3.3 Testing for goodness of fit using chi-square (special topic)

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. (Background on the geometric distribution is not necessary.)

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 Table 3.5, 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

Table 3.5: 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 the years 1990-2011 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.

3.3.1 Creating a test statistic for one-way tables

Example 3.18 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?

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 3.19 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 Table 3.6.

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

Table 3.6: Actual and expected make-up of the jurors.

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 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.

3.3.2 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.[13] Therefore:

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

[13]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.

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:

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

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
chi-square
test statistic

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

The final number X^2 summarizes how strongly the observed counts tend to deviate from the null counts. In Section 3.3.4, 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.

3.3.3 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 the 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 3.20 Figure 3.7 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?[14]

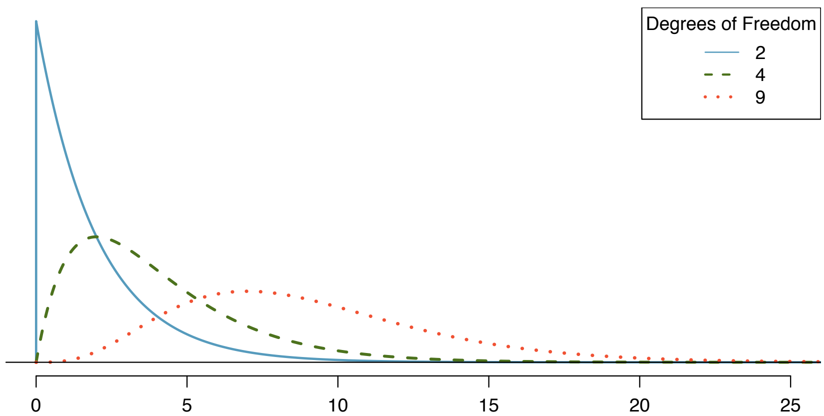

Figure 3.7: Three chi-square distributions with varying degrees of freedom.

Figure 3.7 and Guided Practice 3.20 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. To do so, a new table is needed: the **chi-square table**, partially shown in Table 3.8. A more complete table is presented in Appendix C.3 on page 344. Using this table, we identify a range for the area, and we examine a particular row for distributions with different degrees of freedom. One important quality of this table: the chi-square table only provides upper tail values.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	2	2.41	**3.22**	**4.61**	5.99	7.82	9.21	10.60	13.82
	3	*3.66*	*4.64*	6.25	*7.81*	*9.84*	*11.34*	*12.84*	*16.27*
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

Table 3.8: A section of the chi-square table. A complete table is in Appendix C.3 on page 344.

[14](a) The center becomes larger. If we look carefully, we can see that the center 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.

Example 3.21 Figure 3.9(a) shows a chi-square distribution with 3 degrees of freedom and an upper shaded tail starting at 6.25. Use Table 3.8 to estimate the shaded area.

This distribution has three degrees of freedom, so only the row with 3 degrees of freedom (df) is relevant. This row has been italicized in the table. Next, we see that the value – 6.25 – falls in the column with upper tail area 0.1. That is, the shaded upper tail of Figure 3.9(a) has area 0.1.

Example 3.22 We rarely observe the *exact* value in the table. For instance, Figure 3.9(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, which does not fall in Table 3.8. Find the approximate tail area.

The cutoff 4.3 falls between the second and third columns in the 2 degrees of freedom row. Because these columns correspond to tail areas of 0.2 and 0.1, we can be certain that the area shaded in Figure 3.9(b) is between 0.1 and 0.2.

Example 3.23 Figure 3.9(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.

Looking in the row with 5 df, 5.1 falls below the smallest cutoff for this row (6.06). That means we can only say that the area is *greater than 0.3*.

Guided Practice 3.24 Figure 3.9(d) shows a cutoff of 11.7 on a chi-square distribution with 7 degrees of freedom. Find the area of the upper tail.[15]

Guided Practice 3.25 Figure 3.9(e) shows a cutoff of 10 on a chi-square distribution with 4 degrees of freedom. Find the area of the upper tail.[16]

Guided Practice 3.26 Figure 3.9(f) shows a cutoff of 9.21 with a chi-square distribution with 3 df. Find the area of the upper tail.[17]

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

In Section 3.3.2, 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.

---

[15]The value 11.7 falls between 9.80 and 12.02 in the 7 df row. Thus, the area is between 0.1 and 0.2.
[16]The area is between 0.02 and 0.05.
[17]Between 0.02 and 0.05.

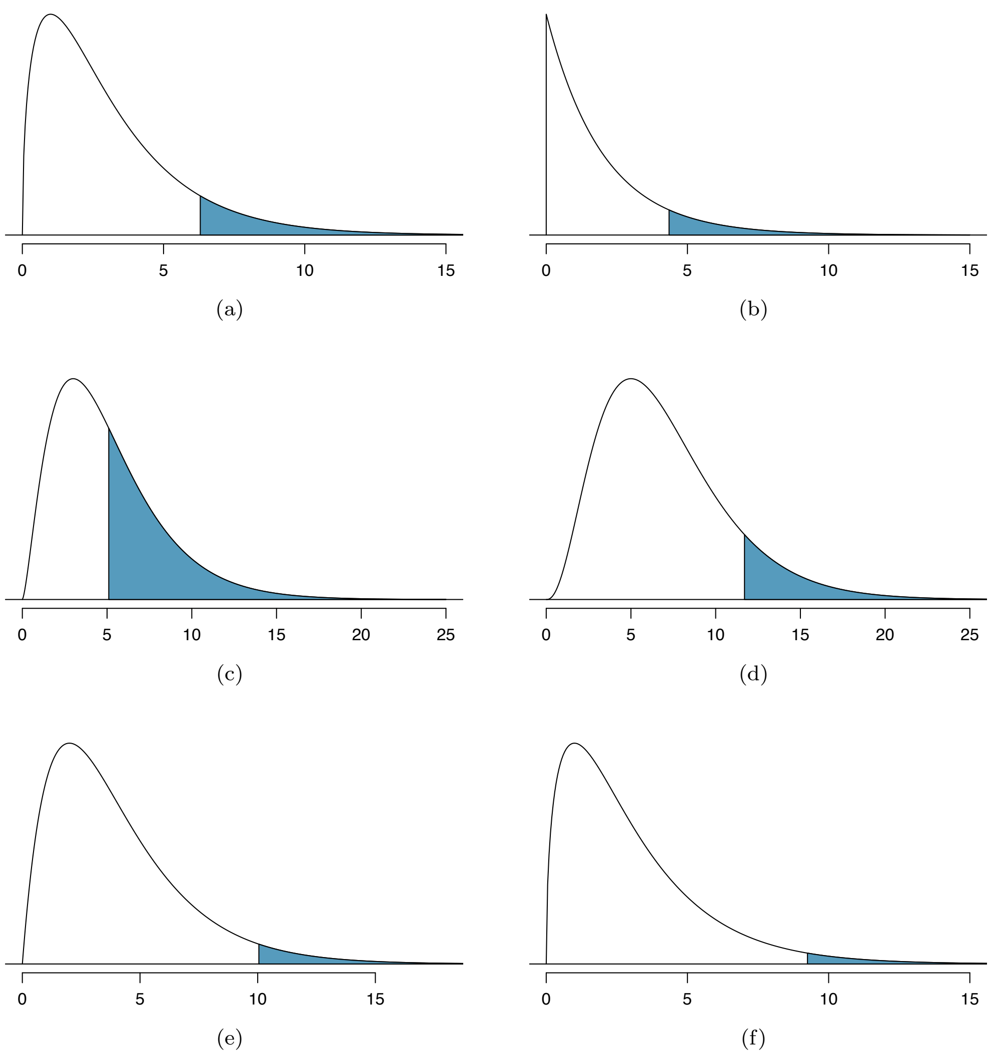

Figure 3.9: **(a)** Chi-square distribution with 3 degrees of freedom, area above 6.25 shaded. **(b)** 2 degrees of freedom, area above 4.3 shaded. **(c)** 5 degrees of freedom, area above 5.1 shaded. **(d)** 7 degrees of freedom, area above 11.7 shaded. **(e)** 4 degrees of freedom, area above 10 shaded. **(f)** 3 degrees of freedom, area above 9.21 shaded.

Example 3.27 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?

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 the normal model 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 3.28 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.

The chi-square distribution and p-value are shown in Figure 3.10. Because larger chi-square values correspond to stronger evidence against the null hypothesis, we shade the upper tail to represent the p-value. Using the chi-square table in Appendix C.3 or the short table on page 137, we can determine that the area is between 0.1 and 0.2. That is, the p-value is larger than 0.1 but smaller than 0.2. 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.

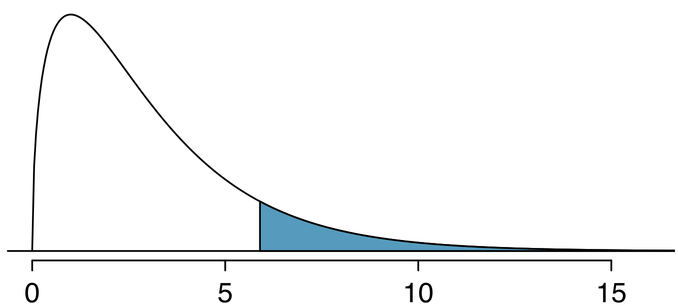

Figure 3.10: The p-value for the juror hypothesis test is shaded in the chi-square distribution with $df = 3$.

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.

TIP: Conditions for the chi-square test

There are three 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.

Degrees of freedom We only apply the chi-square technique when the table is associated with a chi-square distribution with 2 or more degrees of freedom.

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 3.1.

3.3.5 Evaluating goodness of fit for a distribution

We can apply our new 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 1990-2011 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. Table 3.11 shows the number of waiting days for a positive trading day during 1990-2011 for the S&P500.

Days	1	2	3	4	5	6	7+	Total
Observed	1532	760	338	194	74	33	17	2948

Table 3.11: Observed distribution of the waiting time until a positive trading day for the S&P500, 1990-2011.

We consider how many days one must wait until observing an `Up` day on the S&P500 stock exchange. 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 from 1990 to 2011 and summarize the waiting times in Table 3.12 and Figure 3.13. The S&P500 was positive on 53.2% 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 Table 3.12, can be compared to the expected counts from the *Geometric Model* row. The method for computing expected counts is discussed in Table 3.12. 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

Days	1	2	3	4	5	6	7+	Total
Observed	1532	760	338	194	74	33	17	2948
Geometric Model	1569	734	343	161	75	35	31	2948

Table 3.12: 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 ($P(D) = (1 - 0.532)^{D-1}(0.532)$) and multiply by the total number of streaks, 2948. For example, waiting for three days occurs under the geometric model about $0.468^2 \times 0.532 = 11.65\%$ of the time, which corresponds to $0.1165 \times 2948 = 343$ streaks.

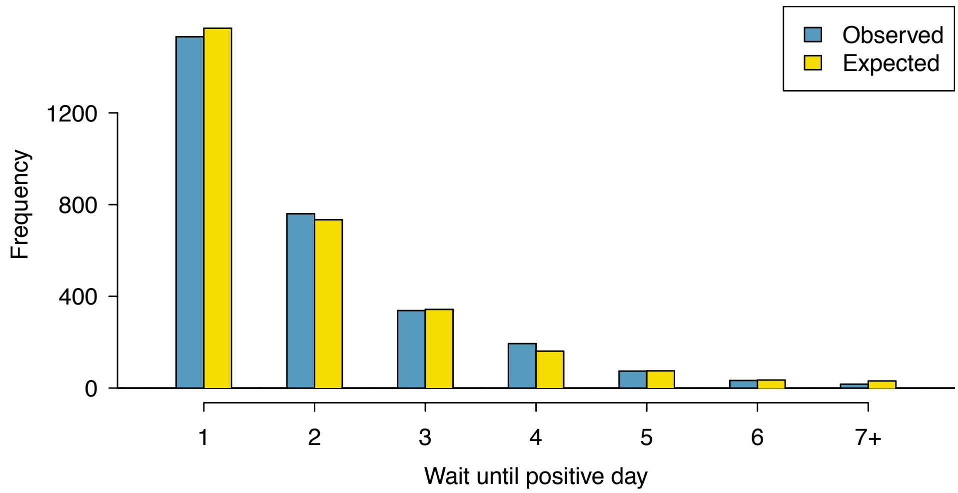

Figure 3.13: Side-by-side bar plot of the observed and expected counts for each waiting time.

counts. That is, this strategy identifies what proportion of the total count we would expect to be in each bin.

Example 3.29 Do you notice any unusually large deviations in the graph? Can you tell if these deviations are due to chance just by looking?

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 Table 3.12.

Guided Practice 3.30 Table 3.12 provides a set of count data for waiting times ($O_1 = 1532$, $O_2 = 760$, ...) and expected counts under the geometric distribution ($E_1 = 1569$, $E_2 = 734$, ...). Compute the chi-square test statistic, X^2.[18]

Guided Practice 3.31 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?[19]

Example 3.32 If the observed counts follow the geometric model, then the chi-square test statistic $X^2 = 15.08$ would closely follow a chi-square distribution with $df = 6$. Using this information, compute a p-value.

Figure 3.14 shows the chi-square distribution, cutoff, and the shaded p-value. If we look up the statistic $X^2 = 15.08$ in Appendix C.3, we find that the p-value is between 0.01 and 0.02. In other words, we have sufficient evidence to reject the notion that the wait times follow a geometric distribution, i.e. trading days are not independent and past days may help predict what the stock market will do today.

[18] $X^2 = \frac{(1532-1569)^2}{1569} + \frac{(760-734)^2}{734} + \cdots + \frac{(17-31)^2}{31} = 15.08$

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

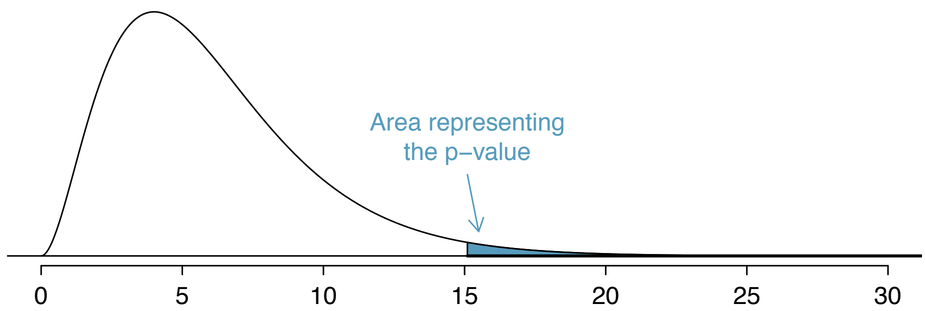

Figure 3.14: Chi-square distribution with 6 degrees of freedom. The p-value for the stock analysis is shaded.

Example 3.33 In Example 3.32, we rejected the null hypothesis that the trading days are independent. Why is this so important?

Because the data provided strong evidence that the geometric distribution is not appropriate, we reject the claim that trading days are independent. While it is not obvious how to exploit this information, it suggests there are some hidden patterns in the data that could be interesting and possibly useful to a stock trader.

3.4 Testing for independence in two-way tables (special topic)

Google is constantly running experiments to test new search algorithms. For example, Google might test three algorithms using a sample of 10,000 google.com search queries. Table 3.15 shows an example of 10,000 queries split into three algorithm groups.[20] The group sizes were specified before the start of the experiment to be 5000 for the current algorithm and 2500 for each test algorithm.

Search algorithm	current	test 1	test 2	Total
Counts	5000	2500	2500	10000

Table 3.15: Google experiment breakdown of test subjects into three search groups.

Example 3.34 What is the ultimate goal of the Google experiment? What are the null and alternative hypotheses, in regular words?

The ultimate goal is to see whether there is a difference in the performance of the algorithms. The hypotheses can be described as the following:

H_0: The algorithms each perform equally well.

H_A: The algorithms do not perform equally well.

[20]Google regularly runs experiments in this manner to help improve their search engine. It is entirely possible that if you perform a search and so does your friend, that you will have different search results. While the data presented in this section resemble what might be encountered in a real experiment, these data are simulated.

In this experiment, the explanatory variable is the search algorithm. However, an outcome variable is also needed. This outcome variable should somehow reflect whether the search results align with the user's interests. One possible way to quantify this is to determine whether (1) the user clicked one of the links provided and did not try a new search, or (2) the user performed a related search. Under scenario (1), we might think that the user was satisfied with the search results. Under scenario (2), the search results probably were not relevant, so the user tried a second search.

Table 3.16 provides the results from the experiment. These data are very similar to the count data in Section 3.3. However, now the different combinations of two variables are binned in a *two-way* table. In examining these data, we want to evaluate whether there is strong evidence that at least one algorithm is performing better than the others. To do so, we apply a chi-square test to this two-way table. The ideas of this test are similar to those ideas in the one-way table case. However, degrees of freedom and expected counts are computed a little differently than before.

Search algorithm	current	test 1	test 2	Total
No new search	3511	1749	1818	7078
New search	1489	751	682	2922
Total	5000	2500	2500	10000

Table 3.16: Results of the Google search algorithm experiment.

What is so different about one-way tables and two-way tables?
A one-way table describes counts for each outcome in a single variable. A two-way table describes counts for *combinations* of outcomes for two variables. When we consider a two-way table, we often would like to know, are these variables related in any way? That is, are they dependent (versus independent)?

The hypothesis test for this Google experiment is really about assessing whether there is statistically significant evidence that the choice of the algorithm affects whether a user performs a second search. In other words, the goal is to check whether the `search` variable is independent of the `algorithm` variable.

3.4.1 Expected counts in two-way tables

Example 3.35 From the experiment, we estimate the proportion of users who were satisfied with their initial search (no new search) as $7078/10000 = 0.7078$. If there really is no difference among the algorithms and 70.78% of people are satisfied with the search results, how many of the 5000 people in the "current algorithm" group would be expected to not perform a new search?

About 70.78% of the 5000 would be satisfied with the initial search:

$$0.7078 \times 5000 = 3539 \text{ users}$$

That is, if there was no difference between the three groups, then we would expect 3539 of the current algorithm users not to perform a new search.

Guided Practice 3.36 Using the same rationale described in Example 3.35, about how many users in each test group would not perform a new search if the algorithms were equally helpful?[21]

We can compute the expected number of users who would perform a new search for each group using the same strategy employed in Example 3.35 and Guided Practice 3.36. These expected counts were used to construct Table 3.17, which is the same as Table 3.16, except now the expected counts have been added in parentheses.

Search algorithm	current		test 1		test 2		Total
No new search	3511	**(3539)**	1749	**(1769.5)**	1818	**(1769.5)**	7078
New search	1489	**(1461)**	751	**(730.5)**	682	**(730.5)**	2922
Total	5000		2500		2500		10000

Table 3.17: The observed counts and the **(expected counts)**.

The examples and guided practice above provided some help in computing expected counts. In general, expected counts for a two-way table may be computed using the row totals, column totals, and the table total. For instance, if there was no difference between the groups, then about 70.78% of each column should be in the first row:

$$0.7078 \times (\text{column 1 total}) = 3539$$
$$0.7078 \times (\text{column 2 total}) = 1769.5$$
$$0.7078 \times (\text{column 3 total}) = 1769.5$$

Looking back to how the fraction 0.7078 was computed – as the fraction of users who did not perform a new search (7078/10000) – these three expected counts could have been computed as

$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 1 total}) = 3539$$
$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 2 total}) = 1769.5$$
$$\left(\frac{\text{row 1 total}}{\text{table total}}\right)(\text{column 3 total}) = 1769.5$$

This leads us to a general formula for computing expected counts in a two-way table when we would like to test whether there is strong evidence of an association between the column variable and row variable.

Computing expected counts in a two-way table

To identify the expected count for the i^{th} row and j^{th} column, compute

$$\text{Expected Count}_{\text{row } i,\text{ col } j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{table total}}$$

[21] We would expect $0.7078 * 2500 = 1769.5$. It is okay that this is a fraction.

3.4.2 The chi-square test for two-way tables

The chi-square test statistic for a two-way table is found the same way it is found for a one-way table. For each table count, compute

General formula	$\dfrac{(\text{observed count} \; - \; \text{expected count})^2}{\text{expected count}}$
Row 1, Col 1	$\dfrac{(3511 - 3539)^2}{3539} = 0.222$
Row 1, Col 2	$\dfrac{(1749 - 1769.5)^2}{1769.5} = 0.237$
$\vdots$	$\vdots$
Row 2, Col 3	$\dfrac{(682 - 730.5)^2}{730.5} = 3.220$

Adding the computed value for each cell gives the chi-square test statistic X^2:

$$X^2 = 0.222 + 0.237 + \cdots + 3.220 = 6.120$$

Just like before, this test statistic follows a chi-square distribution. However, the degrees of freedom are computed a little differently for a two-way table.[22] For two way tables, the degrees of freedom is equal to

$$df = (\text{number of rows minus 1}) \times (\text{number of columns minus 1})$$

In our example, the degrees of freedom parameter is

$$df = (2 - 1) \times (3 - 1) = 2$$

If the null hypothesis is true (i.e. the algorithms are equally useful), then the test statistic $X^2 = 6.12$ closely follows a chi-square distribution with 2 degrees of freedom. Using this information, we can compute the p-value for the test, which is depicted in Figure 3.18.

Computing degrees of freedom for a two-way table

When applying the chi-square test to a two-way table, we use

$$df = (R - 1) \times (C - 1)$$

where R is the number of rows in the table and C is the number of columns.

TIP: Use two-proportion methods for 2-by-2 contingency tables

When analyzing 2-by-2 contingency tables, use the two-proportion methods introduced in Section 3.2.

[22] Recall: in the one-way table, the degrees of freedom was the number of cells minus 1.

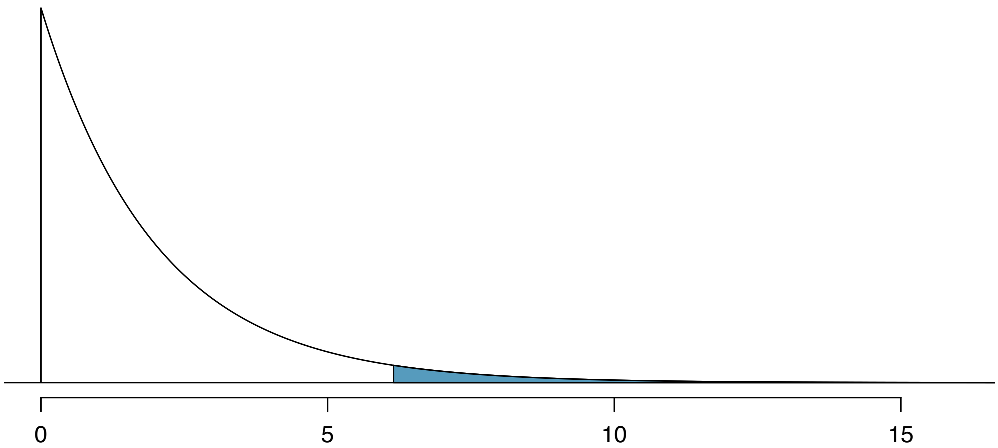

Figure 3.18: Computing the p-value for the Google hypothesis test.

		Congress		
	Obama	Democrats	Republicans	Total
Approve	842	736	541	2119
Disapprove	616	646	842	2104
Total	1458	1382	1383	4223

Table 3.19: Pew Research poll results of a March 2012 poll.

Example 3.37 Compute the p-value and draw a conclusion about whether the search algorithms have different performances.

Looking in Appendix C.3 on page 344, we examine the row corresponding to 2 degrees of freedom. The test statistic, $X^2 = 6.120$, falls between the fourth and fifth columns, which means the p-value is between 0.02 and 0.05. Because we typically test at a significance level of $\alpha = 0.05$ and the p-value is less than 0.05, the null hypothesis is rejected. That is, the data provide convincing evidence that there is some difference in performance among the algorithms.

Example 3.38 Table 3.19 summarizes the results of a Pew Research poll.[23] We would like to determine if there are actually differences in the approval ratings of Barack Obama, Democrats in Congress, and Republicans in Congress. What are appropriate hypotheses for such a test?

H_0: There is no difference in approval ratings between the three groups.

H_A: There is some difference in approval ratings between the three groups, e.g. perhaps Obama's approval differs from Democrats in Congress.

[23]See the Pew Research website: www.people-press.org/2012/03/14/romney-leads-gop-contest-trails-in-matchup-with-obama. The counts in Table 3.19 are approximate.

Guided Practice 3.39 A chi-square test for a two-way table may be used to test the hypotheses in Example 3.38. As a first step, compute the expected values for each of the six table cells.[24]

Guided Practice 3.40 Compute the chi-square test statistic.[25]

Guided Practice 3.41 Because there are 2 rows and 3 columns, the degrees of freedom for the test is $df = (2-1)\times(3-1) = 2$. Use $X^2 = 106.4$, $df = 2$, and the chi-square table on page 344 to evaluate whether to reject the null hypothesis.[26]

---

[24] The expected count for row one / column one is found by multiplying the row one total (2119) and column one total (1458), then dividing by the table total (4223): $\frac{2119\times 1458}{3902} = 731.6$. Similarly for the first column and the second row: $\frac{2104\times 1458}{4223} = 726.4$. Column 2: 693.5 and 688.5. Column 3: 694.0 and 689.0

[25] For each cell, compute $\frac{(\text{obs}-\text{exp})^2}{exp}$. For instance, the first row and first column: $\frac{(842-731.6)^2}{731.6} = 16.7$. Adding the results of each cell gives the chi-square test statistic: $X^2 = 16.7 + \cdots + 34.0 = 106.4$.

[26] The test statistic is larger than the right-most column of the $df = 2$ row of the chi-square table, meaning the p-value is less than 0.001. That is, we reject the null hypothesis because the p-value is less than 0.05, and we conclude that Americans' approval has differences among Democrats in Congress, Republicans in Congress, and the president.