

### 9.2.3: 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.

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 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.

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

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

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.

<sup>12</sup>Compute the test statistic:

$$Z = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{0.067 - 0}{0.026} = 2.58 \quad (9.2.3.1)$$

We leave the picture to you. Looking up  $Z = 2.58$  in the normal probability table: 0.9951. However this is the lower tail, and the upper tail represents the  $p$ -value:  $1 - 0.9951 = 0.0049$ . We reject the null hypothesis and conclude that dogs getting cancer and owners using 2,4-D are associated.

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.

#### Creating a test statistic for one-way tables

##### Example 9.2.3.1:

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?

##### Solution

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.

##### Exercise 9.2.3.1

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?

## Answer

Answers can be found in Table 6.6.

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

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

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.

## The chi-square test statistic

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

$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}} \quad (9.2.3.2)$$

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:

$$Z_1 = \frac{\text{observed white count} - \text{null white count}}{\text{SE of observed white count}} \quad (9.2.3.3)$$

The standard error for the point estimate of the count in binned data is the square root of the count under the null.<sup>13</sup> Therefore:

$$Z_1 = \frac{205 - 198}{\sqrt{198}} = 0.50 \quad (9.2.3.4)$$

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:

$$Z_2 = \frac{\overset{Black}{26 - 19.25}}{\sqrt{19.25}} = 1.54 \quad Z_3 = \frac{\overset{Hispanic}{25 - 33}}{\sqrt{33}} = -1.39 \quad Z_4 = \frac{\overset{Other}{19 - 24.75}}{\sqrt{24.75}} = -1.16 \quad (9.2.3.5)$$

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:

$$|Z_1| + |Z_2| + |Z_3| + |Z_4| = 4.58 \quad (9.2.3.6)$$

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:

$$Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 = 5.89 \quad (9.2.3.7)$$

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:

$$X^2 = \frac{(\text{observed count}_1 - \text{null count}_1)^2}{\text{null count}_1} + \dots + \frac{(\text{observed count}_4 - \text{null count}_4)^2}{\text{null count}_4} \quad (9.2.3.8)$$

<sup>13</sup>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 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.

### 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.

#### Exercise 9.2.3.1

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?<sup>14</sup>

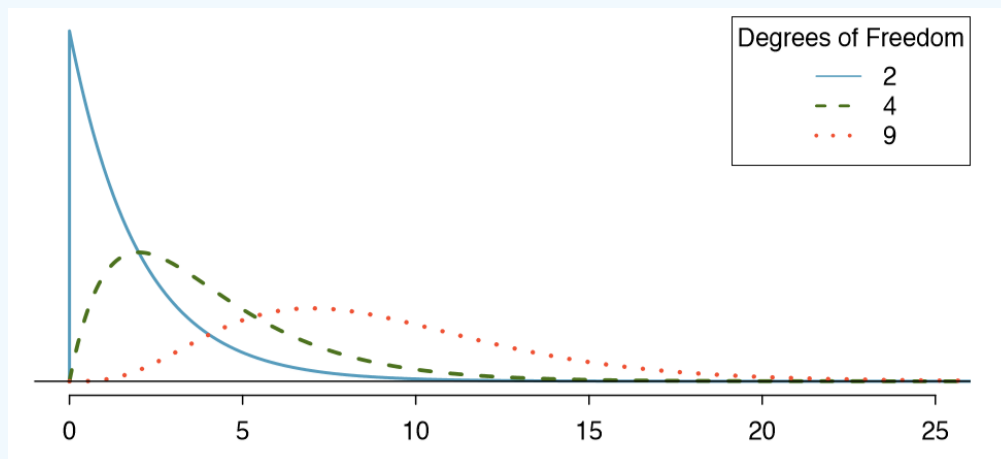


Figure 6.7: Three chi-square distributions with varying degrees of freedom.

#### Answer

<sup>14</sup>(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.

Figure 6.7 and Exercise 6.20 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 increases.

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 6.8. A more complete table is presented in Appendix B.3 on page 412. This table is very similar to the t table from Sections 5.3 and 5.4: we identify a range for the area, and we examine a particular row for distributions with different degrees of freedom. One important difference from the t table is that the chi-square table only provides upper tail values.

Table 6.8: A section of the chi-square table. A complete table is in Appendix B.3 on page 412.

Upper tail	0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df 1								
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

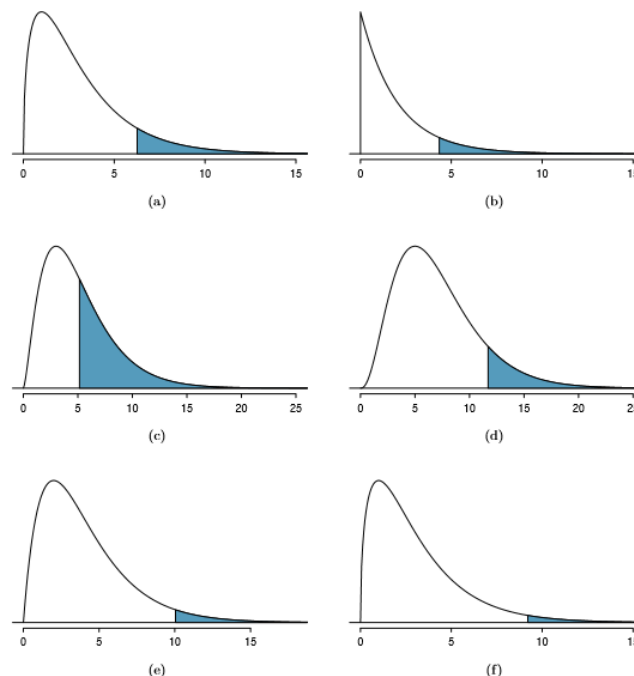


Figure 6.9: (a) Chi-square distribution with 3 degrees of freedom, area above 6.25 shaded. (b) 2 degrees of freedom, area above 4.3 shaded. (c) 5 degrees of freedom, area above 5.1 shaded. (d) 7 degrees of freedom, area above 11.7 shaded. (e) 4 degrees of freedom, area above 10 shaded. (f) 3 degrees of freedom, area above 9.21 shaded.

**Example 6.21** Figure 6.9(a) shows a chi-square distribution with 3 degrees of freedom and an upper shaded tail starting at 6.25. Use Table 6.8 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 6.9(a) has area 0.1.

**Example 6.22** We rarely observe the exact value in the table. For instance, Figure 6.9(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 6.8. 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 6.9(b) is between 0.1 and 0.2.

**Example 6.23** Figure 6.9(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.

**Exercise 6.24** Figure 6.9(d) shows a cutoff of 11.7 on a chi-square distribution with 7 degrees of freedom. Find the area of the upper tail.<sup>15</sup>

**Exercise 6.25** Figure 6.9(e) shows a cutoff of 10 on a chi-square distribution with 4 degrees of freedom. Find the area of the upper tail.<sup>16</sup>

**Exercise 6.26** Figure 6.9(f) shows a cutoff of 9.21 with a chi-square distribution with 3 df. Find the area of the upper tail.<sup>17</sup>

<sup>15</sup>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.

<sup>16</sup>The area is between 0.02 and 0.05.

<sup>17</sup>Between 0.02 and 0.05.

## 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 9.2.3.1:

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

#### Solution

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 9.2.3.1:

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 6.10. 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 B.3 or the short table on page 277, 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.

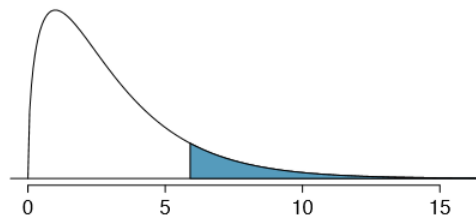


Figure 6.10: The p-value for the juror hypothesis test is shaded in the chi-square distribution with  $df = 3$ .

### Chi-square test for one-way table

Suppose we are to evaluate whether there is convincing evidence that a set of observed counts  $O_1, O_2, \dots, O_k$  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  $E_1, E_2, \dots, E_k$ . 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:

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \dots + \frac{(O_k - E_k)^2}{E_k} \quad (9.2.3.9)$$

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 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 6.1.

### Evaluating goodness of fit for a distribution

Section 3.3 would be useful background reading for this example, but it is not a prerequisite. 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

(9.2.3.10)

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 6.11 shows the number of waiting days for a positive trading day during 1990-2011 for the S&P500.

Table 6.11: Observed distribution of the waiting time until a positive trading day for the S&P500, 1990-2011.

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

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 6.12 and Figure 6.13. 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 6.12, can be compared to the expected counts from the Geometric Model row. The method for computing expected counts is discussed in Table 6.12. In general, the expected counts are determined by (1) identifying the null proportion associated with each

Table 6.12: 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.

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

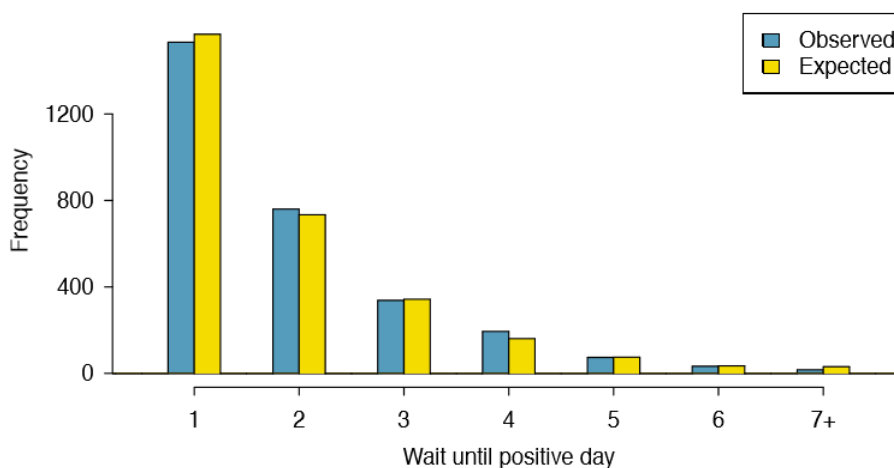


Figure 6.13: Side-by-side bar plot of the observed and expected counts for each waiting time.

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.

**Example 6.29** 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 6.12.

**Exercise 6.30** Table 6.12 provides a set of count data for waiting times ( $O_1 = 1532, O_2 = 760, \dots$ ) and expected counts under the geometric distribution ( $E_1 = 1569, E_2 = 734, \dots$ ). Compute the chi-square test statistic,  $X^2$ .<sup>18</sup>

**Exercise 6.31** 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?<sup>19</sup>

**Example 6.32** 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 6.14 shows the chi-square distribution, cutoff, and the shaded p-value. If we look up the statistic  $X^2 = 15.08$  in Appendix B.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

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

<sup>19</sup>There are  $k = 7$  groups, so we use  $df = k - 1 = 6$ .

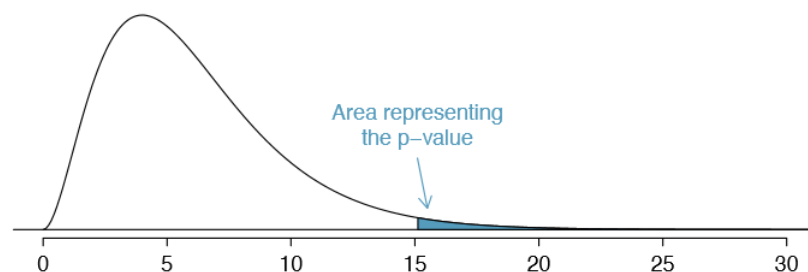


Figure 6.14: Chi-square distribution with 6 degrees of freedom. The p-value for the stock analysis is shaded.

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.

**Example 6.33** In Example 6.32, 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.

This page titled 9.2.3: Testing for Goodness of Fit using Chi-Square (Special Topic) is shared under a CC BY-SA 3.0 license and was authored, remixed, and/or curated by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel via source content that was edited to the style and standards of the LibreTexts platform.

- 6.3: Testing for Goodness of Fit using Chi-Square (Special Topic) by David Diez, Christopher Barr, & Mine Çetinkaya-Rundel is licensed CC BY-SA 3.0. Original source: <https://www.openintro.org/book/os>.