tell us how unusual our sample is. If H_0 is true:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{71.8 - 70.43}{1.22} = 1.12$$

A Z score of just 1.12 is not very unusual (we typically use a threshold of ± 2 to decide what is unusual), so there is not strong evidence against the claim that the heights are representative. This does not mean the heights are actually representative, only that this very small sample does not necessarily show otherwise.

TIP: Relaxing the nearly normal condition
As the sample size becomes larger, it is reasonable to *slowly* relax the nearly normal assumption on the data when dealing with small samples. By the time the sample size reaches 30, the data must show strong skew for us to be concerned about the normality of the sampling distribution.

4.3 Hypothesis testing

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.

In addition to considering run times in this section, we consider a topic near and dear to most students: sleep. A recent study found that college students average about 7 hours of sleep per night.[15] However, researchers at a rural college are interested in showing that their students sleep longer than seven hours on average. We investigate this topic in Section 4.3.4.

4.3.1 Hypothesis testing framework

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 if the `run10Samp` data set provides strong evidence that the participants in 2012 were faster or slower than those runners in 2006, versus the other possibility that there has been no change.[16] We simplify these three options into two competing **hypotheses**:

H_0: The average 10 mile run time was the same for 2006 and 2012.

H_A: The average 10 mile run time for 2012 was *different* than that of 2006.

We call H_0 the null hypothesis and H_A the alternative hypothesis.

H_0
null hypothesis

H_A
alternative hypothesis

Null and alternative hypotheses
The **null hypothesis** (H_0) often represents either a skeptical perspective or a claim to be tested. The **alternative hypothesis** (H_A) represents an alternative claim under consideration and is often represented by a range of possible parameter values.

[15] http://theloquitur.com/?p=1161

[16] While we could answer this question by examining the entire population data (`run10`), we only consider the sample data (`run10Samp`), which is more realistic since we rarely have access to population data.

The null hypothesis often represents a skeptical position or a perspective of no difference. The alternative hypothesis often represents a new perspective, such as the possibility that there has been a change.

TIP: Hypothesis testing framework
The skeptic will not reject the null hypothesis (H_0), unless the evidence in favor of the alternative hypothesis (H_A) is so strong that she rejects H_0 in favor of H_A.

The hypothesis testing framework is a very general tool, and we often use it without a second thought. If a person makes a somewhat unbelievable claim, we are initially skeptical. However, if there is sufficient evidence that supports the claim, we set aside our skepticism and reject the null hypothesis in favor of the alternative. The hallmarks of hypothesis testing are also found in the US court system.

⊙ **Exercise 4.20** A US court considers two possible claims about a defendant: she is either innocent or guilty. If we set these claims up in a hypothesis framework, which would be the null hypothesis and which the alternative?[17]

Jurors examine the evidence to see whether it convincingly shows a defendant is guilty. Even if the jurors leave unconvinced of guilt beyond a reasonable doubt, this does not mean they believe the defendant is innocent. This is also the case with hypothesis testing: *even if we fail to reject the null hypothesis, we typically do not accept the null hypothesis as true.* Failing to find strong evidence for the alternative hypothesis is not equivalent to accepting the null hypothesis.

In the example with the Cherry Blossom Run, the null hypothesis represents no difference in the average time from 2006 to 2012. The alternative hypothesis represents something new or more interesting: there was a difference, either an increase or a decrease. These hypotheses can be described in mathematical notation using μ_{12} as the average run time for 2012:

H_0: $\mu_{12} = 93.29$

H_A: $\mu_{12} \neq 93.29$

where 93.29 minutes (93 minutes and about 17 seconds) is the average 10 mile time for all runners in the 2006 Cherry Blossom Run. Using this mathematical notation, the hypotheses can now be evaluated using statistical tools. We call 93.29 the **null value** since it represents the value of the parameter if the null hypothesis is true. We will use the `run10Samp` data set to evaluate the hypothesis test.

4.3.2 Testing hypotheses using confidence intervals

We can start the evaluation of the hypothesis setup by comparing 2006 and 2012 run times using a point estimate from the 2012 sample: $\bar{x}_{12} = 95.61$ minutes. This estimate suggests the average time is actually longer than the 2006 time, 93.29 minutes. However, to evaluate whether this provides strong evidence that there has been a change, we must consider the uncertainty associated with $\bar{x}_{12}$.

[17] The jury considers whether the evidence is so convincing (strong) that there is no reasonable doubt regarding the person's guilt; in such a case, the jury rejects innocence (the null hypothesis) and concludes the defendant is guilty (alternative hypothesis).

We learned in Section 4.1 that there is fluctuation from one sample to another, and it is very unlikely that the sample mean will be exactly equal to our parameter; we should not expect $\bar{x}_{12}$ to exactly equal μ_{12}. Given that $\bar{x}_{12} = 95.61$, it might still be possible that the population average in 2012 has remained unchanged from 2006. The difference between $\bar{x}_{12}$ and 93.29 could be due to *sampling variation*, i.e. the variability associated with the point estimate when we take a random sample.

In Section 4.2, confidence intervals were introduced as a way to find a range of plausible values for the population mean. Based on `run10Samp`, a 95% confidence interval for the 2012 population mean, μ_{12}, was calculated as

$$(92.45, 98.77)$$

Because the 2006 mean, 93.29, falls in the range of plausible values, we cannot say the null hypothesis is implausible. That is, we failed to reject the null hypothesis, H_0.

TIP: Double negatives can sometimes be used in statistics
In many statistical explanations, we use double negatives. For instance, we might say that the null hypothesis is *not implausible* or we *failed to reject* the null hypothesis. Double negatives are used to communicate that while we are not rejecting a position, we are also not saying it is correct.

Example 4.21 Next consider whether there is strong evidence that the average age of runners has changed from 2006 to 2012 in the Cherry Blossom Run. In 2006, the average age was 36.13 years, and in the 2012 `run10Samp` data set, the average was 35.05 years with a standard deviation of 8.97 years for 100 runners.

First, set up the hypotheses:

H_0: The average age of runners has not changed from 2006 to 2012, $\mu_{age} = 36.13$.
H_A: The average age of runners has changed from 2006 to 2012, $\mu_{age} \neq 36.13$.

We have previously verified conditions for this data set. The normal model may be applied to $\bar{y}$ and the estimate of SE should be very accurate. Using the sample mean and standard error, we can construct a 95% confidence interval for μ_{age} to determine if there is sufficient evidence to reject H_0:

$$\bar{y} \pm 1.96 \times \frac{s}{\sqrt{100}} \quad \rightarrow \quad 35.05 \pm 1.96 \times 0.90 \quad \rightarrow \quad (33.29, 36.81)$$

This confidence interval contains the *null value*, 36.13. Because 36.13 is not implausible, we cannot reject the null hypothesis. We have not found strong evidence that the average age is different than 36.13 years.

Exercise 4.22 Colleges frequently provide estimates of student expenses such as housing. A consultant hired by a community college claimed that the average student housing expense was \$650 per month. What are the null and alternative hypotheses to test whether this claim is accurate?[18]

[18] H_0: The average cost is \$650 per month, $\mu = \$650$.
H_A: The average cost is different than \$650 per month, $\mu \neq \$650$.

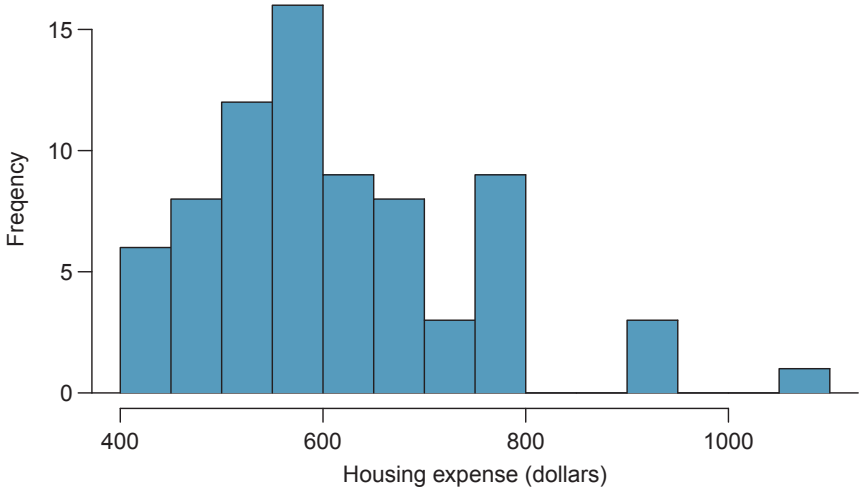


Figure 4.11: Sample distribution of student housing expense. These data are moderately skewed, roughly determined using the outliers on the right.

⊙ **Exercise 4.23** The community college decides to collect data to evaluate the $650 per month claim. They take a random sample of 75 students at their school and obtain the data represented in Figure 4.11. Can we apply the normal model to the sample mean?[19]

● **Example 4.24** The sample mean for student housing is $611.63 and the sample standard deviation is $132.85. Construct a 95% confidence interval for the population mean and evaluate the hypotheses of Exercise 4.22.

The standard error associated with the mean may be estimated using the sample standard deviation divided by the square root of the sample size. Recall that $n = 75$ students were sampled.

$$SE = \frac{s}{\sqrt{n}} = \frac{132.85}{\sqrt{75}} = 15.34$$

You showed in Exercise 4.23 that the normal model may be applied to the sample mean. This ensures a 95% confidence interval may be accurately constructed:

$$\bar{x} \pm z^\star SE \quad \rightarrow \quad 611.63 \pm 1.96 \times 15.34 \quad \rightarrow \quad (581.56, 641.70)$$

Because the null value $650 is not in the confidence interval, a true mean of $650 is implausible and we reject the null hypothesis. The data provide statistically significant evidence that the actual average housing expense is less than $650 per month.

[19] Applying the normal model requires that certain conditions are met. Because the data are a simple random sample and the sample (presumably) represents no more than 10% of all students at the college, the observations are independent. The sample size is also sufficiently large ($n = 75$) and the data exhibit only moderate skew. Thus, the normal model may be applied to the sample mean.

4.3.3 Decision errors

Hypothesis tests are not flawless. Just think of the court system: innocent people are sometimes wrongly convicted and the guilty sometimes walk free. Similarly, we can make a wrong decision in statistical hypothesis tests. However, the difference is that we have the tools necessary to quantify how often we make such errors.

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a statement about which one might be true, but we might choose incorrectly. There are four possible scenarios in a hypothesis test, which are summarized in Table 4.12.

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	Type 1 Error
	H_A true	Type 2 Error	okay

Table 4.12: Four different scenarios for hypothesis tests.

A **Type 1 Error** is rejecting the null hypothesis when H_0 is actually true. A **Type 2 Error** is failing to reject the null hypothesis when the alternative is actually true.

⊙ **Exercise 4.25** In a US court, the defendant is either innocent (H_0) or guilty (H_A). What does a Type 1 Error represent in this context? What does a Type 2 Error represent? Table 4.12 may be useful.[20]

⊙ **Exercise 4.26** How could we reduce the Type 1 Error rate in US courts? What influence would this have on the Type 2 Error rate?[21]

⊙ **Exercise 4.27** How could we reduce the Type 2 Error rate in US courts? What influence would this have on the Type 1 Error rate?[22]

Exercises 4.25-4.27 provide an important lesson: if we reduce how often we make one type of error, we generally make more of the other type.

Hypothesis testing is built around rejecting or failing to reject the null hypothesis. That is, we do not reject H_0 unless we have strong evidence. But what precisely does *strong evidence* mean? As a general rule of thumb, for those cases where the null hypothesis is actually true, we do not want to incorrectly reject H_0 more than 5% of the time. This corresponds to a **significance level** of 0.05. We often write the significance level using α (the Greek letter *alpha*): $\alpha = 0.05$. We discuss the appropriateness of different significance levels in Section 4.3.6.

α
significance level of a hypothesis test

If we use a 95% confidence interval to test a hypothesis where the null hypothesis is true, we will make an error whenever the point estimate is at least 1.96 standard errors

[20] If the court makes a Type 1 Error, this means the defendant is innocent (H_0 true) but wrongly convicted. A Type 2 Error means the court failed to reject H_0 (i.e. failed to convict the person) when she was in fact guilty (H_A true).

[21] To lower the Type 1 Error rate, we might raise our standard for conviction from "beyond a reasonable doubt" to "beyond a conceivable doubt" so fewer people would be wrongly convicted. However, this would also make it more difficult to convict the people who are actually guilty, so we would make more Type 2 Errors.

[22] To lower the Type 2 Error rate, we want to convict more guilty people. We could lower the standards for conviction from "beyond a reasonable doubt" to "beyond a little doubt". Lowering the bar for guilt will also result in more wrongful convictions, raising the Type 1 Error rate.

away from the population parameter. This happens about 5% of the time (2.5% in each tail). Similarly, using a 99% confidence interval to evaluate a hypothesis is equivalent to a significance level of $\alpha = 0.01$.

A confidence interval is, in one sense, simplistic in the world of hypothesis tests. Consider the following two scenarios:

- The null value (the parameter value under the null hypothesis) is in the 95% confidence interval but just barely, so we would not reject H_0. However, we might like to somehow say, quantitatively, that it was a close decision.
- The null value is very far outside of the interval, so we reject H_0. However, we want to communicate that, not only did we reject the null hypothesis, but it wasn't even close. Such a case is depicted in Figure 4.13.

In Section 4.3.4, we introduce a tool called the *p-value* that will be helpful in these cases. The p-value method also extends to hypothesis tests where confidence intervals cannot be easily constructed or applied.

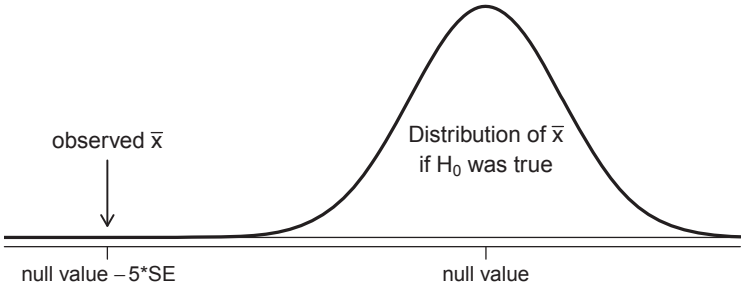


Figure 4.13: It would be helpful to quantify the strength of the evidence against the null hypothesis. In this case, the evidence is extremely strong.

4.3.4 Formal testing using p-values

The p-value is a way of quantifying the strength of the evidence against the null hypothesis and in favor of the alternative. Formally the *p-value* is a conditional probability.

p-value
The **p-value** is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis is true. We typically use a summary statistic of the data, in this chapter the sample mean, to help compute the p-value and evaluate the hypotheses.

⊙ **Exercise 4.28** A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. Researchers at a rural school are interested in showing that students at their school sleep longer than seven hours on average, and they would like to demonstrate this using a sample of students. What would be an appropriate skeptical position for this research?[23]

[23] A skeptic would have no reason to believe that sleep patterns at this school are different than the sleep patterns at another school.

We can set up the null hypothesis for this test as a skeptical perspective: the students at this school average 7 hours of sleep per night. The alternative hypothesis takes a new form reflecting the interests of the research: the students average more than 7 hours of sleep. We can write these hypotheses as

H_0: $\mu = 7$.

H_A: $\mu > 7$.

Using $\mu > 7$ as the alternative is an example of a **one-sided** hypothesis test. In this investigation, there is no apparent interest in learning whether the mean is less than 7 hours.[24] Earlier we encountered a **two-sided** hypothesis where we looked for any clear difference, greater than or less than the null value.

Always use a two-sided test unless it was made clear prior to data collection that the test should be one-sided. Switching a two-sided test to a one-sided test after observing the data is dangerous because it can inflate the Type 1 Error rate.

TIP: One-sided and two-sided tests
If the researchers are only interested in showing an increase or a decrease, but not both, use a one-sided test. If the researchers would be interested in any difference from the null value – an increase or decrease – then the test should be two-sided.

TIP: Always write the null hypothesis as an equality
We will find it most useful if we always list the null hypothesis as an equality (e.g. $\mu = 7$) while the alternative always uses an inequality (e.g. $\mu \neq 7$, $\mu > 7$, or $\mu < 7$).

The researchers at the rural school conducted a simple random sample of $n = 110$ students on campus. They found that these students averaged 7.42 hours of sleep and the standard deviation of the amount of sleep for the students was 1.75 hours. A histogram of the sample is shown in Figure 4.14.

Before we can use a normal model for the sample mean or compute the standard error of the sample mean, we must verify conditions. (1) Because this is a simple random sample from less than 10% of the student body, the observations are independent. (2) The sample size in the sleep study is sufficiently large since it is greater than 30. (3) The data show moderate skew in Figure 4.14 and the presence of a couple of outliers. This skew and the outliers (which are not too extreme) are acceptable for a sample size of $n = 110$. With these conditions verified, the normal model can be safely applied to $\bar{x}$ and the estimated standard error will be very accurate.

⊙ **Exercise 4.29** What is the standard deviation associated with $\bar{x}$? That is, estimate the standard error of $\bar{x}$.[25]

The hypothesis test will be evaluated using a significance level of $\alpha = 0.05$. We want to consider the data under the scenario that the null hypothesis is true. In this case, the sample mean is from a distribution that is nearly normal and has mean 7 and standard deviation of about 0.17. Such a distribution is shown in Figure 4.15.

[24] This is entirely based on the interests of the researchers. Had they been only interested in the opposite case – showing that their students were actually averaging fewer than seven hours of sleep but not interested in showing more than 7 hours – then our setup would have set the alternative as $\mu < 7$.

[25] The standard error can be estimated from the sample standard deviation and the sample size: $SE_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \frac{1.75}{\sqrt{110}} = 0.17$.

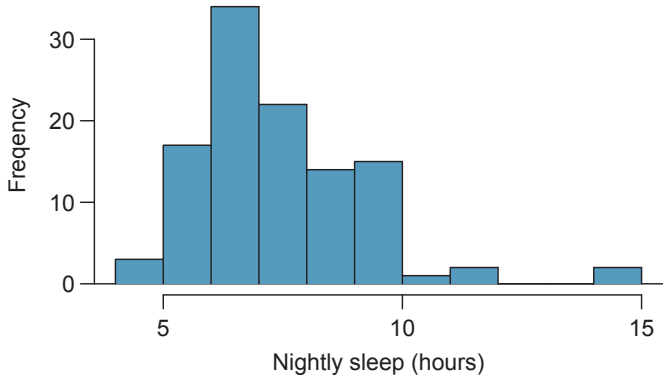


Figure 4.14: Distribution of a night of sleep for 110 college students. These data are moderately skewed.

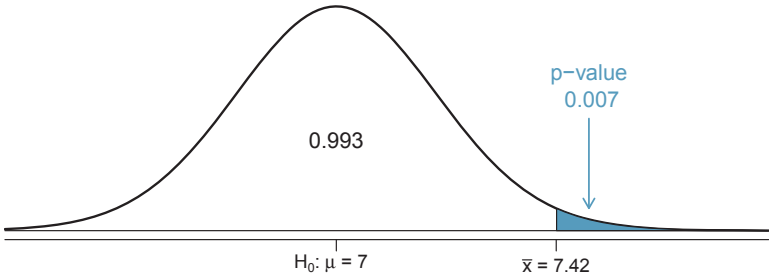


Figure 4.15: If the null hypothesis is true, then the sample mean $\bar{x}$ came from this nearly normal distribution. The right tail describes the probability of observing such a large sample mean if the null hypothesis is true.

The shaded tail in Figure 4.15 represents the chance of observing such a large mean, conditional on the null hypothesis being true. That is, the shaded tail represents the p-value. We shade all means larger than our sample mean, $\bar{x} = 7.42$, because they are more favorable to the alternative hypothesis than the observed mean.

We compute the p-value by finding the tail area of this normal distribution, which we learned to do in Section 3.1. First compute the Z score of the sample mean, $\bar{x} = 7.42$:

$$Z = \frac{\bar{x} - \text{null value}}{SE_{\bar{x}}} = \frac{7.42 - 7}{0.17} = 2.47$$

Using the normal probability table, the lower unshaded area is found to be 0.993. Thus the shaded area is $1 - 0.993 = 0.007$. *If the null hypothesis is true, the probability of observing such a large sample mean for a sample of 110 students is only 0.007.* That is, if the null hypothesis is true, we would not often see such a large mean.

We evaluate the hypotheses by comparing the p-value to the significance level. Because the p-value is less than the significance level (p-value $= 0.007 < 0.05 = \alpha$), we reject the null hypothesis. What we observed is so unusual with respect to the null hypothesis that it casts serious doubt on H_0 and provides strong evidence favoring H_A.

p-value as a tool in hypothesis testing
The p-value quantifies how strongly the data favor H_A over H_0. A small p-value (usually < 0.05) corresponds to sufficient evidence to reject H_0 in favor of H_A.

TIP: It is useful to first draw a picture to find the p-value
It is useful to draw a picture of the distribution of $\bar{x}$ as though H_0 was true (i.e. μ equals the null value), and shade the region (or regions) of sample means that are at least as favorable to the alternative hypothesis. These shaded regions represent the p-value.

The ideas below review the process of evaluating hypothesis tests with p-values:

- The null hypothesis represents a skeptic's position or a position of no difference. We reject this position only if the evidence strongly favors H_A.
- A small p-value means that if the null hypothesis is true, there is a low probability of seeing a point estimate at least as extreme as the one we saw. We interpret this as strong evidence in favor of the alternative.
- We reject the null hypothesis if the p-value is smaller than the significance level, α, which is usually 0.05. Otherwise, we fail to reject H_0.
- We should always state the conclusion of the hypothesis test in plain language so non-statisticians can also understand the results.

The p-value is constructed in such a way that we can directly compare it to the significance level (α) to determine whether or not to reject H_0. This method ensures that the Type 1 Error rate does not exceed the significance level standard.

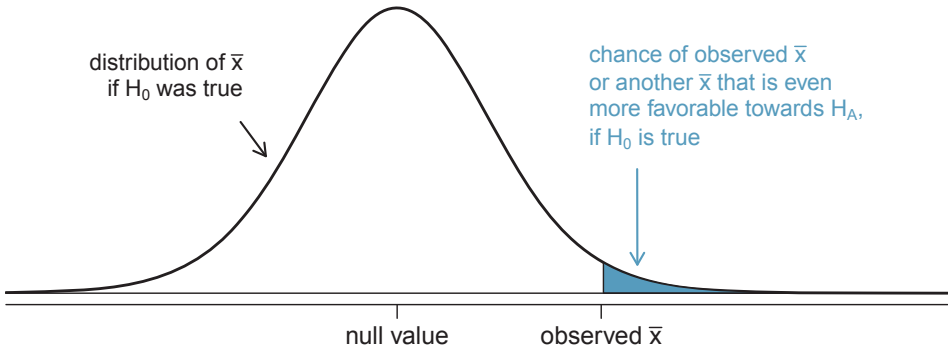


Figure 4.16: To identify the p-value, the distribution of the sample mean is considered as if the null hypothesis was true. Then the p-value is defined and computed as the probability of the observed $\bar{x}$ or an $\bar{x}$ even more favorable to H_A under this distribution.

⊙ **Exercise 4.30** If the null hypothesis is true, how often should the p-value be less than 0.05?[26]

[26] About 5% of the time. If the null hypothesis is true, then the data only has a 5% chance of being in the 5% of data most favorable to H_A.

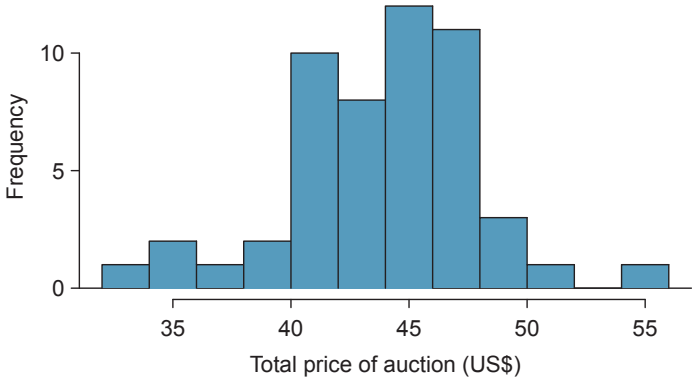


Figure 4.17: A histogram of the total auction prices for 52 Ebay auctions.

⊙ **Exercise 4.31** Suppose we had used a significance level of 0.01 in the sleep study. Would the evidence have been strong enough to reject the null hypothesis? (The p-value was 0.007.) What if the significance level was $\alpha = 0.001$? [27]

⊙ **Exercise 4.32** Ebay might be interested in showing that buyers on its site tend to pay less than they would for the corresponding new item on Amazon. We'll research this topic for one particular product: a video game called *Mario Kart* for the Nintendo Wii. During early October 2009, Amazon sold this game for $46.99. Set up an appropriate (one-sided!) hypothesis test to check the claim that Ebay buyers pay less during auctions at this same time.[28]

⊙ **Exercise 4.33** During early October, 2009, 52 Ebay auctions were recorded for *Mario Kart*.[29] The total prices for the auctions are presented using a histogram in Figure 4.17, and we may like to apply the normal model to the sample mean. Check the three conditions required for applying the normal model: (1) independence, (2) at least 30 observations, and (3) the data are not strongly skewed.[30]

● **Example 4.34** The average sale price of the 52 Ebay auctions for *Wii Mario Kart* was $44.17 with a standard deviation of $4.15. Does this provide sufficient evidence to reject the null hypothesis in Exercise 4.32? Use a significance level of $\alpha = 0.01$.

The hypotheses were set up and the conditions were checked in Exercises 4.32 and 4.33. The next step is to find the standard error of the sample mean and produce a sketch

[27] We reject the null hypothesis whenever $p\text{-}value < \alpha$. Thus, we would still reject the null hypothesis if $\alpha = 0.01$ but not if the significance level had been $\alpha = 0.001$.

[28] The skeptic would say the average is the same on Ebay, and we are interested in showing the average price is lower.

H_0: The average auction price on Ebay is equal to (or more than) the price on Amazon. We write only the equality in the statistical notation: $\mu_{ebay} = 46.99$.

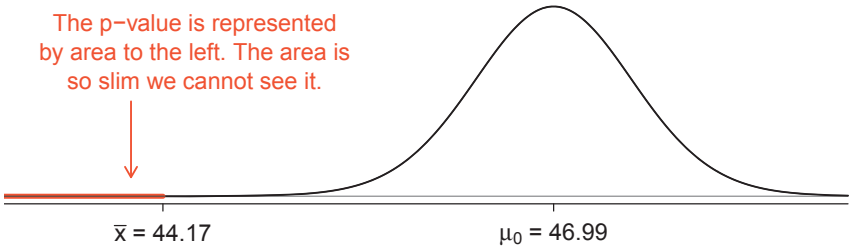
H_A: The average price on Ebay is less than the price on Amazon, $\mu_{ebay} < 46.99$.

[29] These data were collected by OpenIntro staff.

[30] (1) The independence condition is unclear. *We will make the assumption that the observations are independent, which we should report with any final results.* (2) The sample size is sufficiently large: $n = 52 \geq 30$. (3) The data distribution is not strongly skewed; it is approximately symmetric.

to help find the p-value.

$$SE_{\bar{x}} = s/\sqrt{n} = 4.15/\sqrt{52} = 0.5755$$



Because the alternative hypothesis says we are looking for a smaller mean, we shade the lower tail. We find this shaded area by using the Z score and normal probability table: $Z = \frac{44.17-46.99}{0.5755} = -4.90$, which has area less than 0.0002. The area is so small we cannot really see it on the picture. This lower tail area corresponds to the p-value.

Because the p-value is so small – specifically, smaller than $\alpha = 0.01$ – this provides sufficiently strong evidence to reject the null hypothesis in favor of the alternative. The data provide statistically significant evidence that the average price on Ebay is lower than Amazon's asking price.

4.3.5 Two-sided hypothesis testing with p-values

We now consider how to compute a p-value for a two-sided test. In one-sided tests, we shade the single tail in the direction of the alternative hypothesis. For example, when the alternative had the form $\mu > 7$, then the p-value was represented by the upper tail (Figure 4.16). When the alternative was $\mu < 46.99$, the p-value was the lower tail (Exercise 4.32). In a two-sided test, *we shade two tails* since evidence in either direction is favorable to H_A.

⊙ **Exercise 4.35** Earlier we talked about a research group investigating whether the students at their school slept longer than 7 hours each night. Let's consider a second group of researchers who want to evaluate whether the students at their college differ from the norm of 7 hours. Write the null and alternative hypotheses for this investigation.[31]

● **Example 4.36** The second college randomly samples 72 students and finds a mean of $\bar{x} = 6.83$ hours and a standard deviation of $s = 1.8$ hours. Does this provide strong evidence against H_0 in Exercise 4.35? Use a significance level of $\alpha = 0.05$.

First, we must verify assumptions. (1) A simple random sample of less than 10% of the student body means the observations are independent. (2) The sample size is 72, which is greater than 30. (3) Based on the earlier distribution and what we already know about college student sleep habits, the distribution is probably not strongly skewed.

Next we can compute the standard error ($SE_{\bar{x}} = \frac{s}{\sqrt{n}} = 0.21$) of the estimate and create a picture to represent the p-value, shown in Figure 4.18. Both tails are shaded.

[31] Because the researchers are interested in any difference, they should use a two-sided setup: $H_0 : \mu = 7$, $H_A : \mu \neq 7$.

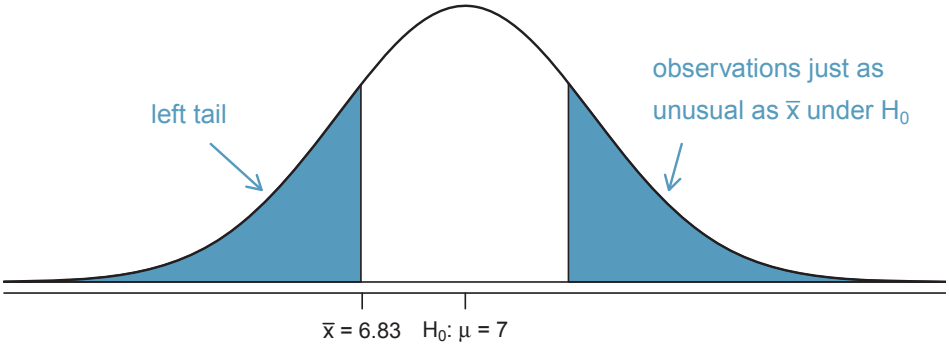


Figure 4.18: H_A is two-sided, so *both* tails must be counted for the p-value.

An estimate of 7.17 or more provides at least as strong of evidence against the null hypothesis and in favor of the alternative as the observed estimate, $\bar{x} = 6.83$.

We can calculate the tail areas by first finding the lower tail corresponding to $\bar{x}$:

$$Z = \frac{6.83 - 7.00}{0.21} = -0.81 \quad \overset{table}{\rightarrow} \quad \text{left tail} = 0.2090$$

Because the normal model is symmetric, the right tail will have the same area as the left tail. The p-value is found as the sum of the two shaded tails:

$$\text{p-value} = \text{left tail} + \text{right tail} = 2 \times (\text{left tail}) = 0.4180$$

This p-value is relatively large (larger than $\alpha = 0.05$), so we should not reject H_0. That is, if H_0 is true, it would not be very unusual to see a sample mean this far from 7 hours simply due to sampling variation. Thus, we do not have sufficient evidence to conclude that the mean is different than 7 hours.

Example 4.37 It is never okay to change two-sided tests to one-sided tests after observing the data. In this example we explore the consequences of ignoring this advice. Using $\alpha = 0.05$, we show that freely switching from two-sided tests to one-sided tests will cause us to make twice as many Type 1 Errors as intended.

Suppose the sample mean was larger than the null value, μ_0 (e.g. μ_0 would represent 7 if H_0: $\mu = 7$). Then if we can flip to a one-sided test, we would use H_A: $\mu > \mu_0$. Now if we obtain any observation with a Z score greater than 1.65, we would reject H_0. If the null hypothesis is true, we incorrectly reject the null hypothesis about 5% of the time when the sample mean is above the null value, as shown in Figure 4.19.

Suppose the sample mean was smaller than the null value. Then if we change to a one-sided test, we would use H_A: $\mu < \mu_0$. If $\bar{x}$ had a Z score smaller than -1.65, we would reject H_0. If the null hypothesis is true, then we would observe such a case about 5% of the time.

By examining these two scenarios, we can determine that we will make a Type 1 Error $5\% + 5\% = 10\%$ of the time if we are allowed to swap to the "best" one-sided test for the data. This is twice the error rate we prescribed with our significance level: $\alpha = 0.05$ (!).

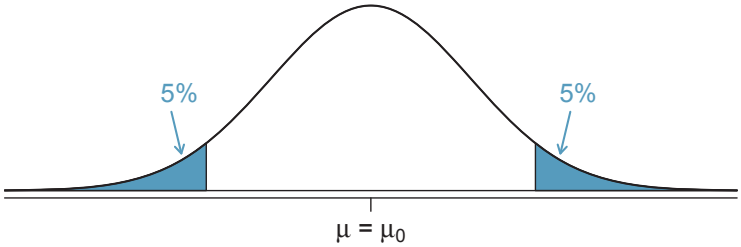


Figure 4.19: The shaded regions represent areas where we would reject H_0 under the bad practices considered in Example 4.37 when $\alpha = 0.05$.

Caution: One-sided hypotheses are allowed only *before* seeing data
After observing data, it is tempting to turn a two-sided test into a one-sided test. Avoid this temptation. Hypotheses must be set up *before* observing the data. If they are not, the test must be two-sided.

4.3.6 Choosing a significance level

Choosing a significance level for a test is important in many contexts, and the traditional level is 0.05. However, it is often helpful to adjust the significance level based on the application. We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.

If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring H_A before we would reject H_0.

If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false. We will discuss this particular case in greater detail in Section 4.6.

Significance levels should reflect consequences of errors
The significance level selected for a test should reflect the consequences associated with Type 1 and Type 2 Errors.

Example 4.38 A car manufacturer is considering a higher quality but more expensive supplier for window parts in its vehicles. They sample a number of parts from their current supplier and also parts from the new supplier. They decide that if the high quality parts will last more than 12% longer, it makes financial sense to switch to this more expensive supplier. Is there good reason to modify the significance level in such a hypothesis test?

The null hypothesis is that the more expensive parts last no more than 12% longer while the alternative is that they do last more than 12% longer. This decision is just one of the many regular factors that have a marginal impact on the car and company. A significance level of 0.05 seems reasonable since neither a Type 1 or Type 2 error should be dangerous or (relatively) much more expensive.

Example 4.39 The same car manufacturer is considering a slightly more expensive supplier for parts related to safety, not windows. If the durability of these safety components is shown to be better than the current supplier, they will switch manufacturers. Is there good reason to modify the significance level in such an evaluation?

The null hypothesis would be that the suppliers' parts are equally reliable. Because safety is involved, the car company should be eager to switch to the slightly more expensive manufacturer (reject H_0) even if the evidence of increased safety is only moderately strong. A slightly larger significance level, such as $\alpha = 0.10$, might be appropriate.

Exercise 4.40 A part inside of a machine is very expensive to replace. However, the machine usually functions properly even if this part is broken, so the part is replaced only if we are extremely certain it is broken based on a series of measurements. Identify appropriate hypotheses for this test (in plain language) and suggest an appropriate significance level.[32]

4.4 Examining the Central Limit Theorem

The normal model for the sample mean tends to be very good when the sample consists of at least 30 independent observations and the population data are not strongly skewed. The Central Limit Theorem provides the theory that allows us to make this assumption.

Central Limit Theorem, informal definition
The distribution of $\bar{x}$ is approximately normal. The approximation can be poor if the sample size is small, but it improves with larger sample sizes.

The Central Limit Theorem states that when the sample size is small, the normal approximation may not be very good. However, as the sample size becomes large, the normal approximation improves. We will investigate three cases to see roughly when the approximation is reasonable.

We consider three data sets: one from a *uniform* distribution, one from an *exponential* distribution, and the other from a *log-normal* distribution. These distributions are shown in the top panels of Figure 4.20. The uniform distribution is symmetric, the exponential distribution may be considered as having moderate skew since its right tail is relatively short (few outliers), and the log-normal distribution is strongly skewed and will tend to produce more apparent outliers.

The left panel in the $n = 2$ row represents the sampling distribution of $\bar{x}$ if it is the sample mean of two observations from the uniform distribution shown. The dashed line represents the closest approximation of the normal distribution. Similarly, the center and right panels of the $n = 2$ row represent the respective distributions of $\bar{x}$ for data from exponential and log-normal distributions.

[32] Here the null hypothesis is that the part is not broken, and the alternative is that it is broken. If we don't have sufficient evidence to reject H_0, we would not replace the part. It sounds like failing to fix the part if it is broken (H_0 false, H_A true) is not very problematic, and replacing the part is expensive. Thus, we should require very strong evidence against H_0 before we replace the part. Choose a small significance level, such as $\alpha = 0.01$.