Chapter 4

Inference for numerical data

Chapters 2 and 3 introduced us to inference for proportions using the normal model, and in Section 3.3, we encountered the chi-square distribution, which is useful for working with categorical data with many levels. In this chapter, our focus will be on numerical data, where we will encounter two more distributions: the t distribution (looks a lot like the normal distribution) and the F distribution. Our general approach will be:

1. Determine which point estimate or test statistic is useful.
2. Identify an appropriate distribution for the point estimate or test statistic.
3. Apply the hypothesis and confidence interval techniques from Chapter 2 using the distribution from step 2.

4.1 One-sample means with the t distribution

The sampling distribution associated with a sample mean or difference of two sample means is, if certain conditions are satisfied, nearly normal. However, this becomes more complex when the sample size is small, where *small* here typically means a sample size smaller than 30 observations. For this reason, we'll use a new distribution called the t distribution that will often work for both small and large samples of numerical data.

4.1.1 Two examples using the normal distribution

Before we get started with the t distribution, let's take a look at two applications where it is okay to use the normal model for the sample mean. For the case of a single mean, the standard error of the sample mean can be calculated as

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

where σ is the population standard deviation and n is the sample size. Generally we use the sample standard deviation, denoted by s, in place of the population standard deviation when we compute the standard error:

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

If we look at this formula, there are some characteristics that we can think about intuitively.

- If we examine the standard error formula, we would see that a larger s corresponds to a larger SE. This makes intuitive sense: if the data are more volatile, then we'll be less certain of the location of the true mean, so the standard error should be bigger. On the other hand, if the observations all fall very close together, then s will be small, and the sample mean should be a more precise estimate of the true mean.
- In the formula, the larger the sample size n, the smaller the standard error. This matches our intuition: we expect estimates to be more precise when we have more data, so the standard error SE should get smaller when n gets bigger.

As we did with proportions, we'll also need to check a few conditions before using the normal model. We'll forgo describing those details until later this section, but these conditions have been verified for the two examples below.

Example 4.1 We've taken a random sample of 100 runners from a race called the Cherry Blossom Run in Washington, DC, which was a race with 16,924 participants.[1] The sample data for the 100 runners is summarized in Table 4.1, histograms of the run time and age of participants are in Figure 4.2, and summary statistics are available in Table 4.3. Create a 95% confidence interval for the average time it takes runners in the Cherry Blossom Run to complete the race.

We can use the same confidence interval formula for the mean that we used for a proportion:

$$\text{point estimate} \ \pm \ 1.96 \times SE$$

In this case, the best estimate of the overall mean is the sample mean, $\bar{x} = 95.61$ minutes. The standard error can be calculated using sample standard deviation ($s = 15.78$), the sample size ($n = 100$), and the standard error formula:

$$SE = \frac{s}{\sqrt{n}} = \frac{15.78}{\sqrt{100}} = 1.578$$

Finally, we can calculate a 95% confidence interval:

$$\text{point estimate} \ \pm \ z^{\star} \times SE \quad \rightarrow \quad 95.61 \pm 1.96 \times 1.578 \quad \rightarrow \quad (92.52, 98.70)$$

We are 95% confident that the average time for all runners in the 2012 Cherry Blossom Run is between 92.52 and 98.70 minutes.

ID	time	age	gender	state
1	88.31	59	M	MD
2	100.67	32	M	VA
3	109.52	33	F	VA
⋮	⋮	⋮	⋮	⋮
100	89.49	26	M	DC

Table 4.1: Four observations for the `run10Samp` data set, which represents a simple random sample of 100 runners from the 2012 Cherry Blossom Run.

[1] See www.cherryblossom.org.

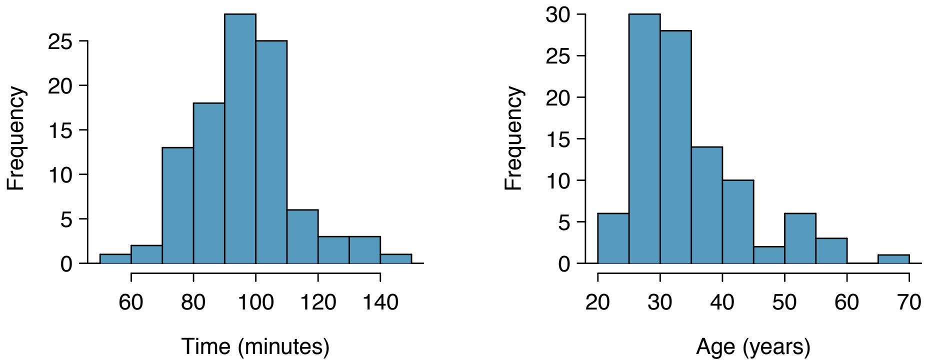


Figure 4.2: Histograms of `time` and `age` for the sample Cherry Blossom Run data. The average time is in the mid-90s, and the average age is in the mid-30s. The age distribution is moderately skewed to the right.

	`time`	`age`
sample mean	95.61	35.05
sample median	95.37	32.50
sample st. dev.	15.78	8.97

Table 4.3: Point estimates and parameter values for the `time` variable.

Guided Practice 4.2 Use the data to calculate a 90% confidence interval for the average age of participants in the 2012 Cherry Blossom Run. The conditions for applying the normal model have already been verified.[2]

Example 4.3 The nutrition label on a bag of potato chips says that a one ounce (28 gram) serving of potato chips has 130 calories and contains ten grams of fat, with three grams of saturated fat. A random sample of 35 bags yielded a sample mean of 134 calories with a standard deviation of 17 calories. Is there evidence that the nutrition label does not provide an accurate measure of calories in the bags of potato chips? The conditions necessary for applying the normal model have been checked and are satisfied.

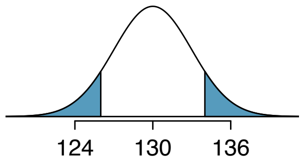
The question has been framed in terms of two possibilities: the nutrition label accurately lists the correct average calories per bag of chips or it does not, which may be framed as a hypothesis test:

H_0: The average is listed correctly. $\mu = 130$

H_A: The nutrition label is incorrect. $\mu \neq 130$

The observed average is $\bar{x} = 134$ and the standard error may be calculated as $SE = \frac{17}{\sqrt{35}} = 2.87$. First, we draw a picture summarizing this scenario.

[2] As before, we identify the point estimate, $\bar{x} = 35.05$, and the standard error, $SE = 8.97/\sqrt{100} = 0.897$. Next, we apply the formula for a 90% confidence interval, which uses $z^\star = 1.65$: $35.05 \pm 1.65 \times 0.897 \to (33.57, 36.53)$. We are 90% confident that the average age of all participants in the 2012 Cherry Blossom Run is between 33.57 and 36.53 years.



We can compute a test statistic as the Z score:

$$Z = \frac{134 - 130}{2.87} = 1.39$$

The upper-tail area is 0.0823, so the p-value is $2 \times 0.0823 = 0.1646$. Since the p-value is larger than 0.05, we do not reject the null hypothesis. That is, there is not enough evidence to show the nutrition label has incorrect information.

The normal model works well when the sample size is larger than about 30. For smaller sample sizes, we run into a problem: our estimate of s, which is used to compute the standard error, isn't as reliable when the sample size is small. To solve this problem, we'll use a new distribution: the t distribution.

4.1.2 Introducing the t distribution

A t distribution, shown as a solid line in Figure 4.4, has a bell shape that looks very similar to a normal distribution (dotted line). However, its tails are thicker, which means observations are more likely to fall beyond two standard deviations from the mean than under the normal distribution.[3] When our sample is small, the value s used to compute the standard error isn't very reliable. The extra thick tails of the t distribution are exactly the correction we need to resolve this problem.

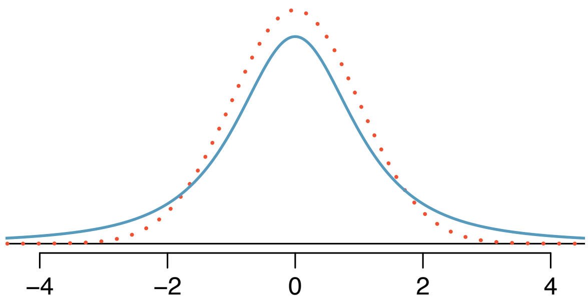


Figure 4.4: Comparison of a t distribution (solid line) and a normal distribution (dotted line).

The t distribution, always centered at zero, has a single parameter: degrees of freedom. The **degrees of freedom (df)** describe the precise form of the bell-shaped t distribution. Several t distributions are shown in Figure 4.5 with various degrees of freedom. When there are more degrees of freedom, the t distribution looks very much like the standard normal distribution.

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

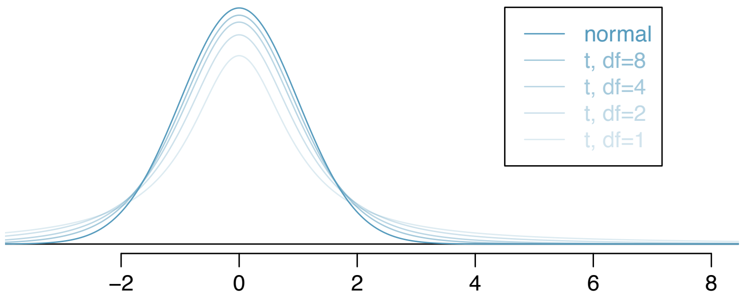


Figure 4.5: The larger the degrees of freedom, the more closely the t distribution resembles the standard normal model.

Degrees of freedom (df)
The degrees of freedom describe the shape of the t distribution. The larger the degrees of freedom, the more closely the distribution approximates the normal model.

When the degrees of freedom is about 30 or more, the t distribution is nearly indistinguishable from the normal distribution, e.g. see Figure 4.5. In Section 4.1.3, we relate degrees of freedom to sample size.

We will find it very useful to become familiar with the t distribution, because it plays a very similar role to the normal distribution during inference for numerical data. We use a **t table**, partially shown in Table 4.6, in place of the normal probability table for small sample numerical data. A larger table is presented in Appendix C.2 on page 342. Alternatively, we could use statistical software to get this same information.

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

Table 4.6: An abbreviated look at the t table. Each row represents a different t distribution. The columns describe the cutoffs for specific tail areas. The row with $df = 18$ has been **highlighted**.

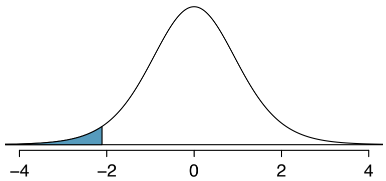


Figure 4.7: The t distribution with 18 degrees of freedom. The area below -2.10 has been shaded.

Each row in the t table represents a t distribution with different degrees of freedom. The columns correspond to tail probabilities. For instance, if we know we are working with the t distribution with $df = 18$, we can examine row 18, which is **highlighted** in Table 4.6. If we want the value in this row that identifies the cutoff for an upper tail of 10%, we can look in the column where *one tail* is 0.100. This cutoff is 1.33. If we had wanted the cutoff for the lower 10%, we would use -1.33. Just like the normal distribution, all t distributions are symmetric.

Example 4.4 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 4.7 and shade the area below -2.10. To find this area, we identify the appropriate row: $df = 18$. Then we identify the column containing the absolute value of -2.10; it is the third column. Because we are looking for just one tail, we examine the top line of the table, which shows that a one tail area for a value in the third row corresponds to 0.025. About 2.5% of the distribution falls below -2.10. In the next example we encounter a case where the exact t value is not listed in the table.

Example 4.5 A t distribution with 20 degrees of freedom is shown in the left panel of Figure 4.8. Estimate the proportion of the distribution falling above 1.65.

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

Example 4.6 A t distribution with 2 degrees of freedom is shown in the right panel of Figure 4.8. Estimate the proportion of the distribution falling more than 3 units from the mean (above or below).

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

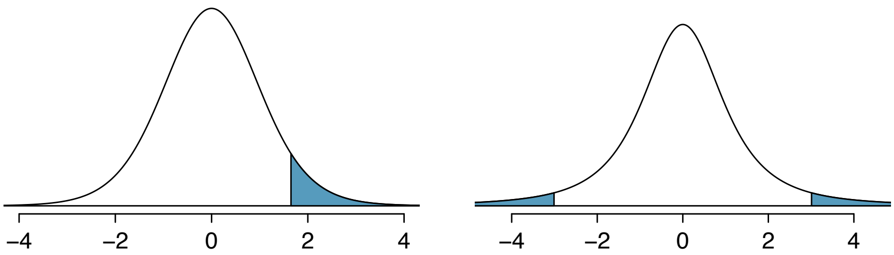


Figure 4.8: Left: The t distribution with 20 degrees of freedom, with the area above 1.65 shaded. Right: The t distribution with 2 degrees of freedom, with the area further than 3 units from 0 shaded.

Guided Practice 4.7 What proportion of the t distribution with 19 degrees of freedom falls above -1.79 units?[4]

4.1.3 Applying the t distribution to the single-mean situation

When estimating the mean and standard error from a sample of numerical data, the t distribution is a little more accurate than the normal model. This is true for both small and large samples, though the benefits for larger samples are limited.

> **Using the t distribution**
> Use the t distribution for inference of the sample mean when observations are independent and nearly normal. You may relax the nearly normal condition as the sample size increases. For example, the data distribution may be moderately skewed when the sample size is at least 30.

Before applying the t distribution for inference about a single mean, we check two conditions.

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

Observations come from a nearly normal distribution. This second condition is difficult to verify with small data sets. We often (i) take a look at a plot of the data for obvious departures from the normal model, usually in the form of prominent outliers, and (ii) consider whether any previous experiences alert us that the data may not be nearly normal. However, if the sample size is somewhat large, then we can relax this condition, e.g. moderate skew is acceptable when the sample size is 30 or more, and strong skew is acceptable when the size is about 60 or more.

When examining a sample mean and estimated standard error from a sample of n independent and nearly normal observations, we use a t distribution with $n - 1$ degrees of freedom (df). For example, if the sample size was 19, then we would use the t distribution

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

with $df = 19 - 1 = 18$ degrees of freedom and proceed in the same way as we did in Chapter 3, except that *now we use the t table*.

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

4.1.4 One sample t confidence intervals

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

Figure 4.9: A Risso's dolphin.

Photo by Mike Baird (http://www.bairdphotos.com/).

Here we identify a confidence interval for the average mercury content in dolphin muscle using a sample of 19 Risso's dolphins from the Taiji area in Japan.[5] The data are summarized in Table 4.10. The minimum and maximum observed values can be used to evaluate whether or not there are obvious outliers or skew.

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

Table 4.10: Summary of mercury content in the muscle of 19 Risso's dolphins from the Taiji area. Measurements are in μg/wet g (micrograms of mercury per wet gram of muscle).

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

Example 4.8 Are the independence and normality conditions satisfied for this data set?

The observations are a simple random sample and consist of less than 10% of the population, therefore independence is reasonable. Ideally we would see a visualization of the data to check for skew and outliers. However, we can instead examine the summary statistics in Table 4.10, which do not suggest any skew or outliers. All observations are within 2.5 standard deviations of the mean. Based on this evidence, the normality assumption seems reasonable.

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

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

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

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

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

- For a 95% confidence interval, we want to find the cutoff $t^\star_{18}$ such that 95% of the t distribution is between -$t^\star_{18}$ and $t^\star_{18}$.
- We look in the t table on page 167, find the column with area totaling 0.05 in the two tails (third column), and then the row with 18 degrees of freedom: $t^\star_{18} = 2.10$.

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

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

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

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

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

where $\bar{x}$ is the sample mean, $t^\star_{df}$ corresponds to the confidence level and degrees of freedom, and SE is the standard error as estimated by the sample. The normality condition may be relaxed for larger sample sizes.

Guided Practice 4.9 The FDA's webpage provides some data on mercury content of fish.[6] 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?[7]

Example 4.10 Estimate the standard error of the sample mean using the data summaries in Guided Practice 4.9. If we are to use the t distribution to create a 90% confidence interval for the actual mean of the mercury content, identify the degrees of freedom we should use and also find $t^{\star}_{df}$.

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

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

Guided Practice 4.11 Using the results of Guided Practice 4.9 and Example 4.10, compute a 90% confidence interval for the average mercury content of croaker white fish (Pacific).[8]

4.1.5 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 Run, comparing runners in 2006 and 2012. Technological advances in shoes, training, and diet might suggest runners would be faster in 2012. An opposing viewpoint might say that with the average body mass index on the rise, people tend to run slower. In fact, all of these components might be influencing run time.

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

Guided Practice 4.12 What are appropriate hypotheses for this context?[9]

Guided Practice 4.13 The data come from a simple random sample from less than 10% of all participants, so the observations are independent. However, should we be worried about skew in the data? A histogram of the differences was shown in the left panel of Figure 4.2 on page 165. [10]

With independence satisfied and skew not a concern, we can proceed with performing a hypothesis test using the t distribution.

---

[6] http://www.fda.gov/food/foodborneillnesscontaminants/metals/ucm115644.htm

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

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

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

[10] With a sample of 100, we should only be concerned if there is extreme skew. The histogram of the data suggest, at worst, slight skew.

⊙ **Guided Practice 4.14** The sample mean and sample standard deviation are 95.61 and 15.78 minutes, respectively. Recall that the sample size is 100. What is the p-value for the test, and what is your conclusion?[11]

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

4.2 Paired data

Are textbooks actually cheaper online? Here we compare the price of textbooks at the University of California, Los Angeles' (UCLA's) bookstore and prices at Amazon.com. Seventy-three UCLA courses were randomly sampled in Spring 2010, representing less than 10% of all UCLA courses.[12] A portion of the data set is shown in Table 4.11.

	dept	course	ucla	amazon	diff
1	Am Ind	C170	27.67	27.95	-0.28
2	Anthro	9	40.59	31.14	9.45
3	Anthro	135T	31.68	32.00	-0.32
4	Anthro	191HB	16.00	11.52	4.48
⋮	⋮	⋮	⋮	⋮	⋮
72	Wom Std	M144	23.76	18.72	5.04
73	Wom Std	285	27.70	18.22	9.48

Table 4.11: Six cases of the `textbooks` data set.

4.2.1 Paired observations

Each textbook has two corresponding prices in the data set: one for the UCLA bookstore and one for Amazon. Therefore, each textbook price from the UCLA bookstore has a natural correspondence with a textbook price from 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.

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

[12] When a class had multiple books, only the most expensive text was considered.

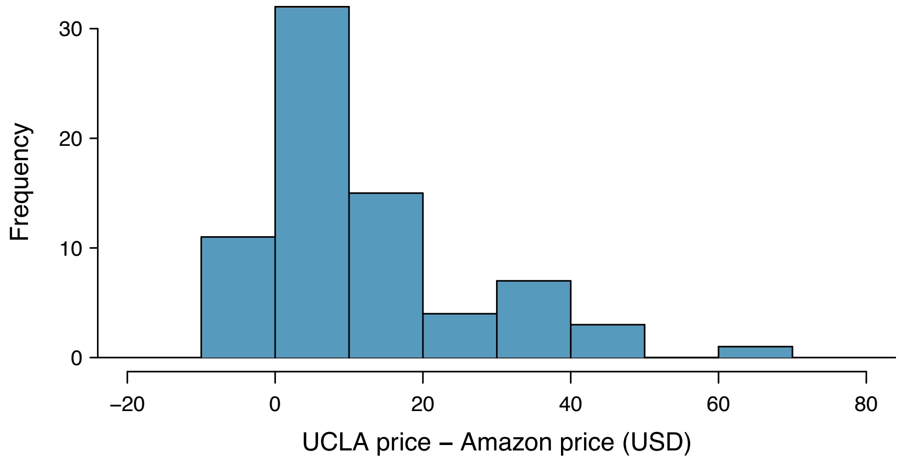


Figure 4.12: Histogram of the difference in price for each book sampled. These data are strongly skewed.

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

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

for each book. It is important that we always subtract using a consistent order; here Amazon prices are always subtracted from UCLA prices. A histogram of these differences is shown in Figure 4.12. Using differences between paired observations is a common and useful way to analyze paired data.

Guided Practice 4.15 The first difference shown in Table 4.11 is computed as $27.67 - 27.95 = -0.28$. Verify the differences are calculated correctly for observations 2 and 3.[13]

4.2.2 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 the last section.

n_{diff}	$\bar{x}_{diff}$	s_{diff}
73	12.76	14.26

Table 4.13: Summary statistics for the price differences. There were 73 books, so there are 73 differences.

[13]Observation 2: $40.59 - 31.14 = 9.45$. Observation 3: $31.68 - 32.00 = -0.32$.

Example 4.16 Set up and implement 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.

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.

Can the t distribution be used for this application? The observations are based on a simple random sample from less than 10% of all books sold at the bookstore, so independence is reasonable. While the distribution is strongly skewed, the sample is reasonably large ($n = 73$), so we can proceed. Because the conditions are reasonably satisfied, we can apply the t distribution to this setting.

We compute the standard error associated with $\bar{x}_{diff}$ using the standard deviation of the differences ($s_{_{diff}} = 14.26$) and the number of differences ($n_{_{diff}} = 73$):

$$SE_{\bar{x}_{diff}} = \frac{s_{diff}}{\sqrt{n_{diff}}} = \frac{14.26}{\sqrt{73}} = 1.67$$

To visualize the p-value, the sampling distribution of $\bar{x}_{diff}$ is drawn as though H_0 is true, which is shown in Figure 4.14. The p-value is represented by the two (very) small tails.

To find the tail areas, we compute the test statistic, which is the T score of $\bar{x}_{diff}$ under the null condition that the actual mean difference is 0:

$$T = \frac{\bar{x}_{diff} - 0}{SE_{x_{diff}}} = \frac{12.76 - 0}{1.67} = 7.59$$

The degrees of freedom are $df = 73 - 1 = 72$. If we examined Appendix C.2 on page 342, we would see that this value is larger than any in the 70 df row (we round down for df when using the table), meaning the two-tailed p-value is less than 0.01. If we used statistical software, we would find the p-value is less than 1-in-10 billion! Because the p-value is less than 0.05, we reject the null hypothesis. We have found convincing evidence that Amazon is, on average, cheaper than the UCLA bookstore for UCLA course textbooks.

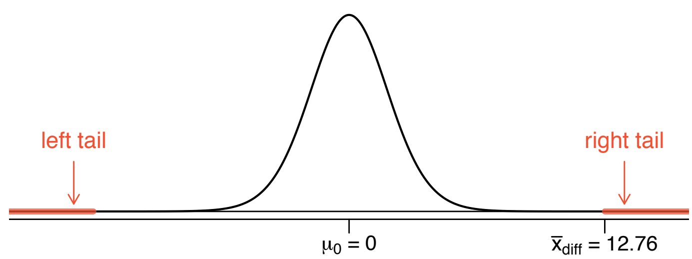


Figure 4.14: Sampling distribution for the mean difference in book prices, if the true average difference is zero.

Guided Practice 4.17 Create a 95% confidence interval for the average price difference between books at the UCLA bookstore and books on Amazon.[14]

In the textbook price example, we applied the t distribution. However, as we mentioned in the last section, the t distribution looks a lot like the normal distribution when the degrees of freedom are larger than about 30. In such cases, including this one, it would be reasonable to use the normal distribution in place of the t distribution.

4.3 Difference of two means

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

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

4.3.1 Confidence interval for a differences of means

Does treatment using embryonic stem cells (ESCs) help improve heart function following a heart attack? Table 4.15 contains summary statistics for an experiment to test ESCs in sheep that had a heart attack. Each of these sheep was randomly assigned to the ESC or control group, and the change in their hearts' pumping capacity was measured in the study. A positive value corresponds to increased pumping capacity, which generally suggests a stronger recovery. Our goal will be to identify a 95% confidence interval for the effect of ESCs on the change in heart pumping capacity relative to the control group.

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

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

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

Table 4.15: Summary statistics of the embryonic stem cell study.

[14]Conditions have already verified and the standard error computed in Example 4.16. To find the interval, identify $t^\star_{72}$ (use $df = 70$ in the table, $t^\star_{70} = 1.99$) and plug it, the point estimate, and the standard error into the confidence interval formula:

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

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