S. Department of Energy, Fuel Economy Data, 2012 Datafile.

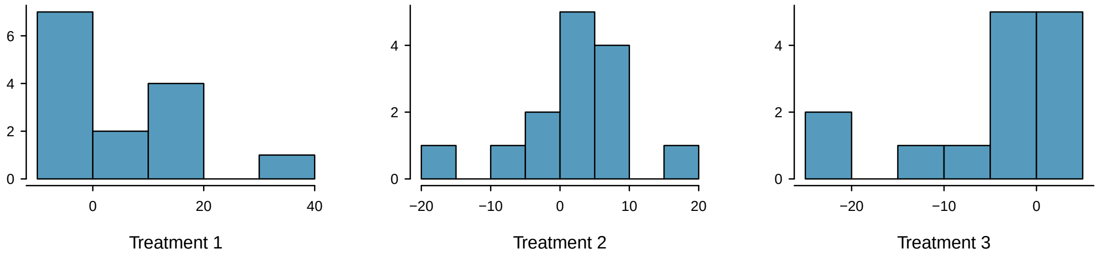
7.31 Prison isolation experiment, Part I. Subjects from Central Prison in Raleigh, NC, volunteered for an experiment involving an "isolation" experience. The goal of the experiment was to find a treatment that reduces subjects' psychopathic deviant T scores. This score measures a person's need for control or their rebellion against control, and it is part of a commonly used mental health test called the Minnesota Multiphasic Personality Inventory (MMPI) test. The experiment had three treatment groups:

(1) Four hours of sensory restriction plus a 15 minute "therapeutic" tape advising that professional help is available.

(2) Four hours of sensory restriction plus a 15 minute "emotionally neutral" tape on training hunting dogs.

(3) Four hours of sensory restriction but no taped message.

Forty-two subjects were randomly assigned to these treatment groups, and an MMPI test was administered before and after the treatment. Distributions of the differences between pre and post treatment scores (pre - post) are shown below, along with some sample statistics. Use this information to independently test the effectiveness of each treatment. Make sure to clearly state your hypotheses, check conditions, and interpret results in the context of the data.[24]



	Tr 1	Tr 2	Tr 3
Mean	6.21	2.86	-3.21
SD	12.3	7.94	8.57
n	14	14	14

7.32 True / False: comparing means. Determine if the following statements are true or false, and explain your reasoning for statements you identify as false.

(a) When comparing means of two samples where $n_1 = 20$ and $n_2 = 40$, we can use the normal model for the difference in means since $n_2 \geq 30$.

(b) As the degrees of freedom increases, the t-distribution approaches normality.

(c) We use a pooled standard error for calculating the standard error of the difference between means when sample sizes of groups are equal to each other.

[24] Prison isolation experiment, stat.duke.edu/resources/datasets/prison-isolation.

7.4 Power calculations for a difference of means

Often times in experiment planning, there are two competing considerations:

- We want to collect enough data that we can detect important effects.
- Collecting data can be expensive, and in experiments involving people, there may be some risk to patients.

In this section, we focus on the context of a clinical trial, which is a health-related experiment where the subject are people, and we will determine an appropriate sample size where we can be 80% sure that we would detect any practically important effects.[25]

7.4.1 Going through the motions of a test

We're going to go through the motions of a hypothesis test. This will help us frame our calculations for determining an appropriate sample size for the study.

EXAMPLE 7.31

Suppose a pharmaceutical company has developed a new drug for lowering blood pressure, and they are preparing a clinical trial (experiment) to test the drug's effectiveness. They recruit people who are taking a particular standard blood pressure medication. People in the control group will continue to take their current medication through generic-looking pills to ensure blinding. Write down the hypotheses for a two-sided hypothesis test in this context.

Generally, clinical trials use a two-sided alternative hypothesis, so below are suitable hypotheses for this context:

H_0: The new drug performs exactly as well as the standard medication.
$\mu_{trmt} - \mu_{ctrl} = 0$.

H_A: The new drug's performance differs from the standard medication.
$\mu_{trmt} - \mu_{ctrl} \neq 0$.

EXAMPLE 7.32

The researchers would like to run the clinical trial on patients with systolic blood pressures between 140 and 180 mmHg. Suppose previously published studies suggest that the standard deviation of the patients' blood pressures will be about 12 mmHg and the distribution of patient blood pressures will be approximately symmetric.[26] If we had 100 patients per group, what would be the approximate standard error for $\bar{x}_{trmt} - \bar{x}_{ctrl}$?

The standard error is calculated as follows:

$$SE_{\bar{x}_{trmt}-\bar{x}_{ctrl}} = \sqrt{\frac{s^2_{trmt}}{n_{trmt}} + \frac{s^2_{ctrl}}{n_{ctrl}}} = \sqrt{\frac{12^2}{100} + \frac{12^2}{100}} = 1.70$$

This may be an imperfect estimate of $SE_{\bar{x}_{trmt}-\bar{x}_{ctrl}}$, since the standard deviation estimate we used may not be perfectly correct for this group of patients. However, it is sufficient for our purposes.

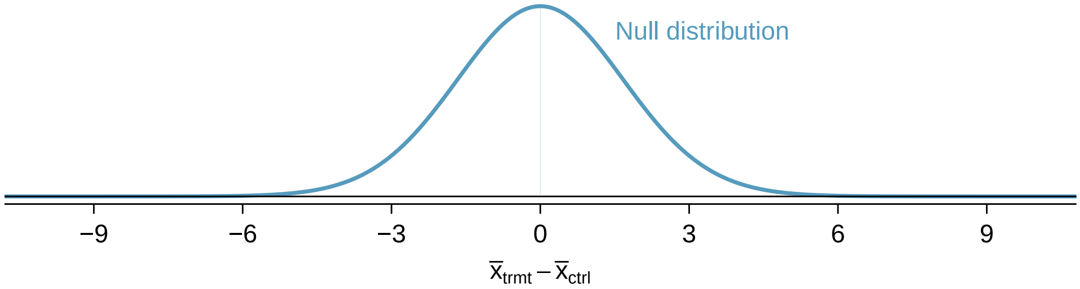
[25] Even though we don't cover it explicitly, similar sample size planning is also helpful for observational studies.

[26] In this particular study, we'd generally measure each patient's blood pressure at the beginning and end of the study, and then the outcome measurement for the study would be the average change in blood pressure. That is, both μ_{trmt} and μ_{ctrl} would represent average differences. This is what you might think of as a 2-sample paired testing structure, and we'd analyze it exactly just like a hypothesis test for a difference in the average change for patients. In the calculations we perform here, we'll suppose that 12 mmHg is the predicted standard deviation of a patient's blood pressure difference over the course of the study.

EXAMPLE 7.33

What does the null distribution of $\bar{x}_{trmt} - \bar{x}_{ctrl}$ look like?

The degrees of freedom are greater than 30, so the distribution of $\bar{x}_{trmt} - \bar{x}_{ctrl}$ will be approximately normal. The standard deviation of this distribution (the standard error) would be about 1.70, and under the null hypothesis, its mean would be 0.



EXAMPLE 7.34

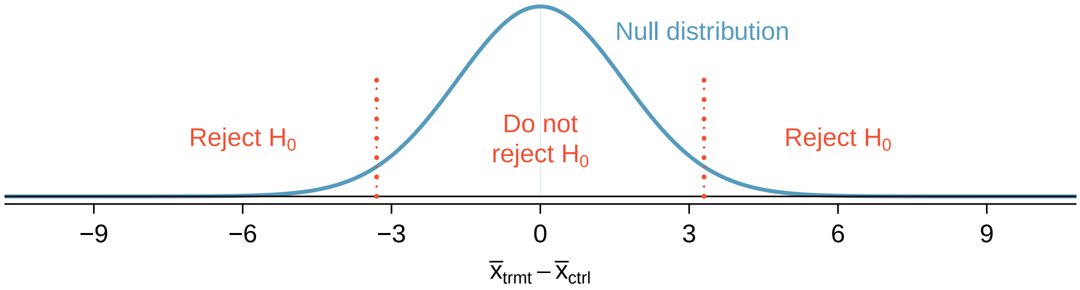
For what values of $\bar{x}_{trmt} - \bar{x}_{ctrl}$ would we reject the null hypothesis?

For $\alpha = 0.05$, we would reject H_0 if the difference is in the lower 2.5% or upper 2.5% tail:

Lower 2.5%: For the normal model, this is 1.96 standard errors below 0, so any difference smaller than $-1.96 \times 1.70 = -3.332$ mmHg.

Upper 2.5%: For the normal model, this is 1.96 standard errors above 0, so any difference larger than $1.96 \times 1.70 = 3.332$ mmHg.

The boundaries of these **rejection regions** are shown below:



Next, we'll perform some hypothetical calculations to determine the probability we reject the null hypothesis, if the alternative hypothesis were actually true.

7.4.2 Computing the power for a 2-sample test

When planning a study, we want to know how likely we are to detect an effect we care about. In other words, if there is a real effect, and that effect is large enough that it has practical value, then what's the probability that we detect that effect? This probability is called the **power**, and we can compute it for different sample sizes or for different *effect sizes*.

We first determine what is a practically significant result. Suppose that the company researchers care about finding any effect on blood pressure that is 3 mmHg or larger vs the standard medication. Here, 3 mmHg is the minimum **effect size** of interest, and we want to know how likely we are to detect this size of an effect in the study.

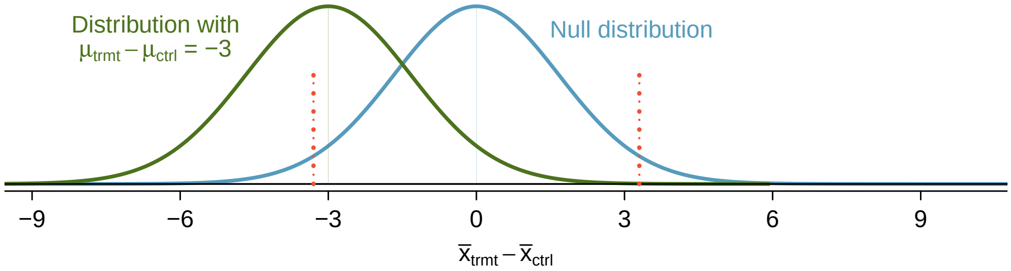
EXAMPLE 7.35

Suppose we decided to move forward with 100 patients per treatment group and the new drug reduces blood pressure by an additional 3 mmHg relative to the standard medication. What is the probability that we detect a drop?

Before we even do any calculations, notice that if $\bar{x}_{trmt} - \bar{x}_{ctrl} = -3$ mmHg, there wouldn't even be sufficient evidence to reject H_0. That's not a good sign.

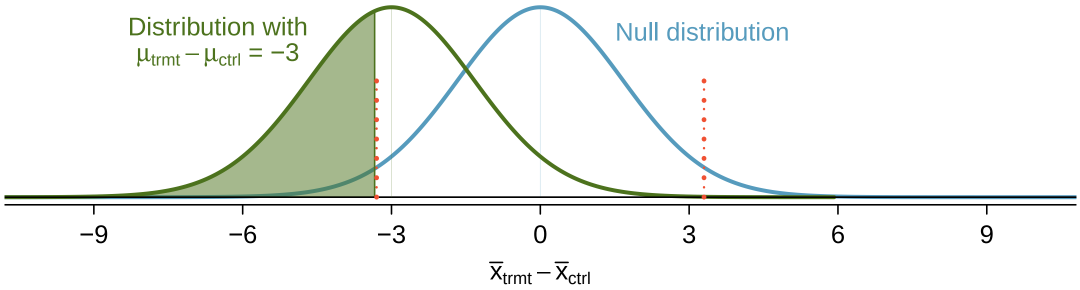
To calculate the probability that we will reject H_0, we need to determine a few things:

- The sampling distribution for $\bar{x}_{trmt} - \bar{x}_{ctrl}$ when the true difference is -3 mmHg. This is the same as the null distribution, except it is shifted to the left by 3:



- The rejection regions, which are outside of the dotted lines above.
- The fraction of the distribution that falls in the rejection region.

In short, we need to calculate the probability that $x < -3.332$ for a normal distribution with mean -3 and standard deviation 1.7. To do so, we first shade the area we want to calculate:



We'll use a normal approximation, which is good approximation when the degrees of freedom is about 30 or more. We'll start by calculating the Z-score and find the tail area using either statistical software or the probability table:

$$Z = \frac{-3.332 - (-3)}{1.7} = -0.20 \qquad \rightarrow \qquad 0.42$$

The power for the test is about 42% when $\mu_{trmt} - \mu_{ctrl} = -3$ and each group has a sample size of 100.

In Example 7.35, we ignored the upper rejection region in the calculation, which was in the opposite direction of the hypothetical truth, i.e. -3. The reasoning? There wouldn't be any value in rejecting the null hypothesis and concluding there was an increase when in fact there was a decrease.

We've also used a normal distribution instead of the t-distribution. This is a convenience, and if the sample size is too small, we'd need to revert back to using the t-distribution. We'll discuss this a bit further at the end of this section.

7.4.3 Determining a proper sample size

In the last example, we found that if we have a sample size of 100 in each group, we can only detect an effect size of 3 mmHg with a probability of about 0.42. Suppose the researchers moved forward and only used 100 patients per group, and the data did not support the alternative hypothesis, i.e. the researchers did not reject H_0. This is a very bad situation to be in for a few reasons:

- In the back of the researchers' minds, they'd all be wondering, *maybe there is a real and meaningful difference, but we weren't able to detect it with such a small sample.*
- The company probably invested hundreds of millions of dollars in developing the new drug, so now they are left with great uncertainty about its potential since the experiment didn't have a great shot at detecting effects that could still be important.
- Patients were subjected to the drug, and we can't even say with much certainty that the drug doesn't help (or harm) patients.
- Another clinical trial may need to be run to get a more conclusive answer as to whether the drug does hold any practical value, and conducting a second clinical trial may take years and many millions of dollars.

We want to avoid this situation, so we need to determine an appropriate sample size to ensure we can be pretty confident that we'll detect any effects that are practically important. As mentioned earlier, a change of 3 mmHg was deemed to be the minimum difference that was practically important. As a first step, we could calculate power for several different sample sizes. For instance, let's try 500 patients per group.

GUIDED PRACTICE 7.36

Calculate the power to detect a change of -3 mmHg when using a sample size of 500 per group.[27]

(a) Determine the standard error (recall that the standard deviation for patients was expected to be about 12 mmHg).

(b) Identify the null distribution and rejection regions.

(c) Identify the alternative distribution when $\mu_{trmt} - \mu_{ctrl} = -3$.

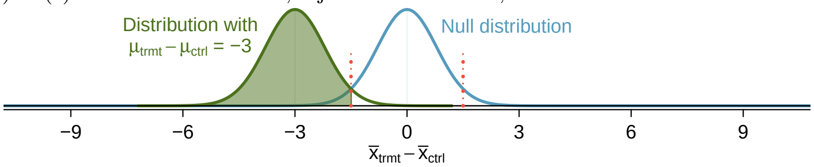
(d) Compute the probability we reject the null hypothesis.

The researchers decided 3 mmHg was the minimum difference that was practically important, and with a sample size of 500, we can be very certain (97.7% or better) that we will detect any such difference. We now have moved to another extreme where we are exposing an unnecessary number of patients to the new drug in the clinical trial. Not only is this ethically questionable, but it would also cost a lot more money than is necessary to be quite sure we'd detect any important effects.

The most common practice is to identify the sample size where the power is around 80%, and sometimes 90%. Other values may be reasonable for a specific context, but 80% and 90% are most commonly targeted as a good balance between high power and not exposing too many patients to a new treatment (or wasting too much money).

We could compute the power of the test at several other possible sample sizes until we find one that's close to 80%, but there's a better way. We should solve the problem backwards.

[27] (a) The standard error is given as $SE = \sqrt{\frac{12^2}{500} + \frac{12^2}{500}} = 0.76$.
(b) & (c) The null distribution, rejection boundaries, and alternative distribution are shown below:



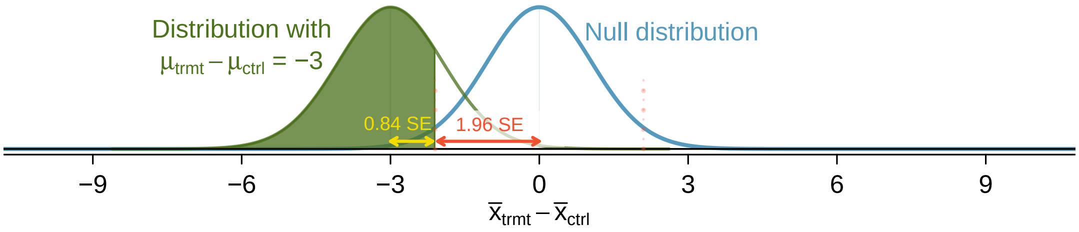
The rejection regions are the areas on the outside of the two dotted lines and are at ±0.76 × 1.96 = ±1.49.
(d) The area of the alternative distribution where $\mu_{trmt} - \mu_{ctrl} = -3$ has been shaded. We compute the Z-score and find the tail area: $Z = \frac{-1.49-(-3)}{0.76} = 1.99 \rightarrow 0.977$. With 500 patients per group, we would be about 97.7% sure (or more) that we'd detect any effects that are at least 3 mmHg in size.

EXAMPLE 7.37

What sample size will lead to a power of 80%? Use $\alpha = 0.05$.

We'll assume we have a large enough sample that the normal distribution is a good approximation for the test statistic, since the normal distribution and the t-distribution look almost identical when the degrees of freedom are moderately large (e.g. $df \geq 30$). If that doesn't turn out to be true, then we'd need to make a correction.

We start by identifying the Z-score that would give us a lower tail of 80%. For a moderately large sample size per group, the Z-score for a lower tail of 80% would be about $Z = 0.84$.



Additionally, the rejection region extends $1.96 \times SE$ from the center of the null distribution for $\alpha = 0.05$. This allows us to calculate the target distance between the center of the null and alternative distributions in terms of the standard error:

$$0.84 \times SE + 1.96 \times SE = 2.8 \times SE$$

In our example, we want the distance between the null and alternative distributions' centers to equal the minimum effect size of interest, 3 mmHg, which allows us to set up an equation between this difference and the standard error:

$$\begin{aligned} 3 &= 2.8 \times SE \\ 3 &= 2.8 \times \sqrt{\frac{12^2}{n} + \frac{12^2}{n}} \\ n &= \frac{2.8^2}{3^2} \times \left(12^2 + 12^2\right) = 250.88 \end{aligned}$$

We should target 251 patients per group in order to achieve 80% power at the 0.05 significance level for this context.

The standard error difference of $2.8 \times SE$ is specific to a context where the targeted power is 80% and the significance level is $\alpha = 0.05$. If the targeted power is 90% or if we use a different significance level, then we'll use something a little different than $2.8 \times SE$.

Had the suggested sample size been relatively small – roughly 30 or smaller – it would have been a good idea to rework the calculations using the degrees of fredom for the smaller sample size under that initial sample size. That is, we would have revised the 0.84 and 1.96 values based on degrees of freedom implied by the initial sample size. The revised sample size target would generally have then been a little larger.

GUIDED PRACTICE 7.38

Suppose the targeted power was 90% and we were using $\alpha = 0.01$. How many standard errors should separate the centers of the null and alternative distribution, where the alternative distribution is centered at the minimum effect size of interest?[28]

GUIDED PRACTICE 7.39

What are some considerations that are important in determining what the power should be for an experiment?[29]

Figure 7.18 shows the power for sample sizes from 20 patients to 5,000 patients when $\alpha = 0.05$ and the true difference is -3. This curve was constructed by writing a program to compute the power for many different sample sizes.

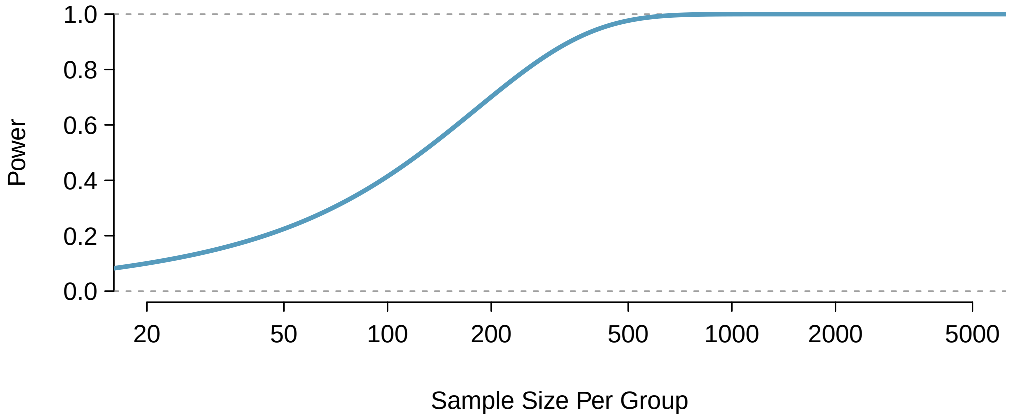


Figure 7.18: The curve shows the power for different sample sizes in the context of the blood pressure example when the true difference is -3. Having more than about 250 to 350 observations doesn't provide much additional value in detecting an effect when $\alpha = 0.05$.

Power calculations for expensive or risky experiments are critical. However, what about experiments that are inexpensive and where the ethical considerations are minimal? For example, if we are doing final testing on a new feature on a popular website, how would our sample size considerations change? As before, we'd want to make sure the sample is big enough. However, suppose the feature has undergone some testing and is known to perform well (e.g. the website's users seem to enjoy the feature). Then it may be reasonable to run a larger experiment if there's value from having a more precise estimate of the feature's effect, such as helping guide the development of the next useful feature.

[28] First, find the Z-score such that 90% of the distribution is below it: $Z = 1.28$. Next, find the cutoffs for the rejection regions: ± 2.58. Then the difference in centers should be about $1.28 \times SE + 2.58 \times SE = 3.86 \times SE$.

[29] Answers will vary, but here are a few important considerations:

- Whether there is any risk to patients in the study.
- The cost of enrolling more patients.
- The potential downside of not detecting an effect of interest.

Exercises

7.33 Increasing corn yield. A large farm wants to try out a new type of fertilizer to evaluate whether it will improve the farm's corn production. The land is broken into plots that produce an average of 1,215 pounds of corn with a standard deviation of 94 pounds per plot. The owner is interested in detecting any average difference of at least 40 pounds per plot. How many plots of land would be needed for the experiment if the desired power level is 90%? Use $\alpha = 0.05$. Assume each plot of land gets treated with either the current fertilizer or the new fertilizer.

7.34 Email outreach efforts. A medical research group is recruiting people to complete short surveys about their medical history. For example, one survey asks for information on a person's family history in regards to cancer. Another survey asks about what topics were discussed during the person's last visit to a hospital. So far, as people sign up, they complete an average of just 4 surveys, and the standard deviation of the number of surveys is about 2.2. The research group wants to try a new interface that they think will encourage new enrollees to complete more surveys, where they will randomize each enrollee to either get the new interface or the current interface. How many new enrollees do they need for each interface to detect an effect size of 0.5 surveys per enrollee, if the desired power level is 80%? Use $\alpha = 0.05$.

7.5 Comparing many means with ANOVA

Sometimes we want to compare means across many groups. We might initially think to do pairwise comparisons. For example, if there were three groups, we might be tempted to compare the first mean with the second, then with the third, and then finally compare the second and third means for a total of three comparisons. However, this strategy can be treacherous. If we have many groups and do many comparisons, it is likely that we will eventually find a difference just by chance, even if there is no difference in the populations. Instead, we should apply a holistic test to check whether there is evidence that at least one pair groups are in fact different, and this is where *ANOVA* saves the day.

7.5.1 Core ideas of ANOVA

In this section, we will learn a new method called **analysis of variance (ANOVA)** and a new test statistic called F. ANOVA uses a single hypothesis test to check whether the means across many groups are equal:

H_0: The mean outcome is the same across all groups. In statistical notation, $\mu_1 = \mu_2 = \cdots = \mu_k$ where μ_i represents the mean of the outcome for observations in category i.

H_A: At least one mean is different.

Generally we must check three conditions on the data before performing ANOVA:

- the observations are independent within and across groups,
- the data within each group are nearly normal, and
- the variability across the groups is about equal.

When these three conditions are met, we may perform an ANOVA to determine whether the data provide strong evidence against the null hypothesis that all the μ_i are equal.

EXAMPLE 7.40

College departments commonly run multiple lectures of the same introductory course each semester because of high demand. Consider a statistics department that runs three lectures of an introductory statistics course. We might like to determine whether there are statistically significant differences in first exam scores in these three classes (A, B, and C). Describe appropriate hypotheses to determine whether there are any differences between the three classes.

E

The hypotheses may be written in the following form:

H_0: The average score is identical in all lectures. Any observed difference is due to chance. Notationally, we write $\mu_A = \mu_B = \mu_C$.

H_A: The average score varies by class. We would reject the null hypothesis in favor of the alternative hypothesis if there were larger differences among the class averages than what we might expect from chance alone.

Strong evidence favoring the alternative hypothesis in ANOVA is described by unusually large differences among the group means. We will soon learn that assessing the variability of the group means relative to the variability among individual observations within each group is key to ANOVA's success.

EXAMPLE 7.41

Examine Figure 7.19. Compare groups I, II, and III. Can you visually determine if the differences in the group centers is due to chance or not? Now compare groups IV, V, and VI. Do these differences appear to be due to chance?

E

Any real difference in the means of groups I, II, and III is difficult to discern, because the data within each group are very volatile relative to any differences in the average outcome. On the other hand, it appears there are differences in the centers of groups IV, V, and VI. For instance, group V appears to have a higher mean than that of the other two groups. Investigating groups IV, V, and VI, we see the differences in the groups' centers are noticeable because those differences are large *relative to the variability in the individual observations within each group*.

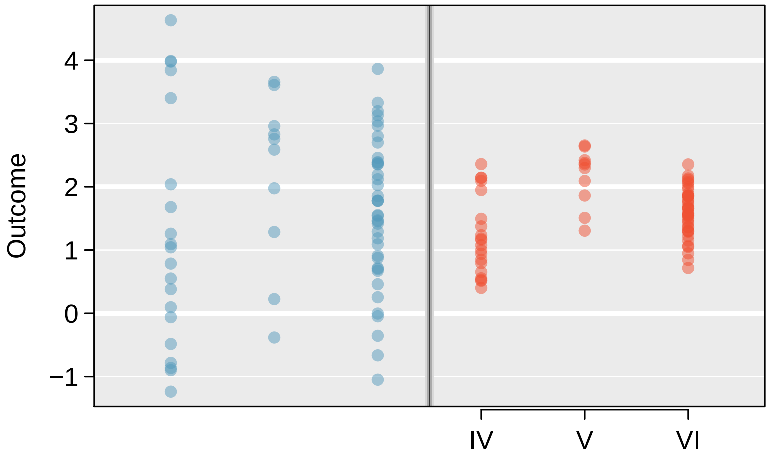


Figure 7.19: Side-by-side dot plot for the outcomes for six groups.

7.5.2 Is batting performance related to player position in MLB?

We would like to discern whether there are real differences between the batting performance of baseball players according to their position: outfielder (OF), infielder (IF), and catcher (C). We will use a data set called bat18, which includes batting records of 429 Major League Baseball (MLB) players from the 2018 season who had at least 100 at bats. Six of the 429 cases represented in bat18 are shown in Figure 7.20, and descriptions for each variable are provided in Figure 7.21. The measure we will use for the player batting performance (the outcome variable) is on-base percentage (OBP). The on-base percentage roughly represents the fraction of the time a player successfully gets on base or hits a home run.

	name	team	position	AB	H	HR	RBI	AVG	OBP
1	Abreu, J	CWS	IF	499	132	22	78	0.265	0.325
2	Acuna Jr., R	ATL	OF	433	127	26	64	0.293	0.366
3	Adames, W	TB	IF	288	80	10	34	0.278	0.348
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		
427	Zimmerman, R	WSH	IF	288	76	13	51	0.264	0.337
428	Zobrist, B	CHC	IF	455	139	9	58	0.305	0.378
429	Zunino, M	SEA	C	373	75	20	44	0.201	0.259

Figure 7.20: Six cases from the bat18 data matrix.

variable	description
`name`	Player name
`team`	The abbreviated name of the player's team
`position`	The player's primary field position (`OF`, `IF`, `C`)
`AB`	Number of opportunities at bat
`H`	Number of hits
`HR`	Number of home runs
`RBI`	Number of runs batted in
`AVG`	Batting average, which is equal to `H`/`AB`
`OBP`	On-base percentage, which is roughly equal to the fraction of times a player gets on base or hits a home run

Figure 7.21: Variables and their descriptions for the `bat18` data set.

GUIDED PRACTICE 7.42

The null hypothesis under consideration is the following: $\mu_{\texttt{OF}} = \mu_{\texttt{IF}} = \mu_{\texttt{C}}$. Write the null and corresponding alternative hypotheses in plain language.[30]

EXAMPLE 7.43

The player positions have been divided into three groups: outfield (`OF`), infield (`IF`), and catcher (`C`). What would be an appropriate point estimate of the on-base percentage by outfielders, $\mu_{\texttt{OF}}$?

A good estimate of the on-base percentage by outfielders would be the sample average of `OBP` for just those players whose position is outfield: $\bar{x}_{OF} = 0.320$.

Figure 7.22 provides summary statistics for each group. A side-by-side box plot for the on-base percentage is shown in Figure 7.23. Notice that the variability appears to be approximately constant across groups; nearly constant variance across groups is an important assumption that must be satisfied before we consider the ANOVA approach.

	`OF`	`IF`	`C`
Sample size (n_i)	160	205	64
Sample mean ($\bar{x}_i$)	0.320	0.318	0.302
Sample SD (s_i)	0.043	0.038	0.038

Figure 7.22: Summary statistics of on-base percentage, split by player position.

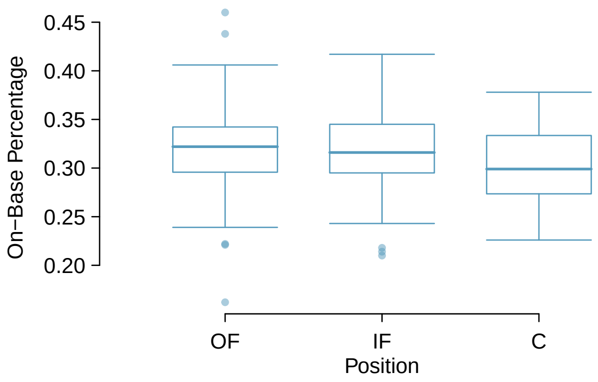


Figure 7.23: Side-by-side box plot of the on-base percentage for 429 players across three groups. With over a hundred players in both the infield and outfield groups, the apparent outliers are not a concern.

[30] H_0: The average on-base percentage is equal across the three positions. H_A: The average on-base percentage varies across some (or all) groups.

EXAMPLE 7.44

The largest difference between the sample means is between the catcher and the outfielder positions. Consider again the original hypotheses:

H_0: $\mu_{\texttt{OF}} = \mu_{\texttt{IF}} = \mu_{\texttt{C}}$

H_A: The average on-base percentage (μ_i) varies across some (or all) groups.

Why might it be inappropriate to run the test by simply estimating whether the difference of $\mu_{\texttt{C}}$ and $\mu_{\texttt{OF}}$ is statistically significant at a 0.05 significance level?

E

The primary issue here is that we are inspecting the data before picking the groups that will be compared. It is inappropriate to examine all data by eye (informal testing) and only afterwards decide which parts to formally test. This is called **data snooping** or **data fishing**. Naturally, we would pick the groups with the large differences for the formal test, and this would leading to an inflation in the Type 1 Error rate. To understand this better, let's consider a slightly different problem.

Suppose we are to measure the aptitude for students in 20 classes in a large elementary school at the beginning of the year. In this school, all students are randomly assigned to classrooms, so any differences we observe between the classes at the start of the year are completely due to chance. However, with so many groups, we will probably observe a few groups that look rather different from each other. If we select only these classes that look so different and then perform a formal test, we will probably make the wrong conclusion that the assignment wasn't random. While we might only formally test differences for a few pairs of classes, we informally evaluated the other classes by eye before choosing the most extreme cases for a comparison.

For additional information on the ideas expressed in Example 7.44, we recommend reading about the **prosecutor's fallacy**.[31]

In the next section we will learn how to use the F statistic and ANOVA to test whether observed differences in sample means could have happened just by chance even if there was no difference in the respective population means.

[31] See, for example, statmodeling.stat.columbia.edu/2007/05/18/the_prosecutors.

7.5.3 Analysis of variance (ANOVA) and the F-test

The method of analysis of variance in this context focuses on answering one question: is the variability in the sample means so large that it seems unlikely to be from chance alone? This question is different from earlier testing procedures since we will *simultaneously* consider many groups, and evaluate whether their sample means differ more than we would expect from natural variation. We call this variability the **mean square between groups** (MSG), and it has an associated degrees of freedom, $df_G = k - 1$ when there are k groups. The MSG can be thought of as a scaled variance formula for means. If the null hypothesis is true, any variation in the sample means is due to chance and shouldn't be too large. Details of MSG calculations are provided in the footnote.[32] However, we typically use software for these computations.

The mean square between the groups is, on its own, quite useless in a hypothesis test. We need a benchmark value for how much variability should be expected among the sample means if the null hypothesis is true. To this end, we compute a pooled variance estimate, often abbreviated as the **mean square error** (MSE), which has an associated degrees of freedom value $df_E = n - k$. It is helpful to think of MSE as a measure of the variability within the groups. Details of the computations of the MSE and a link to an extra online section for ANOVA calculations are provided in the footnote[33] for interested readers.

When the null hypothesis is true, any differences among the sample means are only due to chance, and the MSG and MSE should be about equal. As a test statistic for ANOVA, we examine the fraction of MSG and MSE:

$$F = \frac{MSG}{MSE}$$

The MSG represents a measure of the between-group variability, and MSE measures the variability within each of the groups.

GUIDED PRACTICE 7.45

For the baseball data, $MSG = 0.00803$ and $MSE = 0.00158$. Identify the degrees of freedom associated with MSG and MSE and verify the F statistic is approximately 5.077.[34]

We can use the F statistic to evaluate the hypotheses in what is called an ***F*-test**. A p-value can be computed from the F statistic using an F distribution, which has two associated parameters: df_1 and df_2. For the F statistic in ANOVA, $df_1 = df_G$ and $df_2 = df_E$. An F distribution with 2 and 426 degrees of freedom, corresponding to the F statistic for the baseball hypothesis test, is shown in Figure 7.24.

[32] Let $\bar{x}$ represent the mean of outcomes across all groups. Then the mean square between groups is computed as

$$MSG = \frac{1}{df_G} SSG = \frac{1}{k-1} \sum_{i=1}^{k} n_i \left(\bar{x}_i - \bar{x}\right)^2$$

where SSG is called the **sum of squares between groups** and n_i is the sample size of group i.

[33] Let $\bar{x}$ represent the mean of outcomes across all groups. Then the **sum of squares total** (SST) is computed as

$$SST = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$$

where the sum is over all observations in the data set. Then we compute the **sum of squared errors** (SSE) in one of two equivalent ways:

$$\begin{aligned} SSE &= SST - SSG \\ &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2 \end{aligned}$$

where s_i^2 is the sample variance (square of the standard deviation) of the residuals in group i. Then the MSE is the standardized form of SSE: $MSE = \frac{1}{df_E} SSE$.

For additional details on ANOVA calculations, see www.openintro.org/d?file=stat_extra_anova_calculations

[34] There are $k = 3$ groups, so $df_G = k - 1 = 2$. There are $n = n_1 + n_2 + n_3 = 429$ total observations, so $df_E = n - k = 426$. Then the F statistic is computed as the ratio of MSG and MSE: $F = \frac{MSG}{MSE} = \frac{0.00803}{0.00158} = 5.082 \approx 5.077$. ($F = 5.077$ was computed by using values for MSG and MSE that were not rounded.)

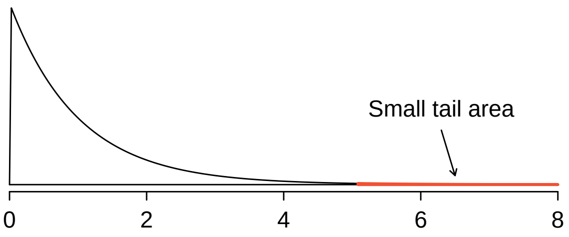


Figure 7.24: An F distribution with $df_1 = 2$ and $df_2 = 426$.

The larger the observed variability in the sample means (MSG) relative to the within-group observations (MSE), the larger F will be and the stronger the evidence against the null hypothesis. Because larger values of F represent stronger evidence against the null hypothesis, we use the upper tail of the distribution to compute a p-value.

THE F STATISTIC AND THE F-TEST

Analysis of variance (ANOVA) is used to test whether the mean outcome differs across 2 or more groups. ANOVA uses a test statistic F, which represents a standardized ratio of variability in the sample means relative to the variability within the groups. If H_0 is true and the model conditions are satisfied, the statistic F follows an F distribution with parameters $df_1 = k - 1$ and $df_2 = n - k$. The upper tail of the F distribution is used to represent the p-value.

EXAMPLE 7.46

The p-value corresponding to the shaded area in Figure 7.24 is equal to about 0.0066. Does this provide strong evidence against the null hypothesis?

The p-value is smaller than 0.05, indicating the evidence is strong enough to reject the null hypothesis at a significance level of 0.05. That is, the data provide strong evidence that the average on-base percentage varies by player's primary field position.

7.5.4 Reading an ANOVA table from software

The calculations required to perform an ANOVA by hand are tedious and prone to human error. For these reasons, it is common to use statistical software to calculate the F statistic and p-value.

An ANOVA can be summarized in a table very similar to that of a regression summary, which we will see in Chapters 8 and 9. Figure 7.25 shows an ANOVA summary to test whether the mean of on-base percentage varies by player positions in the MLB. Many of these values should look familiar; in particular, the F-test statistic and p-value can be retrieved from the last two columns.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
position	2	0.0161	0.0080	5.0766	0.0066
Residuals	426	0.6740	0.0016		

$s_{pooled} = 0.040$ on $df = 423$

Figure 7.25: ANOVA summary for testing whether the average on-base percentage differs across player positions.

7.5.5 Graphical diagnostics for an ANOVA analysis

There are three conditions we must check for an ANOVA analysis: all observations must be independent, the data in each group must be nearly normal, and the variance within each group must be approximately equal.

Independence. If the data are a simple random sample, this condition is satisfied. For processes and experiments, carefully consider whether the data may be independent (e.g. no pairing). For example, in the MLB data, the data were not sampled. However, there are not obvious reasons why independence would not hold for most or all observations.

Approximately normal. As with one- and two-sample testing for means, the normality assumption is especially important when the sample size is quite small when it is ironically difficult to check for non-normality. A histogram of the observations from each group is shown in Figure 7.26. Since each of the groups we're considering have relatively large sample sizes, what we're looking for are major outliers. None are apparent, so this conditions is reasonably met.

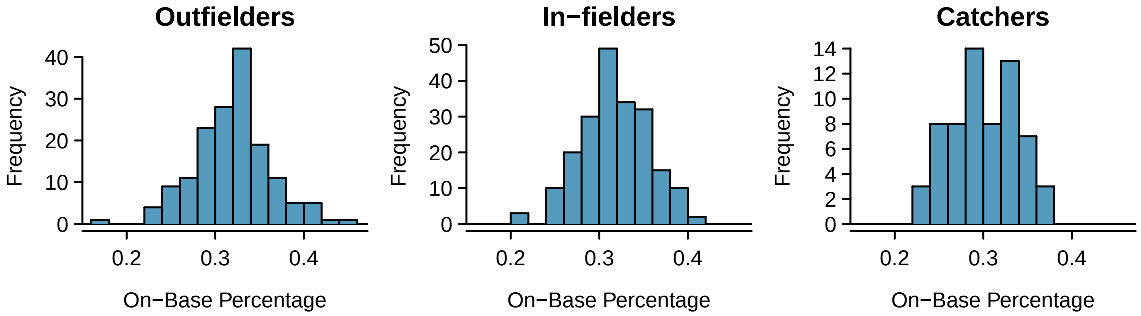


Figure 7.26: Histograms of OBP for each field position.

Constant variance. The last assumption is that the variance in the groups is about equal from one group to the next. This assumption can be checked by examining a side-by-side box plot of the outcomes across the groups, as in Figure 7.23 on page 287. In this case, the variability is similar in the three groups but not identical. We see in Table 7.22 on page 287 that the standard deviation doesn't vary much from one group to the next.

DIAGNOSTICS FOR AN ANOVA ANALYSIS

Independence is always important to an ANOVA analysis. The normality condition is very important when the sample sizes for each group are relatively small. The constant variance condition is especially important when the sample sizes differ between groups.

7.5.6 Multiple comparisons and controlling Type 1 Error rate

When we reject the null hypothesis in an ANOVA analysis, we might wonder, which of these groups have different means? To answer this question, we compare the means of each possible pair of groups. For instance, if there are three groups and there is strong evidence that there are some differences in the group means, there are three comparisons to make: group 1 to group 2, group 1 to group 3, and group 2 to group 3. These comparisons can be accomplished using a two-sample t-test, but we use a modified significance level and a pooled estimate of the standard deviation across groups. Usually this pooled standard deviation can be found in the ANOVA table, e.g. along the bottom of Figure 7.25.

EXAMPLE 7.47

Example 7.40 on page 285 discussed three statistics lectures, all taught during the same semester. Figure 7.27 shows summary statistics for these three courses, and a side-by-side box plot of the data is shown in Figure 7.28. We would like to conduct an ANOVA for these data. Do you see any deviations from the three conditions for ANOVA?

E

In this case (like many others) it is difficult to check independence in a rigorous way. Instead, the best we can do is use common sense to consider reasons the assumption of independence may not hold. For instance, the independence assumption may not be reasonable if there is a star teaching assistant that only half of the students may access; such a scenario would divide a class into two subgroups. No such situations were evident for these particular data, and we believe that independence is acceptable.

The distributions in the side-by-side box plot appear to be roughly symmetric and show no noticeable outliers.

The box plots show approximately equal variability, which can be verified in Figure 7.27, supporting the constant variance assumption.

Class i	A	B	C
n_i	58	55	51
$\bar{x}_i$	75.1	72.0	78.9
s_i	13.9	13.8	13.1

Figure 7.27: Summary statistics for the first midterm scores in three different lectures of the same course.

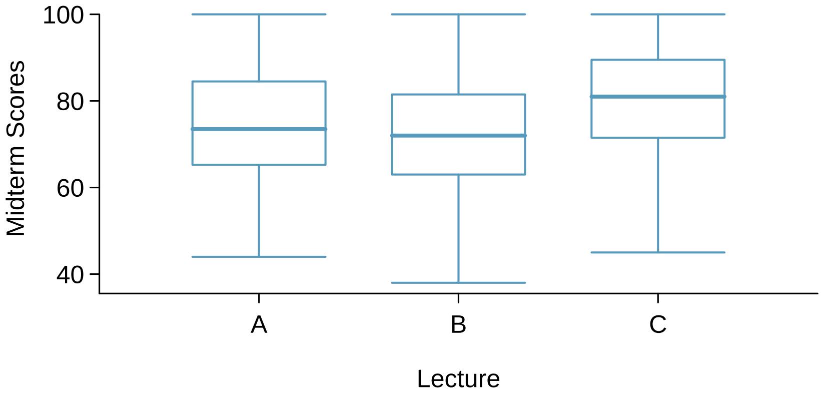


Figure 7.28: Side-by-side box plot for the first midterm scores in three different lectures of the same course.

GUIDED PRACTICE 7.48

ANOVA was conducted for the midterm data, and summary results are shown in Figure 7.29. What should we conclude?[35]

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
lecture	2	1290.11	645.06	3.48	0.0330
Residuals	161	29810.13	185.16		

$s_{pooled} = 13.61$ on $df = 161$

Figure 7.29: ANOVA summary table for the midterm data.

There is strong evidence that the different means in each of the three classes is not simply due to chance. We might wonder, which of the classes are actually different? As discussed in earlier chapters, a two-sample t-test could be used to test for differences in each possible pair of groups. However, one pitfall was discussed in Example 7.44 on page 288: when we run so many tests, the Type 1 Error rate increases. This issue is resolved by using a modified significance level.

MULTIPLE COMPARISONS AND THE BONFERRONI CORRECTION FOR α

The scenario of testing many pairs of groups is called **multiple comparisons**. The **Bonferroni correction** suggests that a more stringent significance level is more appropriate for these tests:

$$\alpha^\star = \alpha / K$$

where K is the number of comparisons being considered (formally or informally). If there are k groups, then usually all possible pairs are compared and $K = \frac{k(k-1)}{2}$.

EXAMPLE 7.49

In Guided Practice 7.48, you found strong evidence of differences in the average midterm grades between the three lectures. Complete the three possible pairwise comparisons using the Bonferroni correction and report any differences.

We use a modified significance level of $\alpha^\star = 0.05/3 = 0.0167$. Additionally, we use the pooled estimate of the standard deviation: $s_{pooled} = 13.61$ on $df = 161$, which is provided in the ANOVA summary table.

Lecture A versus Lecture B: The estimated difference and standard error are, respectively,

$$\bar{x}_A - \bar{x}_B = 75.1 - 72 = 3.1 \qquad\qquad SE = \sqrt{\frac{13.61^2}{58} + \frac{13.61^2}{55}} = 2.56$$

(See Section 7.3.4 on page 273 for additional details.) This results in a T-score of 1.21 on $df = 161$ (we use the df associated with s_{pooled}). Statistical software was used to precisely identify the two-sided p-value since the modified significance level of 0.0167 is not found in the t-table. The p-value (0.228) is larger than $\alpha^* = 0.0167$, so there is not strong evidence of a difference in the means of lectures A and B.

Lecture A versus Lecture C: The estimated difference and standard error are 3.8 and 2.61, respectively. This results in a T score of 1.46 on $df = 161$ and a two-sided p-value of 0.1462. This p-value is larger than α^*, so there is not strong evidence of a difference in the means of lectures A and C.

Lecture B versus Lecture C: The estimated difference and standard error are 6.9 and 2.65, respectively. This results in a T score of 2.60 on $df = 161$ and a two-sided p-value of 0.0102. This p-value is smaller than α^*. Here we find strong evidence of a difference in the means of lectures B and C.

[35] The p-value of the test is 0.0330, less than the default significance level of 0.05. Therefore, we reject the null hypothesis and conclude that the difference in the average midterm scores are not due to chance.

We might summarize the findings of the analysis from Example 7.49 using the following notation:

$$\mu_A \overset{?}{=} \mu_B \qquad \mu_A \overset{?}{=} \mu_C \qquad \mu_B \neq \mu_C$$

The midterm mean in lecture A is not statistically distinguishable from those of lectures B or C. However, there is strong evidence that lectures B and C are different. In the first two pairwise comparisons, we did not have sufficient evidence to reject the null hypothesis. Recall that failing to reject H_0 does not imply H_0 is true.

REJECT H_0 WITH ANOVA BUT FIND NO DIFFERENCES IN GROUP MEANS

It is possible to reject the null hypothesis using ANOVA and then to not subsequently identify differences in the pairwise comparisons. However, *this does not invalidate the ANOVA conclusion.* It only means we have not been able to successfully identify which specific groups differ in their means.

The ANOVA procedure examines the big picture: it considers all groups simultaneously to decipher whether there is evidence that some difference exists. Even if the test indicates that there is strong evidence of differences in group means, identifying with high confidence a specific difference as statistically significant is more difficult.

Consider the following analogy: we observe a Wall Street firm that makes large quantities of money based on predicting mergers. Mergers are generally difficult to predict, and if the prediction success rate is extremely high, that may be considered sufficiently strong evidence to warrant investigation by the Securities and Exchange Commission (SEC). While the SEC may be quite certain that there is insider trading taking place at the firm, the evidence against any single trader may not be very strong. It is only when the SEC considers all the data that they identify the pattern. This is effectively the strategy of ANOVA: stand back and consider all the groups simultaneously.

Exercises

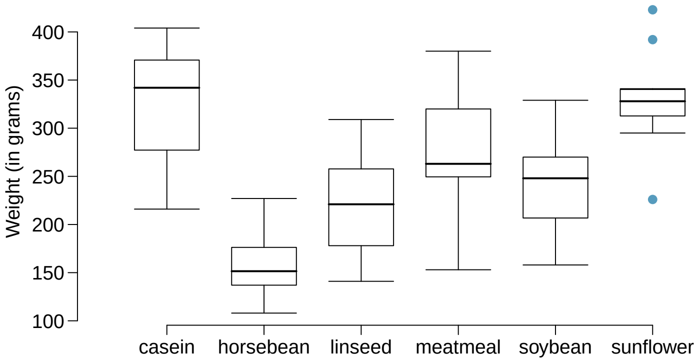
7.35 Fill in the blank. When doing an ANOVA, you observe large differences in means between groups. Within the ANOVA framework, this would most likely be interpreted as evidence strongly favoring the ________________ hypothesis.

7.36 Which test? We would like to test if students who are in the social sciences, natural sciences, arts and humanities, and other fields spend the same amount of time studying for this course. What type of test should we use? Explain your reasoning.

7.37 Chicken diet and weight, Part III. In Exercises 7.27 and 7.29 we compared the effects of two types of feed at a time. A better analysis would first consider all feed types at once: casein, horsebean, linseed, meat meal, soybean, and sunflower. The ANOVA output below can be used to test for differences between the average weights of chicks on different diets.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
feed	5	231,129.16	46,225.83	15.36	0.0000
Residuals	65	195,556.02	3,008.55		

Conduct a hypothesis test to determine if these data provide convincing evidence that the average weight of chicks varies across some (or all) groups. Make sure to check relevant conditions. Figures and summary statistics are shown below.



	Mean	SD	n
casein	323.58	64.43	12
horsebean	160.20	38.63	10
linseed	218.75	52.24	12
meatmeal	276.91	64.90	11
soybean	246.43	54.13	14
sunflower	328.92	48.84	12

7.38 Teaching descriptive statistics. A study compared five different methods for teaching descriptive statistics. The five methods were traditional lecture and discussion, programmed textbook instruction, programmed text with lectures, computer instruction, and computer instruction with lectures. 45 students were randomly assigned, 9 to each method. After completing the course, students took a 1-hour exam.

(a) What are the hypotheses for evaluating if the average test scores are different for the different teaching methods?

(b) What are the degrees of freedom associated with the F-test for evaluating these hypotheses?

(c) Suppose the p-value for this test is 0.0168. What is the conclusion?

7.39 Coffee, depression, and physical activity. Caffeine is the world's most widely used stimulant, with approximately 80% consumed in the form of coffee. Participants in a study investigating the relationship between coffee consumption and exercise were asked to report the number of hours they spent per week on moderate (e.g., brisk walking) and vigorous (e.g., strenuous sports and jogging) exercise. Based on these data the researchers estimated the total hours of metabolic equivalent tasks (MET) per week, a value always greater than 0. The table below gives summary statistics of MET for women in this study based on the amount of coffee consumed.[36]

	Caffeinated coffee consumption					
	$\leq$ 1 cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	$\geq$ 4 cups/day	Total
Mean	18.7	19.6	19.3	18.9	17.5	
SD	21.1	25.5	22.5	22.0	22.0	
n	12,215	6,617	17,234	12,290	2,383	50,739

(a) Write the hypotheses for evaluating if the average physical activity level varies among the different levels of coffee consumption.

(b) Check conditions and describe any assumptions you must make to proceed with the test.

(c) Below is part of the output associated with this test. Fill in the empty cells.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
coffee	□	□	□	□	0.0003
Residuals	□	25,564,819	□		
Total	□	25,575,327			

(d) What is the conclusion of the test?

7.40 Student performance across discussion sections. A professor who teaches a large introductory statistics class (197 students) with eight discussion sections would like to test if student performance differs by discussion section, where each discussion section has a different teaching assistant. The summary table below shows the average final exam score for each discussion section as well as the standard deviation of scores and the number of students in each section.

	Sec 1	Sec 2	Sec 3	Sec 4	Sec 5	Sec 6	Sec 7	Sec 8
n_i	33	19	10	29	33	10	32	31
$\bar{x}_i$	92.94	91.11	91.80	92.45	89.30	88.30	90.12	93.35
s_i	4.21	5.58	3.43	5.92	9.32	7.27	6.93	4.57

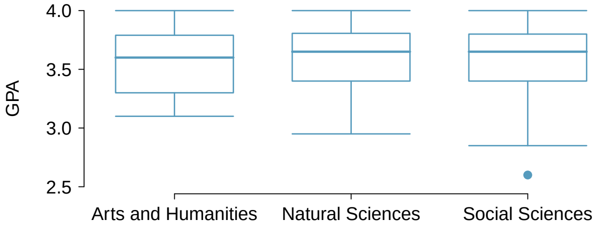
The ANOVA output below can be used to test for differences between the average scores from the different discussion sections.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
section	7	525.01	75.00	1.87	0.0767
Residuals	189	7584.11	40.13		

Conduct a hypothesis test to determine if these data provide convincing evidence that the average score varies across some (or all) groups. Check conditions and describe any assumptions you must make to proceed with the test.

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

7.41 GPA and major. Undergraduate students taking an introductory statistics course at Duke University conducted a survey about GPA and major. The side-by-side box plots show the distribution of GPA among three groups of majors. Also provided is the ANOVA output.



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
major	2	0.03	0.015	0.185	0.8313
Residuals	195	15.77	0.081		

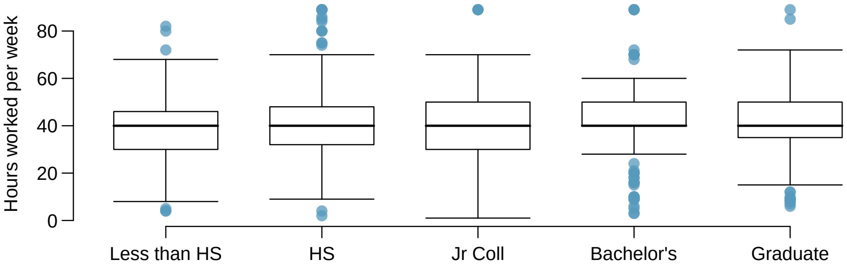
(a) Write the hypotheses for testing for a difference between average GPA across majors.

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

(c) How many students answered these questions on the survey, i.e. what is the sample size?

7.42 Work hours and education. The General Social Survey collects data on demographics, education, and work, among many other characteristics of US residents.[37] Using ANOVA, we can consider educational attainment levels for all 1,172 respondents at once. Below are the distributions of hours worked by educational attainment and relevant summary statistics that will be helpful in carrying out this analysis.

	Educational attainment					
	Less than HS	HS	Jr Coll	Bachelor's	Graduate	Total
Mean	38.67	39.6	41.39	42.55	40.85	40.45
SD	15.81	14.97	18.1	13.62	15.51	15.17
n	121	546	97	253	155	1,172



(a) Write hypotheses for evaluating whether the average number of hours worked varies across the five groups.

(b) Check conditions and describe any assumptions you must make to proceed with the test.

(c) Below is part of the output associated with this test. Fill in the empty cells.

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
degree			501.54		0.0682
Residuals		267,382			
Total					

(d) What is the conclusion of the test?

7.43 True / False: ANOVA, Part I. Determine if the following statements are true or false in ANOVA, and explain your reasoning for statements you identify as false.

(a) As the number of groups increases, the modified significance level for pairwise tests increases as well.

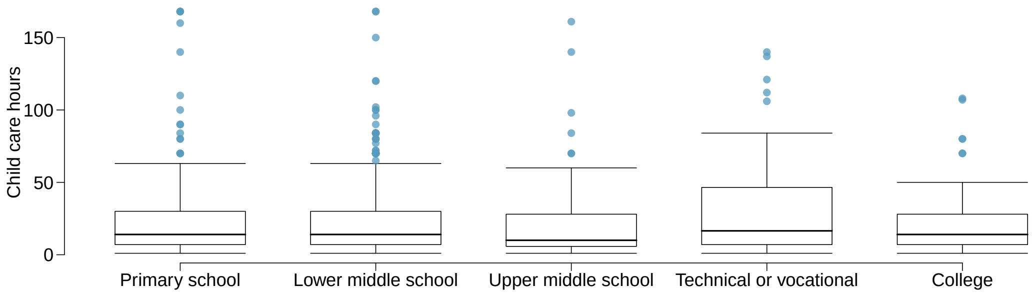
(b) As the total sample size increases, the degrees of freedom for the residuals increases as well.

(c) The constant variance condition can be somewhat relaxed when the sample sizes are relatively consistent across groups.

(d) The independence assumption can be relaxed when the total sample size is large.

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

7.44 Child care hours. The China Health and Nutrition Survey aims to examine the effects of the health, nutrition, and family planning policies and programs implemented by national and local governments.[38] It, for example, collects information on number of hours Chinese parents spend taking care of their children under age 6. The side-by-side box plots below show the distribution of this variable by educational attainment of the parent. Also provided below is the ANOVA output for comparing average hours across educational attainment categories.



	Df	Sum Sq	Mean Sq	F value	Pr(>F)
education	4	4142.09	1035.52	1.26	0.2846
Residuals	794	653047.83	822.48		

(a) Write the hypotheses for testing for a difference between the average number of hours spent on child care across educational attainment levels.

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

7.45 Prison isolation experiment, Part II. Exercise 7.31 introduced an experiment that was conducted with the goal of identifying a treatment that reduces subjects' psychopathic deviant T scores, where this score measures a person's need for control or his rebellion against control. In Exercise 7.31 you evaluated the success of each treatment individually. An alternative analysis involves comparing the success of treatments. The relevant ANOVA output is given below.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatment	2	639.48	319.74	3.33	0.0461
Residuals	39	3740.43	95.91		

$s_{pooled} = 9.793$ on $df = 39$

(a) What are the hypotheses?

(b) What is the conclusion of the test? Use a 5% significance level.

(c) If in part (b) you determined that the test is significant, conduct pairwise tests to determine which groups are different from each other. If you did not reject the null hypothesis in part (b), recheck your answer. Summary statistics for each group are provided below

	Tr 1	Tr 2	Tr 3
Mean	6.21	2.86	-3.21
SD	12.3	7.94	8.57
n	14	14	14

7.46 True / False: ANOVA, Part II. Determine if the following statements are true or false, and explain your reasoning for statements you identify as false.

If the null hypothesis that the means of four groups are all the same is rejected using ANOVA at a 5% significance level, then ...

(a) we can then conclude that all the means are different from one another.

(b) the standardized variability between groups is higher than the standardized variability within groups.

(c) the pairwise analysis will identify at least one pair of means that are significantly different.

(d) the appropriate α to be used in pairwise comparisons is 0.05 / 4 = 0.0125 since there are four groups.

[38] UNC Carolina Population Center, China Health and Nutrition Survey, 2006.

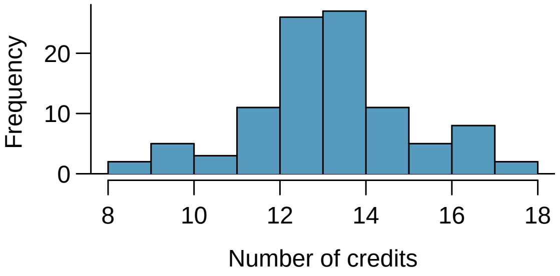
Chapter exercises

7.47 Gaming and distracted eating, Part I. A group of researchers are interested in the possible effects of distracting stimuli during eating, such as an increase or decrease in the amount of food consumption. To test this hypothesis, they monitored food intake for a group of 44 patients who were randomized into two equal groups. The treatment group ate lunch while playing solitaire, and the control group ate lunch without any added distractions. Patients in the treatment group ate 52.1 grams of biscuits, with a standard deviation of 45.1 grams, and patients in the control group ate 27.1 grams of biscuits, with a standard deviation of 26.4 grams. Do these data provide convincing evidence that the average food intake (measured in amount of biscuits consumed) is different for the patients in the treatment group? Assume that conditions for inference are satisfied.[39]

7.48 Gaming and distracted eating, Part II. The researchers from Exercise 7.47 also investigated the effects of being distracted by a game on how much people eat. The 22 patients in the treatment group who ate their lunch while playing solitaire were asked to do a serial-order recall of the food lunch items they ate. The average number of items recalled by the patients in this group was 4. 9, with a standard deviation of 1.8. The average number of items recalled by the patients in the control group (no distraction) was 6.1, with a standard deviation of 1.8. Do these data provide strong evidence that the average number of food items recalled by the patients in the treatment and control groups are different?

7.49 Sample size and pairing. Determine if the following statement is true or false, and if false, explain your reasoning: If comparing means of two groups with equal sample sizes, always use a paired test.

7.50 College credits. A college counselor is interested in estimating how many credits a student typically enrolls in each semester. The counselor decides to randomly sample 100 students by using the registrar's database of students. The histogram below shows the distribution of the number of credits taken by these students. Sample statistics for this distribution are also provided.



Min	8
Q1	13
Median	14
Mean	13.65
SD	1.91
Q3	15
Max	18

(a) What is the point estimate for the average number of credits taken per semester by students at this college? What about the median?

(b) What is the point estimate for the standard deviation of the number of credits taken per semester by students at this college? What about the IQR?

(c) Is a load of 16 credits unusually high for this college? What about 18 credits? Explain your reasoning.

(d) The college counselor takes another random sample of 100 students and this time finds a sample mean of 14.02 units. Should she be surprised that this sample statistic is slightly different than the one from the original sample? Explain your reasoning.

(e) The sample means given above are point estimates for the mean number of credits taken by all students at that college. What measures do we use to quantify the variability of this estimate? Compute this quantity using the data from the original sample.

[39] R.E. Oldham-Cooper et al. "Playing a computer game during lunch affects fullness, memory for lunch, and later snack intake". In: *The American Journal of Clinical Nutrition* 93.2 (2011), p. 308.

7.51 Hen eggs. The distribution of the number of eggs laid by a certain species of hen during their breeding period has a mean of 35 eggs with a standard deviation of 18.2. Suppose a group of researchers randomly samples 45 hens of this species, counts the number of eggs laid during their breeding period, and records the sample mean. They repeat this 1,000 times, and build a distribution of sample means.

(a) What is this distribution called?

(b) Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.

(c) Calculate the variability of this distribution and state the appropriate term used to refer to this value.

(d) Suppose the researchers' budget is reduced and they are only able to collect random samples of 10 hens. The sample mean of the number of eggs is recorded, and we repeat this 1,000 times, and build a new distribution of sample means. How will the variability of this new distribution compare to the variability of the original distribution?

7.52 Forest management. Forest rangers wanted to better understand the rate of growth for younger trees in the park. They took measurements of a random sample of 50 young trees in 2009 and again measured those same trees in 2019. The data below summarize their measurements, where the heights are in feet:

	2009	2019	Differences
$\bar{x}$	12.0	24.5	12.5
s	3.5	9.5	7.2
n	50	50	50

Construct a 99% confidence interval for the average growth of (what had been) younger trees in the park over 2009-2019.

7.53 Experiment resizing. At a startup company running a new weather app, an engineering team generally runs experiments where a random sample of 1% of the app's visitors in the control group and another 1% were in the treatment group to test each new feature. The team's core goal is to increase a metric called *daily visitors*, which is essentially the number of visitors to the app each day. They track this metric in each experiment arm and as their core experiment metric. In their most recent experiment, the team tested including a new animation when the app started, and the number of daily visitors in this experiment stabilized at +1.2% with a 95% confidence interval of (-0.2%, +2.6%). This means if this new app start animation was launched, the team thinks they might lose as many as 0.2% of daily visitors or gain as many as 2.6% more daily visitors. Suppose you are consulting as the team's data scientist, and after discussing with the team, you and they agree that they should run another experiment that is bigger. You also agree that this new experiment should be able to detect a gain in the daily visitors metric of 1.0% or more with 80% power. Now they turn to you and ask, "How big of an experiment do we need to run to ensure we can detect this effect?"

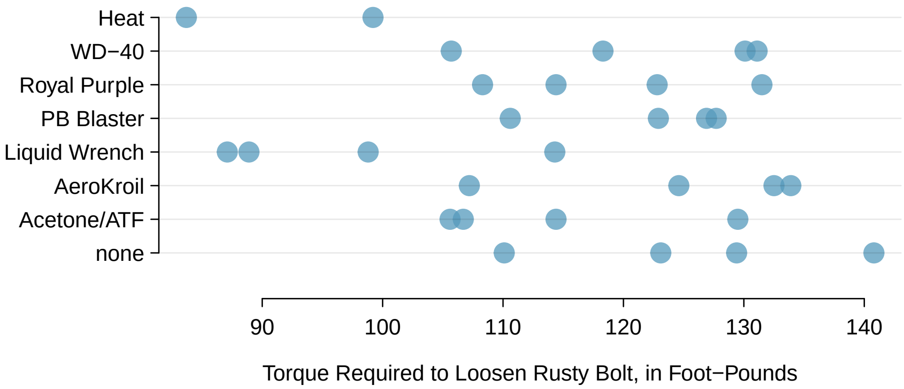
(a) How small must the standard error be if the team is to be able to detect an effect of 1.0% with 80% power and a significance level of $\alpha = 0.05$? You may safely assume the percent change in daily visitors metric follows a normal distribution.

(b) Consider the first experiment, where the point estimate was +1.2% and the 95% confidence interval was (-0.2%, +2.6%). If that point estimate followed a normal distribution, what was the standard error of the estimate?

(c) The ratio of the standard error from part (a) vs the standard error from part (b) should be 1.97. How much bigger of an experiment is needed to shrink a standard error by a factor of 1.97?

(d) Using your answer from part (c) and that the original experiment was a 1% vs 1% experiment to recommend an experiment size to the team.

7.54 Torque on a rusty bolt. Project Farm is a YouTube channel that routinely compares different products. In one episode, the channel evaluated different options for loosening rusty bolts.[40] Eight options were evaluated, including a control group where no treatment was given ("none" in the graph), to determine which was most effective. For all treatments, there were four bolts tested, except for a treatment of heat with a blow torch, where only two data points were collected. The results are shown in the figure below:



(a) Do you think it is reasonable to apply ANOVA in this case?

(b) Regardless of your answer in part (a), describe hypotheses for ANOVA in this context, and use the table below to carry out the test. Give your conclusion in the context of the data.

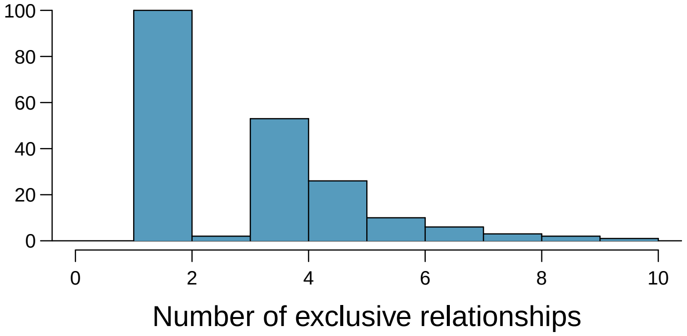
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatment	7	3603.43	514.78	4.03	0.0056
Residuals	22	2812.80	127.85		

(c) The table below are p-values for pairwise *t*-tests comparing each of the different groups. These p-values have not been corrected for multiple comparisons. Which pair of groups appears most likely to represent a difference?

	AeroKroil	Heat	Liquid Wrench	none	PB Blaster	Royal Purple	WD-40
Acetone/ATF	0.2026	0.0308	0.0476	0.1542	0.3294	0.5222	0.3744
AeroKroil		0.0027	0.0025	0.8723	0.7551	0.5143	0.6883
Heat			0.5580	0.0020	0.0050	0.0096	0.0059
Liquid Wrench				0.0017	0.0053	0.0117	0.0065
none					0.6371	0.4180	0.5751
PB Blaster						0.7318	0.9286
Royal Purple							0.8000

(d) There are 28 p-values shown in the table in part (c). Determine if any of them are statistically significant after correcting for multiple comparisons. If so, which one(s)? Explain your answer.

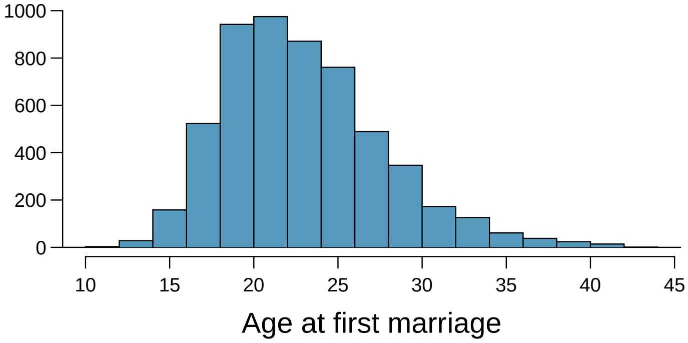
7.55 Exclusive relationships. A survey conducted on a reasonably random sample of 203 undergraduates asked, among many other questions, about the number of exclusive relationships these students have been in. The histogram below shows the distribution of the data from this sample. The sample average is 3.2 with a standard deviation of 1.97.



Estimate the average number of exclusive relationships Duke students have been in using a 90% confidence interval and interpret this interval in context. Check any conditions required for inference, and note any assumptions you must make as you proceed with your calculations and conclusions.

[40] Project Farm on YouTube, youtu.be/xUEob2oAKVs, April 16, 2018.

7.56 Age at first marriage, Part I. The National Survey of Family Growth conducted by the Centers for Disease Control gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health. One of the variables collected on this survey is the age at first marriage. The histogram below shows the distribution of ages at first marriage of 5,534 randomly sampled women between 2006 and 2010. The average age at first marriage among these women is 23.44 with a standard deviation of 4.72.[41]



Estimate the average age at first marriage of women using a 95% confidence interval, and interpret this interval in context. Discuss any relevant assumptions.

7.57 Online communication. A study suggests that the average college student spends 10 hours per week communicating with others online. You believe that this is an underestimate and decide to collect your own sample for a hypothesis test. You randomly sample 60 students from your dorm and find that on average they spent 13.5 hours a week communicating with others online. A friend of yours, who offers to help you with the hypothesis test, comes up with the following set of hypotheses. Indicate any errors you see.

$$H_0 : \bar{x} < 10 \ hours$$
$$H_A : \bar{x} > 13.5 \ hours$$

7.58 Age at first marriage, Part II. Exercise 7.56 presents the results of a 2006 - 2010 survey showing that the average age of women at first marriage is 23.44. Suppose a social scientist thinks this value has changed since the survey was taken. Below is how she set up her hypotheses. Indicate any errors you see.

$$H_0 : \bar{x} \neq 23.44 \ years \ old$$
$$H_A : \bar{x} = 23.44 \ years \ old$$

[41] Centers for Disease Control and Prevention, National Survey of Family Growth, 2010.