6.3 Testing for goodness of fit using chi-square

In this section, we develop a method for assessing a null model when the data are binned. This technique is commonly used in two circumstances:

- Given a sample of cases that can be classified into several groups, determine if the sample is representative of the general population.
- Evaluate whether data resemble a particular distribution, such as a normal distribution or a geometric distribution.

Each of these scenarios can be addressed using the same statistical test: a chi-square test.

In the first case, we consider data from a random sample of 275 jurors in a small county. Jurors identified their racial group, as shown in Figure 6.5, and we would like to determine if these jurors are racially representative of the population. If the jury is representative of the population, then the proportions in the sample should roughly reflect the population of eligible jurors, i.e. registered voters.

Race	White	Black	Hispanic	Other	Total
Representation in juries	205	26	25	19	275
Registered voters	0.72	0.07	0.12	0.09	1.00

Figure 6.5: Representation by race in a city's juries and population.

While the proportions in the juries do not precisely represent the population proportions, it is unclear whether these data provide convincing evidence that the sample is not representative. If the jurors really were randomly sampled from the registered voters, we might expect small differences due to chance. However, unusually large differences may provide convincing evidence that the juries were not representative.

A second application, assessing the fit of a distribution, is presented at the end of this section. Daily stock returns from the S&P500 for 25 years are used to assess whether stock activity each day is independent of the stock's behavior on previous days.

In these problems, we would like to examine all bins simultaneously, not simply compare one or two bins at a time, which will require us to develop a new test statistic.

6.3.1 Creating a test statistic for one-way tables

EXAMPLE 6.22

Of the people in the city, 275 served on a jury. If the individuals are randomly selected to serve on a jury, about how many of the 275 people would we expect to be white? How many would we expect to be black?

(E)

About 72% of the population is white, so we would expect about 72% of the jurors to be white: $0.72 \times 275 = 198$.

Similarly, we would expect about 7% of the jurors to be black, which would correspond to about $0.07 \times 275 = 19.25$ black jurors.

GUIDED PRACTICE 6.23

(G) Twelve percent of the population is Hispanic and 9% represent other races. How many of the 275 jurors would we expect to be Hispanic or from another race? Answers can be found in Figure 6.6.

The sample proportion represented from each race among the 275 jurors was not a precise match for any ethnic group. While some sampling variation is expected, we would expect the

Race	White	Black	Hispanic	Other	Total
Observed data	205	26	25	19	275
Expected counts	198	19.25	33	24.75	275

Figure 6.6: Actual and expected make-up of the jurors.

sample proportions to be fairly similar to the population proportions if there is no bias on juries. We need to test whether the differences are strong enough to provide convincing evidence that the jurors are not a random sample. These ideas can be organized into hypotheses:

H_0: The jurors are a random sample, i.e. there is no racial bias in who serves on a jury, and the observed counts reflect natural sampling fluctuation.

H_A: The jurors are not randomly sampled, i.e. there is racial bias in juror selection.

To evaluate these hypotheses, we quantify how different the observed counts are from the expected counts. Strong evidence for the alternative hypothesis would come in the form of unusually large deviations in the groups from what would be expected based on sampling variation alone.

6.3.2 The chi-square test statistic

In previous hypothesis tests, we constructed a test statistic of the following form:

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

This construction was based on (1) identifying the difference between a point estimate and an expected value if the null hypothesis was true, and (2) standardizing that difference using the standard error of the point estimate. These two ideas will help in the construction of an appropriate test statistic for count data.

Our strategy will be to first compute the difference between the observed counts and the counts we would expect if the null hypothesis was true, then we will standardize the difference:

$$Z_1 = \frac{\text{observed white count} - \text{null white count}}{\text{SE of observed white count}}$$

The standard error for the point estimate of the count in binned data is the square root of the count under the null.[32] Therefore:

$$Z_1 = \frac{205 - 198}{\sqrt{198}} = 0.50$$

The fraction is very similar to previous test statistics: first compute a difference, then standardize it. These computations should also be completed for the black, Hispanic, and other groups:

Black

$$Z_2 = \frac{26 - 19.25}{\sqrt{19.25}} = 1.54$$

Hispanic

$$Z_3 = \frac{25 - 33}{\sqrt{33}} = -1.39$$

Other

$$Z_4 = \frac{19 - 24.75}{\sqrt{24.75}} = -1.16$$

We would like to use a single test statistic to determine if these four standardized differences are irregularly far from zero. That is, Z_1, Z_2, Z_3, and Z_4 must be combined somehow to help determine if they – as a group – tend to be unusually far from zero. A first thought might be to take the absolute value of these four standardized differences and add them up:

$$|Z_1| + |Z_2| + |Z_3| + |Z_4| = 4.58$$

[32] Using some of the rules learned in earlier chapters, we might think that the standard error would be $np(1-p)$, where n is the sample size and p is the proportion in the population. This would be correct if we were looking only at one count. However, we are computing many standardized differences and adding them together. It can be shown – though not here – that the square root of the count is a better way to standardize the count differences.

Indeed, this does give one number summarizing how far the actual counts are from what was expected. However, it is more common to add the squared values:

$$Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 = 5.89$$

Squaring each standardized difference before adding them together does two things:

- Any standardized difference that is squared will now be positive.
- Differences that already look unusual – e.g. a standardized difference of 2.5 – will become much larger after being squared.

The test statistic X^2, which is the sum of the Z^2 values, is generally used for these reasons. We can also write an equation for X^2 using the observed counts and null counts:

$$X^2 = \frac{(\text{observed count}_1 - \text{null count}_1)^2}{\text{null count}_1} + \cdots + \frac{(\text{observed count}_4 - \text{null count}_4)^2}{\text{null count}_4}$$

The final number X^2 summarizes how strongly the observed counts tend to deviate from the null counts. In Section 6.3.4, we will see that if the null hypothesis is true, then X^2 follows a new distribution called a *chi-square distribution*. Using this distribution, we will be able to obtain a p-value to evaluate the hypotheses.

6.3.3 The chi-square distribution and finding areas

The **chi-square distribution** is sometimes used to characterize data sets and statistics that are always positive and typically right skewed. Recall a normal distribution had two parameters – mean and standard deviation – that could be used to describe its exact characteristics. The chi-square distribution has just one parameter called **degrees of freedom (df)**, which influences the shape, center, and spread of the distribution.

GUIDED PRACTICE 6.24

Figure 6.7 shows three chi-square distributions.
(a) How does the center of the distribution change when the degrees of freedom is larger?
(b) What about the variability (spread)?
(c) How does the shape change?[33]

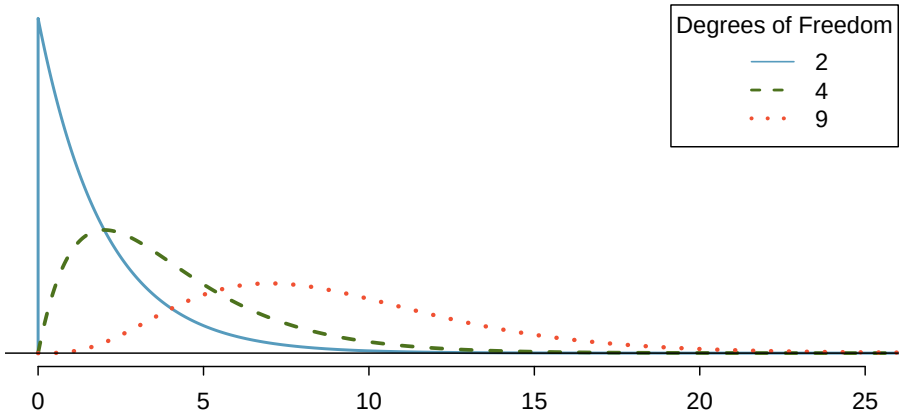


Figure 6.7: Three chi-square distributions with varying degrees of freedom.

[33] (a) The center becomes larger. If took a careful look, we could see that the mean of each distribution is equal to the distribution's degrees of freedom. (b) The variability increases as the degrees of freedom increases. (c) The distribution is very strongly skewed for $df = 2$, and then the distributions become more symmetric for the larger degrees of freedom $df = 4$ and $df = 9$. We would see this trend continue if we examined distributions with even more larger degrees of freedom.

Figure 6.7 and Guided Practice 6.24 demonstrate three general properties of chi-square distributions as the degrees of freedom increases: the distribution becomes more symmetric, the center moves to the right, and the variability inflates.

Our principal interest in the chi-square distribution is the calculation of p-values, which (as we have seen before) is related to finding the relevant area in the tail of a distribution. The most common ways to do this are using computer software, using a graphing calculator, or using a table. For folks wanting to use the table option, we provide an outline of how to read the chi-square table in Appendix C.3, which is also where you may find the table. For the examples below, use your preferred approach to confirm you get the same answers.

EXAMPLE 6.25

Figure 6.8(a) shows a chi-square distribution with 3 degrees of freedom and an upper shaded tail starting at 6.25. Find the shaded area.

(E)

Using statistical software or a graphing calculator, we can find that the upper tail area for a chi-square distribution with 3 degrees of freedom (df) and a cutoff of 6.25 is 0.1001. That is, the shaded upper tail of Figure 6.8(a) has area 0.1.

EXAMPLE 6.26

Figure 6.8(b) shows the upper tail of a chi-square distribution with 2 degrees of freedom. The bound for this upper tail is at 4.3. Find the tail area.

(E)

Using software, we can find that the tail area shaded in Figure 6.8(b) to be 0.1165. If using a table, we would only be able to find a range of values for the tail area: between 0.1 and 0.2.

EXAMPLE 6.27

Figure 6.8(c) shows an upper tail for a chi-square distribution with 5 degrees of freedom and a cutoff of 5.1. Find the tail area.

(E)

Using software, we would obtain a tail area of 0.4038. If using the table in Appendix C.3, we would have identified that the tail area is larger than 0.3 but not be able to give the precise value.

GUIDED PRACTICE 6.28

(G) Figure 6.8(d) shows a cutoff of 11.7 on a chi-square distribution with 7 degrees of freedom. Find the area of the upper tail.[34]

GUIDED PRACTICE 6.29

(G) Figure 6.8(e) shows a cutoff of 10 on a chi-square distribution with 4 degrees of freedom. Find the area of the upper tail.[35]

GUIDED PRACTICE 6.30

(G) Figure 6.8(f) shows a cutoff of 9.21 with a chi-square distribution with 3 df. Find the area of the upper tail.[36]

[34] The area is 0.1109. If using a table, we would identify that it falls between 0.1 and 0.2.
[35] Precise value: 0.0404. If using the table: between 0.02 and 0.05.
[36] Precise value: 0.0266. If using the table: between 0.02 and 0.05.

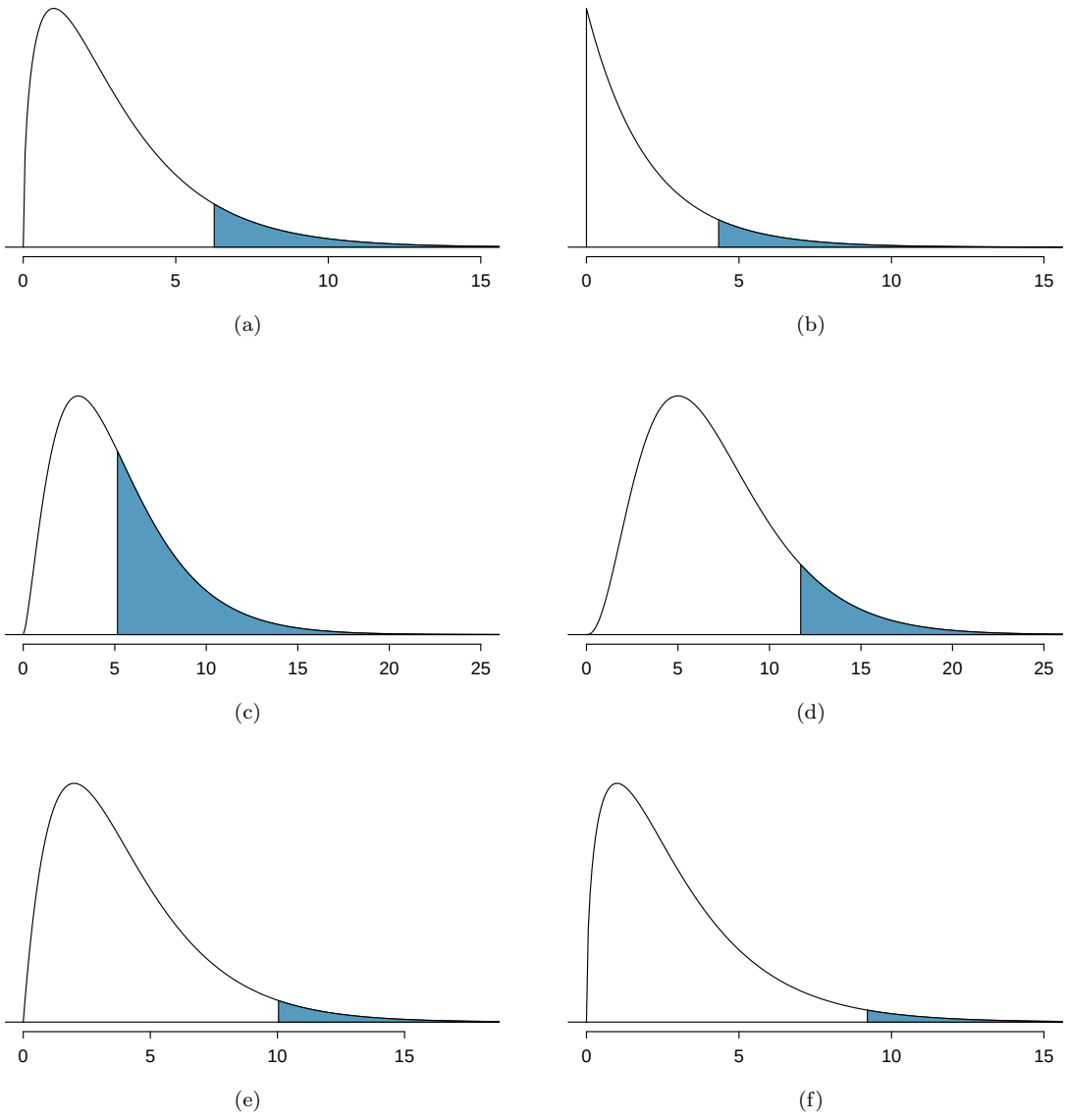


Figure 6.8: **(a)** Chi-square distribution with 3 degrees of freedom, area above 6.25 shaded. **(b)** 2 degrees of freedom, area above 4.3 shaded. **(c)** 5 degrees of freedom, area above 5.1 shaded. **(d)** 7 degrees of freedom, area above 11.7 shaded. **(e)** 4 degrees of freedom, area above 10 shaded. **(f)** 3 degrees of freedom, area above 9.21 shaded.

6.3.4 Finding a p-value for a chi-square distribution

In Section 6.3.2, we identified a new test statistic (X^2) within the context of assessing whether there was evidence of racial bias in how jurors were sampled. The null hypothesis represented the claim that jurors were randomly sampled and there was no racial bias. The alternative hypothesis was that there was racial bias in how the jurors were sampled.

We determined that a large X^2 value would suggest strong evidence favoring the alternative hypothesis: that there was racial bias. However, we could not quantify what the chance was of observing such a large test statistic ($X^2 = 5.89$) if the null hypothesis actually was true. This is where the chi-square distribution becomes useful. If the null hypothesis was true and there was no racial bias, then X^2 would follow a chi-square distribution, with three degrees of freedom in this case. Under certain conditions, the statistic X^2 follows a chi-square distribution with $k - 1$ degrees of freedom, where k is the number of bins.

EXAMPLE 6.31

How many categories were there in the juror example? How many degrees of freedom should be associated with the chi-square distribution used for X^2?

(E)

In the jurors example, there were $k = 4$ categories: white, black, Hispanic, and other. According to the rule above, the test statistic X^2 should then follow a chi-square distribution with $k - 1 = 3$ degrees of freedom if H_0 is true.

Just like we checked sample size conditions to use a normal distribution in earlier sections, we must also check a sample size condition to safely apply the chi-square distribution for X^2. Each expected count must be at least 5. In the juror example, the expected counts were 198, 19.25, 33, and 24.75, all easily above 5, so we can apply the chi-square model to the test statistic, $X^2 = 5.89$.

EXAMPLE 6.32

If the null hypothesis is true, the test statistic $X^2 = 5.89$ would be closely associated with a chi-square distribution with three degrees of freedom. Using this distribution and test statistic, identify the p-value.

(E)

The chi-square distribution and p-value are shown in Figure 6.9. Because larger chi-square values correspond to stronger evidence against the null hypothesis, we shade the upper tail to represent the p-value. Using statistical software (or the table in Appendix C.3), we can determine that the area is 0.1171. Generally we do not reject the null hypothesis with such a large p-value. In other words, the data do not provide convincing evidence of racial bias in the juror selection.

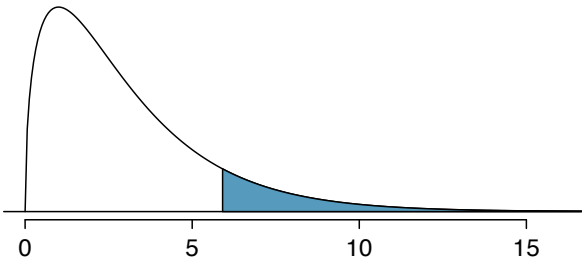


Figure 6.9: The p-value for the juror hypothesis test is shaded in the chi-square distribution with $df = 3$.

CHI-SQUARE TEST FOR ONE-WAY TABLE

Suppose we are to evaluate whether there is convincing evidence that a set of observed counts O_1, O_2, ..., O_k in k categories are unusually different from what might be expected under a null hypothesis. Call the *expected counts* that are based on the null hypothesis E_1, E_2, ..., E_k. If each expected count is at least 5 and the null hypothesis is true, then the test statistic below follows a chi-square distribution with $k - 1$ degrees of freedom:

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k}$$

The p-value for this test statistic is found by looking at the upper tail of this chi-square distribution. We consider the upper tail because larger values of X^2 would provide greater evidence against the null hypothesis.

CONDITIONS FOR THE CHI-SQUARE TEST

There are two conditions that must be checked before performing a chi-square test:

Independence. Each case that contributes a count to the table must be independent of all the other cases in the table.

Sample size / distribution. Each particular scenario (i.e. cell count) must have at least 5 expected cases.

Failing to check conditions may affect the test's error rates.

When examining a table with just two bins, pick a single bin and use the one-proportion methods introduced in Section 6.1.

6.3.5 Evaluating goodness of fit for a distribution

Section 4.2 would be useful background reading for this example, but it is not a prerequisite.

We can apply the chi-square testing framework to the second problem in this section: evaluating whether a certain statistical model fits a data set. Daily stock returns from the S&P500 for 10 can be used to assess whether stock activity each day is independent of the stock's behavior on previous days. This sounds like a very complex question, and it is, but a chi-square test can be used to study the problem. We will label each day as `Up` or `Down` (`D`) depending on whether the market was up or down that day. For example, consider the following changes in price, their new labels of up and down, and then the number of days that must be observed before each `Up` day:

Change in price	2.52	-1.46	0.51	-4.07	3.36	1.10	-5.46	-1.03	-2.99	1.71
Outcome	Up	D	Up	D	Up	Up	D	D	D	Up
Days to Up	1	-	2	-	2	1	-	-	-	4

If the days really are independent, then the number of days until a positive trading day should follow a geometric distribution. The geometric distribution describes the probability of waiting for the k^{th} trial to observe the first success. Here each up day (Up) represents a success, and down (D) days represent failures. In the data above, it took only one day until the market was up, so the first wait time was 1 day. It took two more days before we observed our next `Up` trading day, and two more for the third `Up` day. We would like to determine if these counts (1, 2, 2, 1, 4, and so on) follow the geometric distribution. Figure 6.10 shows the number of waiting days for a positive trading day during 10 years for the S&P500.

Days	1	2	3	4	5	6	7+	Total
Observed	717	369	155	69	28	14	10	1362

Figure 6.10: Observed distribution of the waiting time until a positive trading day for the S&P500.

We consider how many days one must wait until observing an `Up` day on the S&P500 stock index. If the stock activity was independent from one day to the next and the probability of a positive trading day was constant, then we would expect this waiting time to follow a *geometric distribution*. We can organize this into a hypothesis framework:

H_0: The stock market being up or down on a given day is independent from all other days. We will consider the number of days that pass until an `Up` day is observed. Under this hypothesis, the number of days until an `Up` day should follow a geometric distribution.

H_A: The stock market being up or down on a given day is not independent from all other days. Since we know the number of days until an `Up` day would follow a geometric distribution under the null, we look for deviations from the geometric distribution, which would support the alternative hypothesis.

There are important implications in our result for stock traders: if information from past trading days is useful in telling what will happen today, that information may provide an advantage over other traders.

We consider data for the S&P500 and summarize the waiting times in Figure 6.11 and Figure 6.12. The S&P500 was positive on 54.5% of those days.

Because applying the chi-square framework requires expected counts to be at least 5, we have *binned* together all the cases where the waiting time was at least 7 days to ensure each expected count is well above this minimum. The actual data, shown in the *Observed* row in Figure 6.11, can be compared to the expected counts from the *Geometric Model* row. The method for computing expected counts is discussed in Figure 6.11. In general, the expected counts are determined by (1) identifying the null proportion associated with each bin, then (2) multiplying each null proportion by the total count to obtain the expected counts. That is, this strategy identifies what proportion of the total count we would expect to be in each bin.

Days	1	2	3	4	5	6	7+	Total
Observed	717	369	155	69	28	14	10	1362
Geometric Model	743	338	154	70	32	14	12	1362

Figure 6.11: Distribution of the waiting time until a positive trading day. The expected counts based on the geometric model are shown in the last row. To find each expected count, we identify the probability of waiting D days based on the geometric model ($P(D) = (1 - 0.545)^{D-1}(0.545)$) and multiply by the total number of streaks, 1362. For example, waiting for three days occurs under the geometric model about $0.455^2 \times 0.545 = 11.28\%$ of the time, which corresponds to $0.1128 \times 1362 = 154$ streaks.

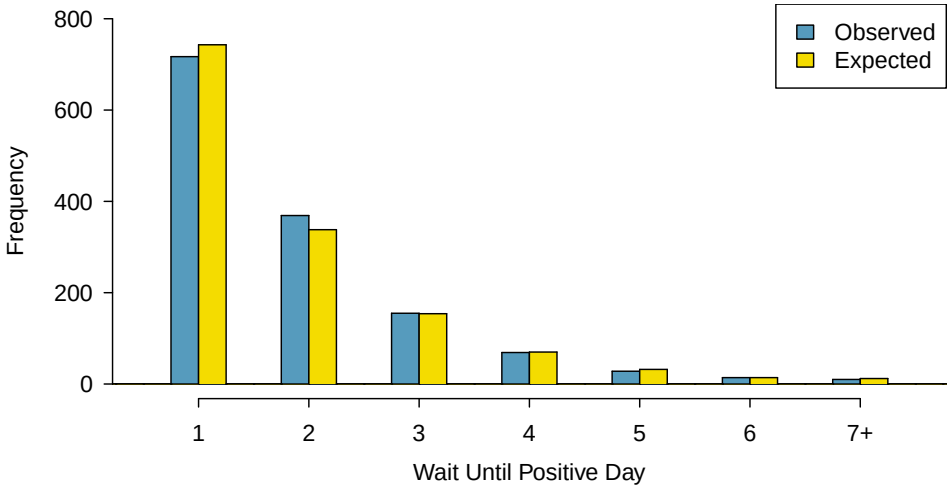


Figure 6.12: Side-by-side bar plot of the observed and expected counts for each waiting time.

EXAMPLE 6.33

Do you notice any unusually large deviations in the graph? Can you tell if these deviations are due to chance just by looking?

(E)

It is not obvious whether differences in the observed counts and the expected counts from the geometric distribution are significantly different. That is, it is not clear whether these deviations might be due to chance or whether they are so strong that the data provide convincing evidence against the null hypothesis. However, we can perform a chi-square test using the counts in Figure 6.11.

GUIDED PRACTICE 6.34

(G) Figure 6.11 provides a set of count data for waiting times ($O_1 = 717$, $O_2 = 369$, ...) and expected counts under the geometric distribution ($E_1 = 743$, $E_2 = 338$, ...). Compute the chi-square test statistic, X^2.[37]

GUIDED PRACTICE 6.35

(G) Because the expected counts are all at least 5, we can safely apply the chi-square distribution to X^2. However, how many degrees of freedom should we use?[38]

EXAMPLE 6.36

If the observed counts follow the geometric model, then the chi-square test statistic $X^2 = 4.61$ would closely follow a chi-square distribution with $df = 6$. Using this information, compute a p-value.

(E)

Figure 6.13 shows the chi-square distribution, cutoff, and the shaded p-value. Using software, we can find the p-value: 0.5951. Ultimately, we do not have sufficient evidence to reject the notion that the wait times follow a geometric distribution for the last 10 years of data for the S&P500, i.e. we cannot reject the notion that trading days are independent.

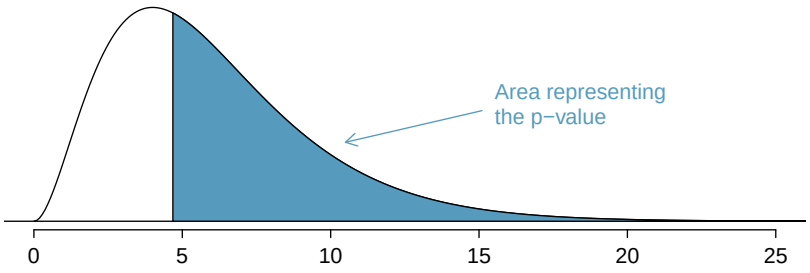


Figure 6.13: Chi-square distribution with 6 degrees of freedom. The p-value for the stock analysis is shaded.

EXAMPLE 6.37

In Example 6.36, we did not reject the null hypothesis that the trading days are independent during the last 10 of data. Why is this so important?

(E)

It may be tempting to think the market is "due" for an `Up` day if there have been several consecutive days where it has been down. However, we haven't found strong evidence that there's any such property where the market is "due" for a correction. At the very least, the analysis suggests any dependence between days is very weak.

[37] $X^2 = \frac{(717-743)^2}{743} + \frac{(369-338)^2}{338} + \cdots + \frac{(10-12)^2}{12} = 4.61$

[38] There are $k = 7$ groups, so we use $df = k - 1 = 6$.

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:

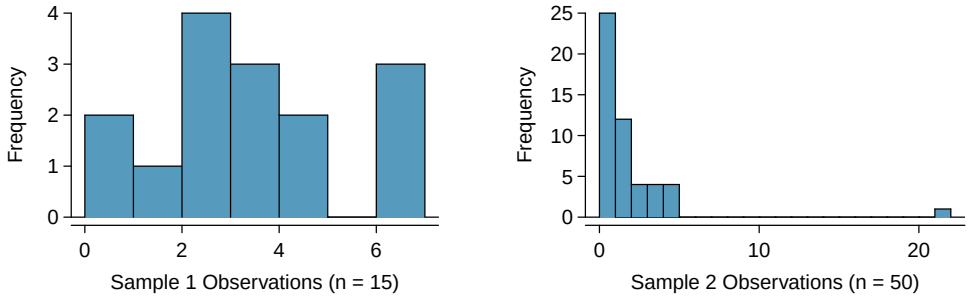
n < 30: If the sample size n is less than 30 and there are no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.

n ≥ 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 $n_2 = 50$.



E

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 on the right, 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.

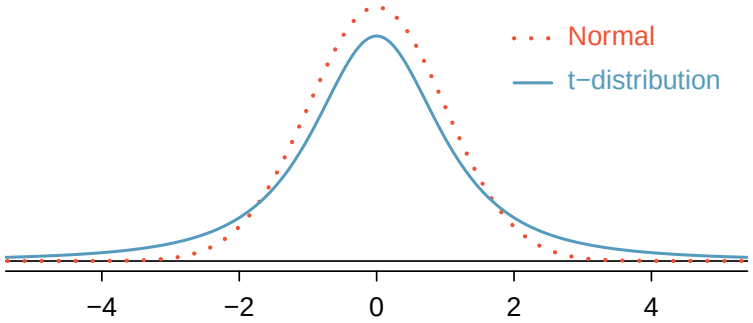


Figure 7.1: Comparison of a t-distribution and a normal distribution.

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.

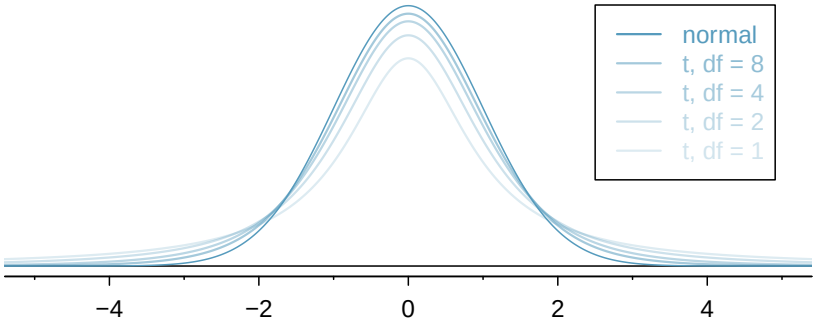


Figure 7.2: The larger the degrees of freedom, the more closely the t-distribution resembles the standard normal distribution.

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.

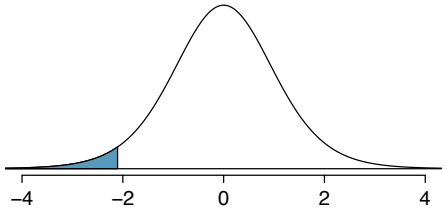


Figure 7.3: The t-distribution with 18 degrees of freedom. The area below -2.10 has been shaded.

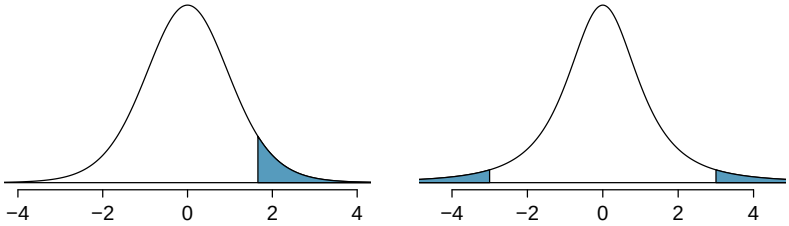


Figure 7.4: Left: The t-distribution with 20 degrees of freedom, with the area above 1.65 shaded. Right: The t-distribution with 2 degrees of freedom, with the area further than 3 units from 0 shaded.

EXAMPLE 7.2

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

(E)

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.

(E)

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).

(E)

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

(G) 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?

(E)

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.

(E)

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%

(E)

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.

(E)

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

(G) 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 $t^\star_{df}$.

(E)

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 finding 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).

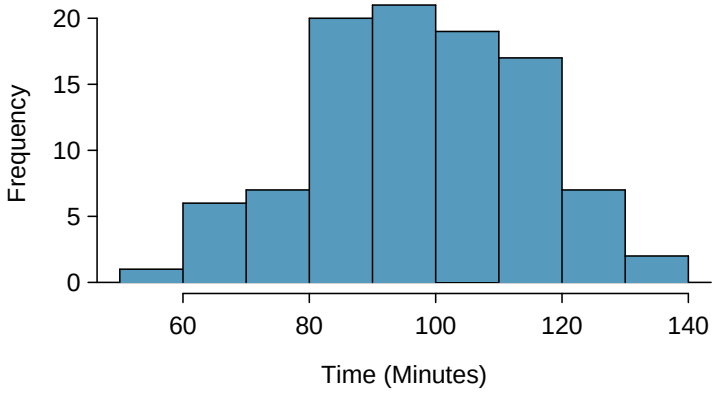


Figure 7.7: A histogram of `time` for the sample Cherry Blossom Race data.

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:

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

(E)

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.

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.

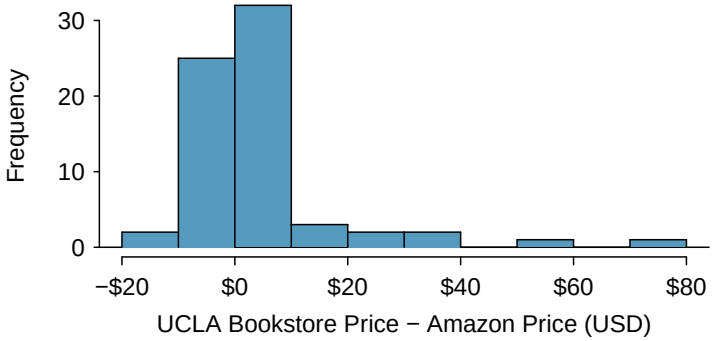


Figure 7.9: Histogram of the difference in price for each book sampled.

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.

(E)

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 ($n_{diff} = 68$):

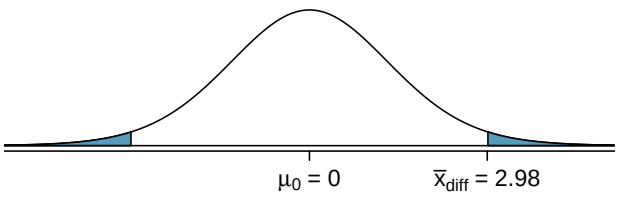
$$SE_{\bar{x}_{diff}} = \frac{s_{diff}}{\sqrt{n_{diff}}} = \frac{13.42}{\sqrt{68}} = 1.63$$

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

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

(E)

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.

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]

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:

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

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.

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

(E) 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.

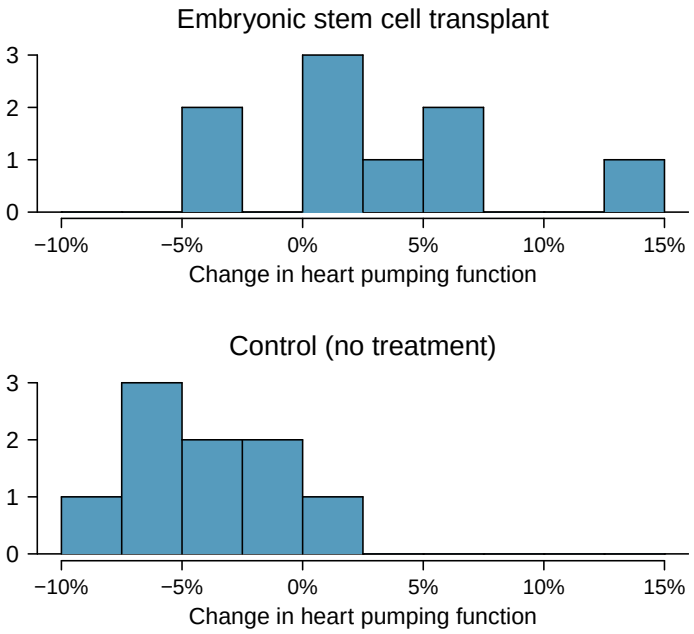


Figure 7.12: Histograms for both the embryonic stem cell and control group.

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:

(E)

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

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.

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

E

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.

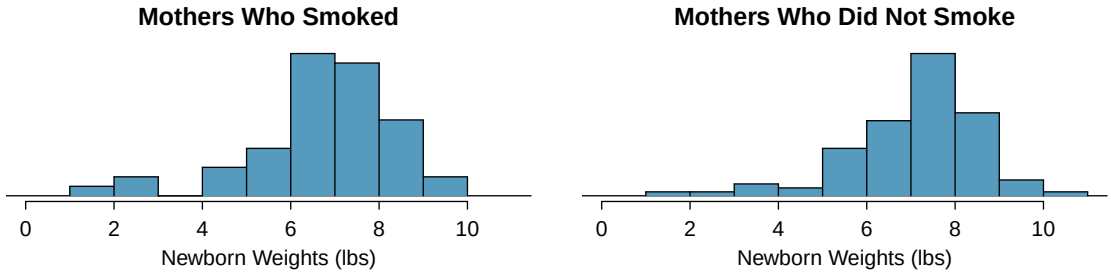


Figure 7.14: The top panel represents birth weights for infants whose mothers smoked. The bottom panel represents the birth weights for infants whose mothers who did not smoke.

GUIDED PRACTICE 7.24

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

G

(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:

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

EXAMPLE 7.25

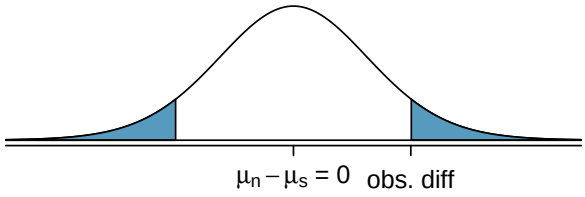
Complete the hypothesis test started in Example 7.23 and Guided Practice 7.24. Use a significance level of $\alpha = 0.05$. For reference, $\bar{x}_n - \bar{x}_s = 0.40$, $SE = 0.26$, and the sample sizes were $n_n = 100$ and $n_s = 50$.

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:

E



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.

GUIDED PRACTICE 7.26

G 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]

GUIDED PRACTICE 7.27

G 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

G 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

G 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.

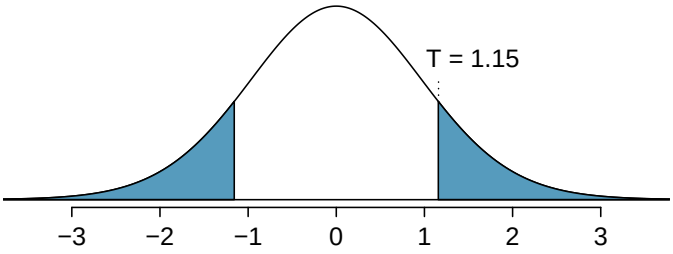


Figure 7.17: The t-distribution with 26 degrees of freedom and the p-value from exam example represented as the shaded areas.

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.

(E) 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:

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

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

$$df = n_1 + n_2 - 2$$

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

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.*

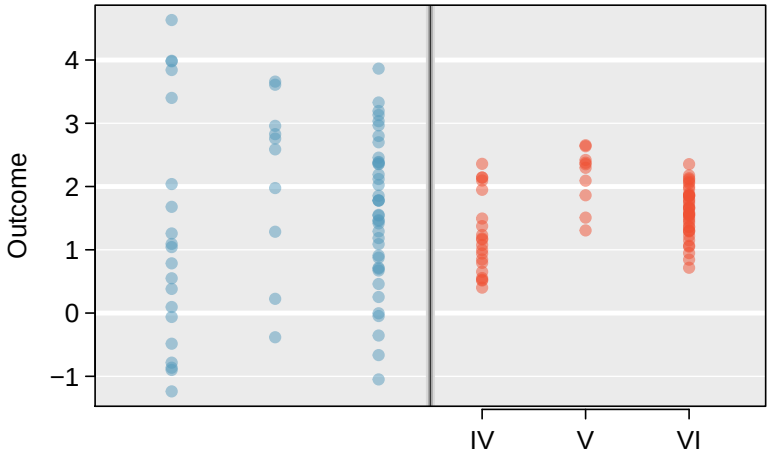


Figure 7.19: Side-by-side dot plot for the outcomes for six groups.

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

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

EXAMPLE 7.43

E The player positions have been divided into four groups: outfield (OF), infield (IF), and catcher (C). What would be an appropriate point estimate of the on-base percentage by outfielders, μ_{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.

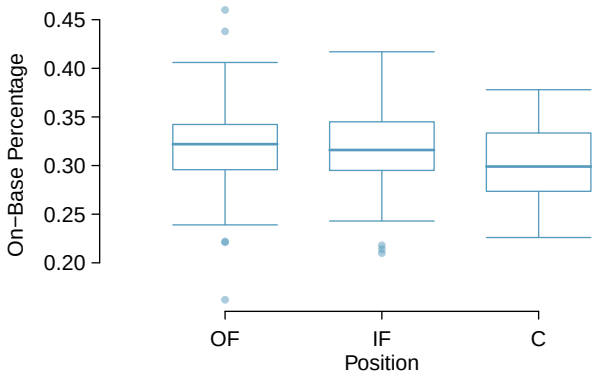


Figure 7.23: Side-by-side box plot of the on-base percentage for 429 players across four groups. There is one prominent outlier visible in the infield group, but with 154 observations in the infield group, this outlier is not a concern.

[30] H_0: The average on-base percentage is equal across the four 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?

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, andrewgelman.com/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

(G) 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.)

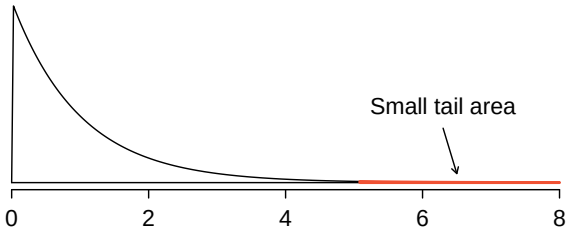


Figure 7.24: An F distribution with $df_1 = 3$ and $df_2 = 323$.

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?

(E)

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.

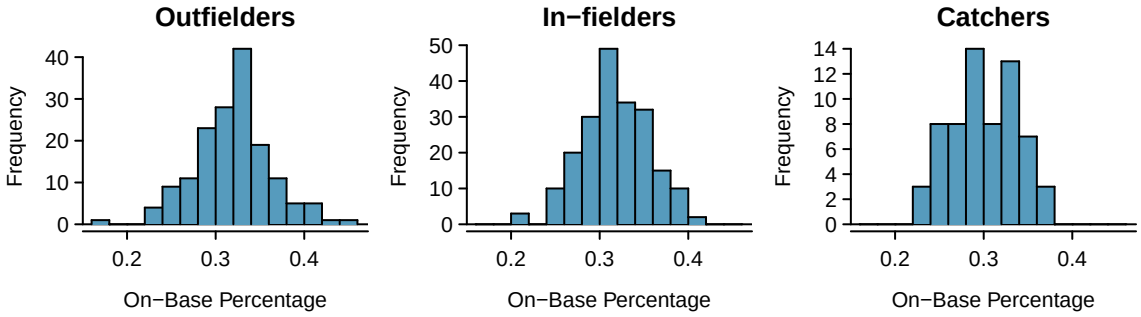


Figure 7.26: Histograms of OBP for each field position.

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 four 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.

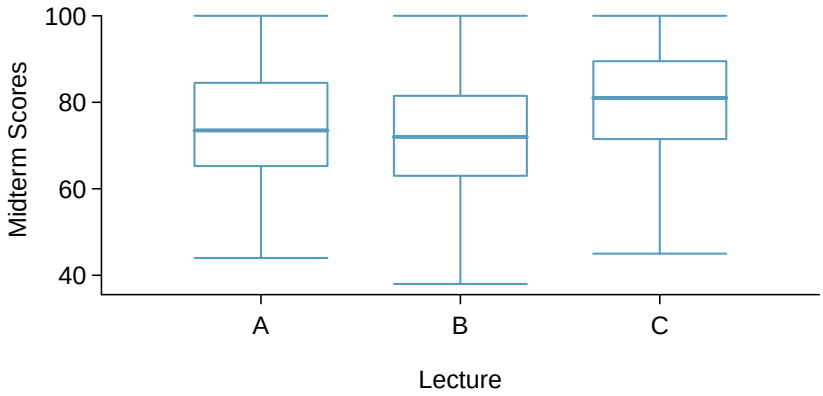


Figure 7.28: Side-by-side box plot for the first midterm scores in three different lectures of the same course.

GUIDED PRACTICE 7.48

G 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,

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

E (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:

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

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.

9.1 Introduction to multiple regression

Multiple regression extends simple two-variable regression to the case that still has one response but many predictors (denoted x_1, x_2, x_3, ...). The method is motivated by scenarios where many variables may be simultaneously connected to an output.

We will consider data about loans from the peer-to-peer lender, Lending Club, which is a data set we first encountered in Chapters 1 and 2. The loan data includes terms of the loan as well as information about the borrower. The outcome variable we would like to better understand is the interest rate assigned to the loan. For instance, all other characteristics held constant, does it matter how much debt someone already has? Does it matter if their income has been verified? Multiple regression will help us answer these and other questions.

The data set `loans` includes results from 10,000 loans, and we'll be looking at a subset of the available variables, some of which will be new from those we saw in earlier chapters. The first six observations in the data set are shown in Figure 9.1, and descriptions for each variable are shown in Figure 9.2. Notice that the past bankruptcy variable (`bankruptcy`) is an indicator variable, where it takes the value 1 if the borrower had a past bankruptcy in their record and 0 if not. Using an indicator variable in place of a category name allows for these variables to be directly used in regression. Two of the other variables are categorical (`income_ver` and `issued`), each of which can take one of a few different non-numerical values; we'll discuss how these are handled in the model in Section 9.1.1.

	interest_rate	income_ver	debt_to_income	credit_util	bankruptcy	term	issued	credit_checks
1	14.07	verified	18.01	0.55	0	60	Mar2018	6
2	12.61	not	5.04	0.15	1	36	Feb2018	1
3	17.09	source_only	21.15	0.66	0	36	Feb2018	4
4	6.72	not	10.16	0.20	0	36	Jan2018	0
5	14.07	verified	57.96	0.75	0	36	Mar2018	7
6	6.72	not	6.46	0.09	0	36	Jan2018	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 9.1: First six rows from the `loans` data set.

variable	description
`interest_rate`	Interest rate for the loan.
`income_ver`	Categorical variable describing whether the borrower's income source and amount have been verified, with levels `verified`, `source_only`, and `not`.
`debt_to_income`	Debt-to-income ratio, which is the percentage of total debt of the borrower divided by their total income.
`credit_util`	Of all the credit available to the borrower, what fraction are they utilizing. For example, the credit utilization on a credit card would be the card's balance divided by the card's credit limit.
`bankruptcy`	An indicator variable for whether the borrower has a past bankruptcy in her record. This variable takes a value of `1` if the answer is "yes" and `0` if the answer is "no".
`term`	The length of the loan, in months.
`issued`	The month and year the loan was issued, which for these loans is always during the first quarter of 2018.
`credit_checks`	Number of credit checks in the last 12 months. For example, when filing an application for a credit card, it is common for the company receiving the application to run a credit check.

Figure 9.2: Variables and their descriptions for the `loans` data set.

9.1.1 Indicator and categorical variables as predictors

Let's start by fitting a linear regression model for interest rate with a single predictor indicating whether or not a person has a bankruptcy in their record:

$$\widehat{rate} = 12.33 + 0.74 \times bankruptcy$$

Results of this model are shown in Figure 9.3.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.3380	0.0533	231.49	<0.0001
bankruptcy	0.7368	0.1529	4.82	<0.0001
				$df = 9998$

Figure 9.3: Summary of a linear model for predicting interest rate based on whether the borrower has a bankruptcy in their record.

EXAMPLE 9.1

Interpret the coefficient for the past bankruptcy variable in the model. Is this coefficient significantly different from 0?

(E) The `bankruptcy` variable takes one of two values: 1 when the borrower has a bankruptcy in their history and 0 otherwise. A slope of 0.74 means that the model predicts a 0.74% higher interest rate for those borrowers with a bankruptcy in their record. (See Section 8.2.8 for a review of the interpretation for two-level categorical predictor variables.) Examining the regression output in Figure 9.3, we can see that the p-value for `bankruptcy` is very close to zero, indicating there is strong evidence the coefficient is different from zero when using this simple one-predictor model.

Suppose we had fit a model using a 3-level categorical variable, such as `income_ver`. The output from software is shown in Figure 9.4. This regression output provides multiple rows for the `income_ver` variable. Each row represents the relative difference for each level of `income_ver`. However, we are missing one of the levels: `not` (for *not verified*). The missing level is called the **reference level**, and it represents the default level that other levels are measured against.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.0995	0.0809	137.18	<0.0001
income_ver: *source_only*	1.4160	0.1107	12.79	<0.0001
income_ver: *verified*	3.2543	0.1297	25.09	<0.0001
				$df = 9998$

Figure 9.4: Summary of a linear model for predicting interest rate based on whether the borrower's income source and amount has been verified. This predictor has three levels, which results in 2 rows in the regression output.

EXAMPLE 9.2

How would we write an equation for this regression model?

(E) The equation for the regression model may be written as a model with two predictors:

$$\widehat{rate} = 11.10 + 1.42 \times \texttt{income\_ver}_{\texttt{source\_only}} + 3.25 \times \texttt{income\_ver}_{\texttt{verified}}$$

We use the notation $\texttt{variable}_{\texttt{level}}$ to represent indicator variables for when the categorical variable takes a particular value. For example, $\texttt{income\_ver}_{\texttt{source\_only}}$ would take a value of 1 if `income_ver` was `source_only` for a loan, and it would take a value of 0 otherwise. Likewise, $\texttt{income\_ver}_{\texttt{verified}}$ would take a value of 1 if `income_ver` took a value of `verified` and 0 if it took any other value.

The notation used in Example 9.2 may feel a bit confusing. Let's figure out how to use the equation for each level of the `income_ver` variable.

EXAMPLE 9.3

Using the model from Example 9.2, compute the average interest rate for borrowers whose income source and amount are both unverified.

When `income_ver` takes a value of `not`, then both indicator functions in the equation from Example 9.2 are set to zero:

E

$$\begin{aligned}\widehat{rate} &= 11.10 + 1.42 \times 0 + 3.25 \times 0 \\ &= 11.10\end{aligned}$$

The average interest rate for these borrowers is 11.1%. Because the `not` level does not have its own coefficient and it is the reference value, the indicators for the other levels for this variable all drop out.

EXAMPLE 9.4

Using the model from Example 9.2, compute the average interest rate for borrowers whose income source is verified but the amount is not.

When `income_ver` takes a value of `source_only`, then the corresponding variable takes a value of 1 while the other ($\texttt{income\_ver}_{\texttt{verified}}$) is 0:

E

$$\begin{aligned}\widehat{rate} &= 11.10 + 1.42 \times 1 + 3.25 \times 0 \\ &= 12.52\end{aligned}$$

The average interest rate for these borrowers is 12.52%.

GUIDED PRACTICE 9.5

G

Compute the average interest rate for borrowers whose income source and amount are both verified.[1]

PREDICTORS WITH SEVERAL CATEGORIES

When fitting a regression model with a categorical variable that has k levels where $k > 2$, software will provide a coefficient for $k - 1$ of those levels. For the last level that does not receive a coefficient, this is the **reference level**, and the coefficients listed for the other levels are all considered relative to this reference level.

[1] When `income_ver` takes a value of `verified`, then the corresponding variable takes a value of 1 while the other ($\texttt{income\_ver}_{\texttt{source\_only}}$) is 0:

$$\begin{aligned}\widehat{rate} &= 11.10 + 1.42 \times 0 + 3.25 \times 1 \\ &= 14.35\end{aligned}$$

The average interest rate for these borrowers is 14.35%.

GUIDED PRACTICE 9.6

Interpret the coefficients in the `income_ver` model.[2]

The higher interest rate for borrowers who have verified their income source or amount is surprising. Intuitively, we'd think that a loan would look *less* risky if the borrower's income has been verified. However, note that the situation may be more complex, and there may be confounding variables that we didn't account for. For example, perhaps lender require borrowers with poor credit to verify their income. That is, verifying income in our data set might be a signal of some concerns about the borrower rather than a reassurance that the borrower will pay back the loan. For this reason, the borrower could be deemed higher risk, resulting in a higher interest rate. (What other confounding variables might explain this counter-intuitive relationship suggested by the model?)

GUIDED PRACTICE 9.7

How much larger of an interest rate would we expect for a borrower who has verified their income source and amount vs a borrower whose income source has only been verified?[3]

9.1.2 Including and assessing many variables in a model

The world is complex, and it can be helpful to consider many factors at once in statistical modeling. For example, we might like to use the full context of borrower to predict the interest rate they receive rather than using a single variable. This is the strategy used in **multiple regression**. While we remain cautious about making any causal interpretations using multiple regression on observational data, such models are a common first step in gaining insights or providing some evidence of a causal connection.

We want to construct a model that accounts for not only for any past bankruptcy or whether the borrower had their income source or amount verified, but simultaneously accounts for all the variables in the data set: `income_ver`, `debt_to_income`, `credit_util`, `bankruptcy`, `term`, `issued`, and `credit_checks`.

$$\begin{aligned}\widehat{\texttt{rate}} = \beta_0 &+ \beta_1 \times \texttt{income\_ver}_{\texttt{source\_only}} + \beta_2 \times \texttt{income\_ver}_{\texttt{verified}} + \beta_3 \times \texttt{debt\_to\_income} \\ &+ \beta_4 \times \texttt{credit\_util} + \beta_5 \times \texttt{bankruptcy} + \beta_6 \times \texttt{term} \\ &+ \beta_7 \times \texttt{issued}_{\texttt{Jan2018}} + \beta_8 \times \texttt{issued}_{\texttt{Mar2018}} + \beta_9 \times \texttt{credit\_checks}\end{aligned}$$

This equation represents a holistic approach for modeling all of the variables simultaneously. Notice that there are two coefficients for `income_ver` and also two coefficients for `issued`, since both are 3-level categorical variables.

We estimate the parameters β_0, β_1, β_2, ..., β_9 in the same way as we did in the case of a single predictor. We select b_0, b_1, b_2, ..., b_9 that minimize the sum of the squared residuals:

$$SSE = e_1^2 + e_2^2 + \cdots + e_{10000}^2 = \sum_{i=1}^{10000} e_i^2 = \sum_{i=1}^{10000} (y_i - \hat{y}_i)^2 \tag{9.8}$$

where y_i and $\hat{y}_i$ represent the observed interest rates and their estimated values according to the model, respectively. 10,000 residuals are calculated, one for each observation. We typically use a computer to minimize the sum of squares and compute point estimates, as shown in the sample output in Figure 9.5. Using this output, we identify the point estimates b_i of each β_i, just as we did in the one-predictor case.

[2] Each of the coefficients gives the incremental interest rate for the corresponding level relative to the `not` level, which is the reference level. For example, for a borrower whose income source and amount have been verified, the model predicts that they will have a 3.25% higher interest rate than a borrower who has not had their income source or amount verified.

[3] Relative to the `not` category, the `verified` category has an interest rate of 3.25% higher, while the `source_only` category is only 1.42% higher. Thus, `verified` borrowers will tend to get an interest rate about 3.25%−1.42% = 1.83% higher than `source_only` borrowers.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9251	0.2102	9.16	<0.0001
income_ver: *source_only*	0.9750	0.0991	9.83	<0.0001
income_ver: *verified*	2.5374	0.1172	21.65	<0.0001
debt_to_income	0.0211	0.0029	7.18	<0.0001
credit_util	4.8959	0.1619	30.24	<0.0001
bankruptcy	0.3864	0.1324	2.92	0.0035
term	0.1537	0.0039	38.96	<0.0001
issued: *Jan2018*	0.0276	0.1081	0.26	0.7981
issued: *Mar2018*	-0.0397	0.1065	-0.37	0.7093
credit_checks	0.2282	0.0182	12.51	<0.0001
				$df = 9990$

Figure 9.5: Output for the regression model, where `interest_rate` is the outcome and the variables listed are the predictors.

MULTIPLE REGRESSION MODEL

A multiple regression model is a linear model with many predictors. In general, we write the model as

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

when there are k predictors. We always estimate the β_i parameters using statistical software.

EXAMPLE 9.9

Write out the regression model using the point estimates from Figure 9.5. How many predictors are there in this model?

The fitted model for the interest rate is given by:

$$\widehat{\texttt{rate}} = 1.925 + 0.975 \times \texttt{income\_ver}_{\texttt{source\_only}} + 2.537 \times \texttt{income\_ver}_{\texttt{verified}} + 0.021 \times \texttt{debt\_to\_income} + 4.896 \times \texttt{credit\_util} + 0.386 \times \texttt{bankruptcy} + 0.154 \times \texttt{term} + 0.028 \times \texttt{issued}_{\texttt{Jan2018}} - 0.040 \times \texttt{issued}_{\texttt{Mar2018}} + 0.228 \times \texttt{credit\_checks}$$

If we count up the number of predictor coefficients, we get the *effective* number of predictors in the model: $k = 9$. Notice that the `issued` categorical predictor counts as two, once for the two levels shown in the model. In general, a categorical predictor with p different levels will be represented by $p - 1$ terms in a multiple regression model.

GUIDED PRACTICE 9.10

What does β_4, the coefficient of variable `credit_util`, represent? What is the point estimate of β_4?[4]

[4] β_4 represents the change in interest rate we would expect if someone's credit utilization was 0 and went to 1, all other factors held even. The point estimate is $b_4 = 4.90\%$.

EXAMPLE 9.11

Compute the residual of the first observation in Figure 9.1 on page 343 using the equation identified in Guided Practice 9.9.

(E) To compute the residual, we first need the predicted value, which we compute by plugging values into the equation from Example 9.9. For example, $\texttt{income\_ver}_{\texttt{source\_only}}$ takes a value of 0, $\texttt{income\_ver}_{\texttt{verified}}$ takes a value of 1 (since the borrower's income source and amount were verified), `debt_to_income` was 18.01, and so on. This leads to a prediction of $\widehat{rate}_1 = 18.09$. The observed interest rate was 14.07%, which leads to a residual of $e_1 = 14.07 - 18.09 = -4.02$.

EXAMPLE 9.12

We estimated a coefficient for `bankruptcy` in Section 9.1.1 of $b_4 = 0.74$ with a standard error of $SE_{b_1} = 0.15$ when using simple linear regression. Why is there a difference between that estimate and the estimated coefficient of 0.39 in the multiple regression setting?

(E) If we examined the data carefully, we would see that some predictors are correlated. For instance, when we estimated the connection of the outcome `interest_rate` and predictor `bankruptcy` using simple linear regression, we were unable to control for other variables like whether the borrower had her income verified, the borrower's debt-to-income ratio, and other variables. That original model was constructed in a vacuum and did not consider the full context. When we include all of the variables, underlying and unintentional bias that was missed by these other variables is reduced or eliminated. Of course, bias can still exist from other confounding variables.

Example 9.12 describes a common issue in multiple regression: correlation among predictor variables. We say the two predictor variables are **collinear** (pronounced as *co-linear*) when they are correlated, and this collinearity complicates model estimation. While it is impossible to prevent collinearity from arising in observational data, experiments are usually designed to prevent predictors from being collinear.

GUIDED PRACTICE 9.13

(G) The estimated value of the intercept is 1.925, and one might be tempted to make some interpretation of this coefficient, such as, it is the model's predicted price when each of the variables take value zero: income source is not verified, the borrower has no debt (debt-to-income and credit utilization are zero), and so on. Is this reasonable? Is there any value gained by making this interpretation?[5]

[5] Many of the variables do take a value 0 for at least one data point, and for those variables, it is reasonable. However, one variable never takes a value of zero: `term`, which describes the length of the loan, in months. If `term` is set to zero, then the loan must be paid back immediately; the borrower must give the money back as soon as she receives it, which means it is not a real loan. Ultimately, the interpretation of the intercept in this setting is not insightful.

9.1.3 Adjusted R^2 as a better tool for multiple regression

We first used R^2 in Section 8.2 to determine the amount of variability in the response that was explained by the model:

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in the outcome}} = 1 - \frac{Var(e_i)}{Var(y_i)}$$

where e_i represents the residuals of the model and y_i the outcomes. This equation remains valid in the multiple regression framework, but a small enhancement can make it even more informative when comparing models.

GUIDED PRACTICE 9.14

G The variance of the residuals for the model given in Guided Practice 9.9 is 18.53, and the variance of the total price in all the auctions is 25.01. Calculate R^2 for this model.[6]

This strategy for estimating R^2 is acceptable when there is just a single variable. However, it becomes less helpful when there are many variables. The regular R^2 is a biased estimate of the amount of variability explained by the model when applied to a new sample of data. To get a better estimate, we use the adjusted R^2.

ADJUSTED R^2 AS A TOOL FOR MODEL ASSESSMENT

The **adjusted R^2** is computed as

$$R^2_{adj} = 1 - \frac{s^2_{\text{residuals}}/(n-k-1)}{s^2_{\text{outcome}}/(n-1)} = 1 - \frac{s^2_{\text{residuals}}}{s^2_{\text{outcome}}} \times \frac{n-1}{n-k-1}$$

where n is the number of cases used to fit the model and k is the number of predictor variables in the model. Remember that a categorical predictor with p levels will contribute $p-1$ to the number of variables in the model.

Because k is never negative, the adjusted R^2 will be smaller – often times just a little smaller – than the unadjusted R^2. The reasoning behind the adjusted R^2 lies in the **degrees of freedom** associated with each variance, which is equal to $n-k-1$ for the multiple regression context. If we were to make predictions for *new data* using our current model, we would find that the unadjusted R^2 would tend to be slightly overly optimistic, while the adjusted R^2 formula helps correct this bias.

GUIDED PRACTICE 9.15

G There were $n = 10000$ auctions in the `loans` data set and $k = 9$ predictor variables in the model. Use n, k, and the variances from Guided Practice 9.14 to calculate R^2_{adj} for the interest rate model.[7]

GUIDED PRACTICE 9.16

G Suppose you added another predictor to the model, but the variance of the errors $Var(e_i)$ didn't go down. What would happen to the R^2? What would happen to the adjusted R^2?[8]

Adjusted R^2 could have been used in Chapter 8. However, when there is only $k = 1$ predictors, adjusted R^2 is very close to regular R^2, so this nuance isn't typically important when the model has only one predictor.

[6] $R^2 = 1 - \frac{18.53}{25.01} = 0.2591$.

[7] $R^2_{adj} = 1 - \frac{18.53}{25.01} \times \frac{10000-1}{1000-9-1} = 0.2584$. While the difference is very small, it will be important when we fine tune the model in the next section.

[8] The unadjusted R^2 would stay the same and the adjusted R^2 would go down.

9.5 Introduction to logistic regression

In this section we introduce **logistic regression** as a tool for building models when there is a categorical response variable with two levels, e.g. yes and no. Logistic regression is a type of **generalized linear model** (**GLM**) for response variables where regular multiple regression does not work very well. In particular, the response variable in these settings often takes a form where residuals look completely different from the normal distribution.

GLMs can be thought of as a two-stage modeling approach. We first model the response variable using a probability distribution, such as the binomial or Poisson distribution. Second, we model the parameter of the distribution using a collection of predictors and a special form of multiple regression. Ultimately, the application of a GLM will feel very similar to multiple regression, even if some of the details are different.

9.5.1 Resume data

We will consider experiment data from a study that sought to understand the effect of race and sex on job application callback rates; details of the study and a link to the data set may be found in Appendix B.9. To evaluate which factors were important, job postings were identified in Boston and Chicago for the study, and researchers created many fake resumes to send off to these jobs to see which would elicit a callback. The researchers enumerated important characteristics, such as years of experience and education details, and they used these characteristics to randomly generate the resumes. Finally, they randomly assigned a name to each resume, where the name would imply the applicant's sex and race.

The first names that were used and randomly assigned in this experiment were selected so that they would predominantly be recognized as belonging to Black or White individuals; other races were not considered in this study. While no name would definitively be inferred as pertaining to a Black individual or to a White individual, the researchers conducted a survey to check for racial association of the names; names that did not pass this survey check were excluded from usage in the experiment. You can find the full set of names that did pass the survey test and were ultimately used in the study in Figure 9.20. For example, Lakisha was a name that their survey indicated would be interpreted as a Black woman, while Greg was a name that would generally be interpreted to be associated with a White male.

first_name	race	sex
Aisha	black	female
Allison	white	female
Anne	white	female
Brad	white	male
Brendan	white	male
Brett	white	male
Carrie	white	female
Darnell	black	male
Ebony	black	female
Emily	white	female
Geoffrey	white	male
Greg	white	male

first_name	race	sex
Hakim	black	male
Jamal	black	male
Jay	white	male
Jermaine	black	male
Jill	white	female
Kareem	black	male
Keisha	black	female
Kenya	black	female
Kristen	white	female
Lakisha	black	female
Latonya	black	female
Latoya	black	female

first_name	race	sex
Laurie	white	female
Leroy	black	male
Matthew	white	male
Meredith	white	female
Neil	white	male
Rasheed	black	male
Sarah	white	female
Tamika	black	female
Tanisha	black	female
Todd	white	male
Tremayne	black	male
Tyrone	black	male

Figure 9.20: List of all 36 unique names along with the commonly inferred race and sex associated with these names.

The response variable of interest is whether or not there was a callback from the employer for the applicant, and there were 8 attributes that were randomly assigned that we'll consider, with special interest in the race and sex variables. Race and sex are **protected classes** in the United States, meaning they are not legally permitted factors for hiring or employment decisions. The full set of attributes considered is provided in Figure 9.21.

variable	description
`callback`	Specifies whether the employer called the applicant following submission of the application for the job.
`job_city`	City where the job was located: Boston or Chicago.
`college_degree`	An indicator for whether the resume listed a college degree.
`years_experience`	Number of years of experience listed on the resume.
`honors`	Indicator for the resume listing some sort of honors, e.g. employee of the month.
`military`	Indicator for if the resume listed any military experience.
`email_address`	Indicator for if the resume listed an email address for the applicant.
`race`	Race of the applicant, implied by their first name listed on the resume.
`sex`	Sex of the applicant (limited to only `male` and `female` in this study), implied by the first name listed on the resume.

Figure 9.21: Descriptions for the `callback` variable along with 8 other variables in the `resume` data set. Many of the variables are indicator variables, meaning they take the value 1 if the specified characteristic is present and 0 otherwise.

All of the attributes listed on each resume were randomly assigned. This means that no attributes that might be favorable or detrimental to employment would favor one demographic over another on these resumes. Importantly, due to the experimental nature of this study, we can infer causation between these variables and the callback rate, if the variable is statistically significant. Our analysis will allow us to compare the practical importance of each of the variables relative to each other.

9.5.2 Modeling the probability of an event

Logistic regression is a generalized linear model where the outcome is a two-level categorical variable. The outcome, Y_i, takes the value 1 (in our application, this represents a callback for the resume) with probability p_i and the value 0 with probability $1 - p_i$. Because each observation has a slightly different context, e.g. different education level or a different number of years of experience, the probability p_i will differ for each observation. Ultimately, it is this probability that we model in relation to the predictor variables: we will examine which resume characteristics correspond to higher or lower callback rates.

NOTATION FOR A LOGISTIC REGRESSION MODEL

The outcome variable for a GLM is denoted by Y_i, where the index i is used to represent observation i. In the resume application, Y_i will be used to represent whether resume i received a callback ($Y_i = 1$) or not ($Y_i = 0$).

The predictor variables are represented as follows: $x_{1,i}$ is the value of variable 1 for observation i, $x_{2,i}$ is the value of variable 2 for observation i, and so on.

The logistic regression model relates the probability a resume would receive a callback (p_i) to the predictors $x_{1,i}$, $x_{2,i}$, ..., $x_{k,i}$ through a framework much like that of multiple regression:

$$transformation(p_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} \tag{9.30}$$

We want to choose a transformation in the equation that makes practical and mathematical sense. For example, we want a transformation that makes the range of possibilities on the left hand side of the equation equal to the range of possibilities for the right hand side; if there was no transformation for this equation, the left hand side could only take values between 0 and 1, but the right hand side could take values outside of this range. A common transformation for p_i is the **logit transformation**, which may be written as

$$logit(p_i) = \log_e \left(\frac{p_i}{1 - p_i} \right)$$

The logit transformation is shown in Figure 9.22. Below, we rewrite the equation relating Y_i to its

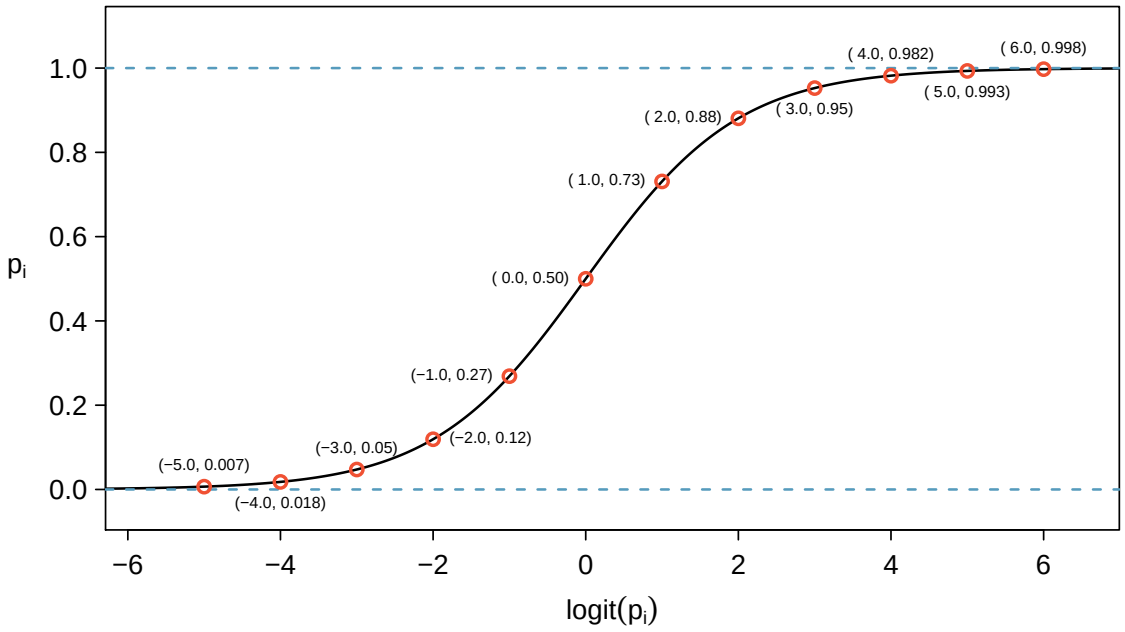


Figure 9.22: Values of p_i against values of $logit(p_i)$.

predictors using the logit transformation of p_i:

$$\log_e\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i}$$

In our resume example, there are 8 predictor variables, so $k = 8$. While the precise choice of a logit function isn't intuitive, it is based on theory that underpins generalized linear models, which is beyond the scope of this book. Fortunately, once we fit a model using software, it will start to feel like we're back in the multiple regression context, even if the interpretation of the coefficients is more complex.

EXAMPLE 9.31

We start by fitting a model with a single predictor: `honors`. This variable indicates whether the applicant had any type of honors listed on their resume, such as employee of the month. The following logistic regression model was fit using statistical software:

$$\log\left(\frac{p_i}{1-p_i}\right) = -2.4998 + 0.8668 \times \texttt{honors}$$

(a) If a resume is randomly selected from the study and it does not have any honors listed, what is the probability resulted in a callback?

(E) (b) What would the probability be if the resume did list some honors?

(a) If a randomly chosen resume from those sent out is considered, and it does not list honors, then `honors` takes value 0 and the right side of the model equation equals -2.4998. Solving for p_i: $\frac{e^{-2.4998}}{1+e^{-2.4998}} = 0.076$. Just as we labeled a fitted value of y_i with a "hat" in single-variable and multiple regression, we do the same for this probability: $\hat{p}_i = 0.076$.

(b) If the resume had listed some honors, then the right side of the model equation is $-2.4998 + 0.8668 \times 1 = -1.6330$, which corresponds to a probability $\hat{p}_i = 0.163$.

Notice that we could examine -2.4998 and -1.6330 in Figure 9.22 to estimate the probability before formally calculating the value.

To convert from values on the logistic regression scale (e.g. -2.4998 and -1.6330 in Example 9.31), use the following formula, which is the result of solving for p_i in the regression model:

$$p_i = \frac{e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}{1 + e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}$$

As with most applied data problems, we substitute the point estimates for the parameters (the β_i) so that we can make use of this formula. In Example 9.31, the probabilities were calculated as

$$\frac{e^{-2.4998}}{1 + e^{-2.4998}} = 0.076 \qquad \frac{e^{-2.4998+0.8668}}{1 + e^{-2.4998+0.8668}} = 0.163$$

While knowing whether a resume listed honors provides some signal when predicting whether or not the employer would call, we would like to account for many different variables at once to understand how each of the different resume characteristics affected the chance of a callback.

9.5.3 Building the logistic model with many variables

We used statistical software to fit the logistic regression model with all 8 predictors described in Figure 9.21. Like multiple regression, the result may be presented in a summary table, which is shown in Figure 9.23. The structure of this table is almost identical to that of multiple regression; the only notable difference is that the p-values are calculated using the normal distribution rather than the t-distribution.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.6632	0.1820	-14.64	<0.0001
job_city: *Chicago*	-0.4403	0.1142	-3.85	0.0001
college_degree	-0.0666	0.1211	-0.55	0.5821
years_experience	0.0200	0.0102	1.96	0.0503
honors	0.7694	0.1858	4.14	<0.0001
military	-0.3422	0.2157	-1.59	0.1127
email_address	0.2183	0.1133	1.93	0.0541
race: *white*	0.4424	0.1080	4.10	<0.0001
sex: *male*	-0.1818	0.1376	-1.32	0.1863

Figure 9.23: Summary table for the full logistic regression model for the resume callback example.

Just like multiple regression, we could trim some variables from the model. Here we'll use a statistic called **Akaike information criterion (AIC)**, which is an analog to how we used adjusted R-squared in multiple regression, and we look for models with a lower AIC through a backward elimination strategy. After using this criteria, the `college_degree` variable is eliminated, giving the smaller model summarized in Figure 9.24, which is what we'll rely on for the remainder of this section.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.7162	0.1551	-17.51	<0.0001
job_city: *Chicago*	-0.4364	0.1141	-3.83	0.0001
years_experience	0.0206	0.0102	2.02	0.0430
honors	0.7634	0.1852	4.12	<0.0001
military	-0.3443	0.2157	-1.60	0.1105
email_address	0.2221	0.1130	1.97	0.0494
race: *white*	0.4429	0.1080	4.10	<0.0001
sex: *male*	-0.1959	0.1352	-1.45	0.1473

Figure 9.24: Summary table for the logistic regression model for the resume callback example, where variable selection has been performed using AIC.

EXAMPLE 9.32

The `race` variable had taken only two levels: `black` and `white`. Based on the model results, was race a meaningful factor for if a prospective employer would call back?

(E) We see that the p-value for this coefficient is very small (very nearly zero), which implies that race played a statistically significant role in whether a candidate received a callback. Additionally, we see that the coefficient shown corresponds to the level of `white`, and it is positive. This positive coefficient reflects a positive gain in callback rate for resumes where the candidate's first name implied they were White. The data provide very strong evidence of racism by prospective employers that favors resumes where the first name is typically interpreted to be White.

The coefficient of $\texttt{race}_{\texttt{white}}$ in the full model in Figure 9.23, is nearly identical to the model shown in Figure 9.24. The predictors in this experiment were thoughtfully laid out so that the coefficient estimates would typically not be much influenced by which other predictors were in the model, which aligned with the motivation of the study to tease out which effects were important to getting a callback. In most observational data, it's common for point estimates to change a little, and sometimes a lot, depending on which other variables are included in the model.

EXAMPLE 9.33

Use the model summarized in Figure 9.24 to estimate the probability of receiving a callback for a job in Chicago where the candidate lists 14 years experience, no honors, no military experience, includes an email address, and has a first name that implies they are a White male.

We can start by writing out the equation using the coefficients from the model, then we can add in the corresponding values of each variable for this individual:

(E)
$$\begin{aligned} &log\left(\frac{p}{1-p}\right) \\ &\quad = -2.7162 - 0.4364 \times \texttt{job\_city}_{\texttt{Chicago}} + 0.0206 \times \texttt{years\_experience} + 0.7634 \times \texttt{honors} \\ &\qquad - 0.3443 \times \texttt{military} + 0.2221 \times \texttt{email} + 0.4429 \times \texttt{race}_{\texttt{white}} - 0.1959 \times \texttt{sex}_{\texttt{male}} \\ &\quad = -2.7162 - 0.4364 \times 1 + 0.0206 \times 14 + 0.7634 \times 0 \\ &\qquad - 0.3443 \times 0 + 0.2221 \times 1 + 0.4429 \times 1 - 0.1959 \times 1 \\ &\quad = -2.3955 \end{aligned}$$

We can now back-solve for p: the chance such an individual will receive a callback is about 8.35%.

EXAMPLE 9.34

Compute the probability of a callback for an individual with a name commonly inferred to be from a Black male but who otherwise has the same characteristics as the one described in Example 9.33.

(E) We can complete the same steps for an individual with the same characteristics who is Black, where the only difference in the calculation is that the indicator variable $\texttt{race}_{\texttt{white}}$ will take a value of `0`. Doing so yields a probability of 0.0553. Let's compare the results with those of Example 9.33.

In practical terms, an individual perceived as White based on their first name would need to apply to $\frac{1}{0.0835} \approx 12$ jobs on average to receive a callback, while an individual perceived as Black based on their first name would need to apply to $\frac{1}{0.0553} \approx 18$ jobs on average to receive a callback. That is, applicants who are perceived as Black need to apply to 50% more employers to receive a callback than someone who is perceived as White based on their first name for jobs like those in the study.

What we've quantified in this section is alarming and disturbing. However, one aspect that makes this racism so difficult to address is that the experiment, as well-designed as it is, cannot send us much signal about which employers are discriminating. It is only possible to say that discrimination is happening, even if we cannot say which particular callbacks – or non-callbacks – represent discrimination. Finding strong evidence of racism for individual cases is a persistent challenge in enforcing anti-discrimination laws.

9.5.4 Diagnostics for the callback rate model

LOGISTIC REGRESSION CONDITIONS

There are two key conditions for fitting a logistic regression model:

1. Each outcome Y_i is independent of the other outcomes.
2. Each predictor x_i is linearly related to logit(p_i) if all other predictors are held constant.

The first logistic regression model condition – independence of the outcomes – is reasonable for the experiment since characteristics of resumes were randomly assigned to the resumes that were sent out.

The second condition of the logistic regression model is not easily checked without a fairly sizable amount of data. Luckily, we have 4870 resume submissions in the data set! Let's first visualize these data by plotting the true classification of the resumes against the model's fitted probabilities, as shown in Figure 9.25.

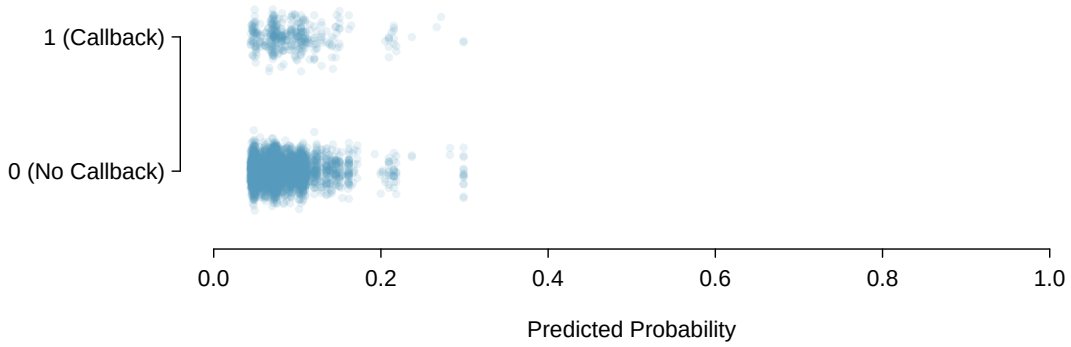


Figure 9.25: The predicted probability that each of the 4870 resumes results in a callback. Noise (small, random vertical shifts) have been added to each point so points with nearly identical values aren't plotted exactly on top of one another.

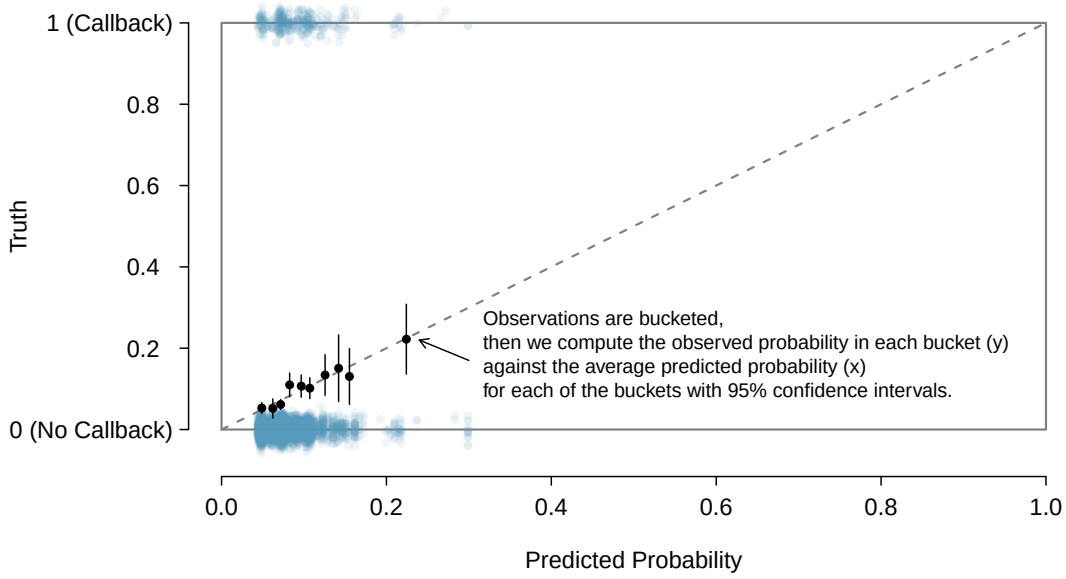


Figure 9.26: The dashed line is within the confidence bound of the 95% confidence intervals of each of the buckets, suggesting the logistic fit is reasonable.

We'd like to assess the quality of the model. For example, we might ask: if we look at resumes that we modeled as having a 10% chance of getting a callback, do we find about 10% of them actually receive a callback? We can check this for groups of the data by constructing a plot as follows:

1. Bucket the data into groups based on their predicted probabilities.
2. Compute the average predicted probability for each group.
3. Compute the observed probability for each group, along with a 95% confidence interval.
4. Plot the observed probabilities (with 95% confidence intervals) against the average predicted probabilities for each group.

The points plotted should fall close to the line $y = x$, since the predicted probabilities should be similar to the observed probabilities. We can use the confidence intervals to roughly gauge whether anything might be amiss. Such a plot is shown in Figure 9.26.

Additional diagnostics may be created that are similar to those featured in Section 9.3. For instance, we could compute residuals as the observed outcome minus the expected outcome ($e_i = Y_i - \hat{p}_i$), and then we could create plots of these residuals against each predictor. We might also create a plot like that in Figure 9.26 to better understand the deviations.

9.5.5 Exploring discrimination between groups of different sizes

Any form of discrimination is concerning, and this is why we decided it was so important to discuss this topic using data. The resume study also only examined discrimination in a single aspect: whether a prospective employer would call a candidate who submitted their resume. There was a 50% higher barrier for resumes simply when the candidate had a first name that was perceived to be from a Black individual. It's unlikely that discrimination would stop there.

EXAMPLE 9.35

Let's consider a sex-imbalanced company that consists of 20% women and 80% men,[22] and we'll suppose that the company is very large, consisting of perhaps 20,000 employees. Suppose when someone goes up for promotion at this company, 5 of their colleagues are randomly chosen to provide feedback on their work.

Now let's imagine that 10% of the people in the company are prejudiced against the other sex. That is, 10% of men are prejudiced against women, and similarly, 10% of women are prejudiced against men.

Who is discriminated against more at the company, men or women?

(E)

Let's suppose we took 100 men who have gone up for promotion in the past few years. For these men, $5 \times 100 = 500$ random colleagues will be tapped for their feedback, of which about 20% will be women (100 women). Of these 100 women, 10 are expected to be biased against the man they are reviewing. Then, of the 500 colleagues reviewing them, men will experience discrimination by about 2% of their colleagues when they go up for promotion.

Let's do a similar calculation for 100 women who have gone up for promotion in the last few years. They will also have 500 random colleagues providing feedback, of which about 400 (80%) will be men. Of these 400 men, about 40 (10%) hold a bias against women. Of the 500 colleagues providing feedback on the promotion packet for these women, 8% of the colleagues hold a bias against the women.

Example 9.35 highlights something profound: even in a hypothetical setting where each demographic has the same degree of prejudice against the other demographic, the smaller group experiences the negative effects more frequently. Additionally, if we would complete a handful of examples like the one above with different numbers, we'd learn that the greater the imbalance in the population groups, the more the smaller group is disproportionately impacted.[23]

Of course, there are other considerable real-world omissions from the hypothetical example. For example, studies have found instances where people from an oppressed group also discriminate against others within their own oppressed group. As another example, there are also instances where a majority group can be oppressed, with apartheid in South Africa being one such historic example. Ultimately, discrimination is complex, and there are many factors at play beyond the mathematics property we observed in Example 9.35.

We close this book on this serious topic, and we hope it inspires you to think about the power of reasoning with data. Whether it is with a formal statistical model or by using critical thinking skills to structure a problem, we hope the ideas you have learned will help you do more and do better in life.

[22] A more thoughtful example would include non-binary individuals.

[23] If a proportion p of a company are women and the rest of the company consists of men, then under the hypothetical situation the ratio of rates of discrimination against women vs men would be given by $\frac{1-p}{p}$; this ratio is always greater than 1 when $p < 0.5$.