

---

---

# MODULE 18

---

## CHI-SQUARE TEST

### Objectives:

1. Identify when a Chi-Square Test is appropriate.
2. Perform the 11 steps of a significance test in a Chi-Square Test situation.

### Contents

18.1 Chi-Square Distribution . . . . .	212
18.2 Chi-Square Test Specifics . . . . .	214
18.3 Chi-Square Test in R . . . . .	219

SITUATIONS WHERE A CATEGORICAL response variable is recorded would be summarized with a frequency or percentage table (see Modules 6 and 9). The appropriate test statistic in these situations is a chi-square rather than a t. The Chi-Square Test statistic follows a chi-square distribution, which is introduced below. The rest of this module is dedicated to the general Chi-Square Test where the distribution of a categorical response variable is compared between two or more populations. The related goodness-of-fit test for a categorical response recorded for only one population is introduced in Module 19.

### 18.1 Chi-Square Distribution

A chi-square ( $\chi^2$ ) distribution is generally right-skewed (Figure 18.1), with the exact shape dictated by the degrees-of-freedom (df; as df increase, the sharpness of the skew decreases; Figure 18.1). In its simplest form, the  $\chi^2$  distribution arises as a sampling distribution for the  $\chi^2$  test statistic,

$$\chi^2 = \sum_{cells} \frac{(Observed - Expected)^2}{Expected}$$

where “Observed” and “Expected” represent the observed and expected individuals in the cells of frequency tables (see Module 6 and Module 9) and “cells” generically represents the number of cells in one of these tables. Thus, the  $\chi^2$  distribution arises from comparing the frequencies in two tables.<sup>1</sup>

Figure 18.1.  $\chi^2$  distributions with varying degrees-of-freedom.

Unlike with the other two distributions that we have seen (normal and t), the  $\chi^2$  distribution always represents the two-tailed situation, although the “two tails” will appear as one tail on the right side of the distribution. The simplest explanation for this characteristic is that the “squaring” in the calculation of the  $\chi^2$  test statistic results in what would be a “negative tail” being “folded over” onto what is the “positive tail.” Thus, all probability (i.e., area) calculations on a  $\chi^2$  distribution represent the two-tailed alternative hypotheses.

◇ Probability calculations on a  $\chi^2$  distribution always pertain to a two-tailed alternative hypothesis.

Proportional areas on a  $\chi^2$  distribution are computed using `distrib()` similar to what was described for normal and t distributions in Modules 7, 12, and 16. The major exceptions for using `distrib()` with a  $\chi^2$  distribution is that `distrib="chisq"` must be used and the degrees-of-freedom must be given to `df=` (how to find the df will be discussed in subsequent sections). In addition, if calculating a p-value, then `lower.tail=FALSE` is always used because the upper-tail probability represents the two-tailed alternative inherent to all Chi-Square Tests. For example, the area right of  $\chi^2 = 6.456$  on a  $\chi^2$  distribution with 2 df is 0.0396 (Figure 18.2).

```
> ( distrib(6.456,distrib="chisq",df=2,lower.tail=FALSE) )  
[1] 0.03963669
```

<sup>1</sup>Subsequent sections demonstrate how this test statistic is used to compare observed frequencies (i.e., from a sample) to a table of expected frequencies (i.e., from a null hypothesis).

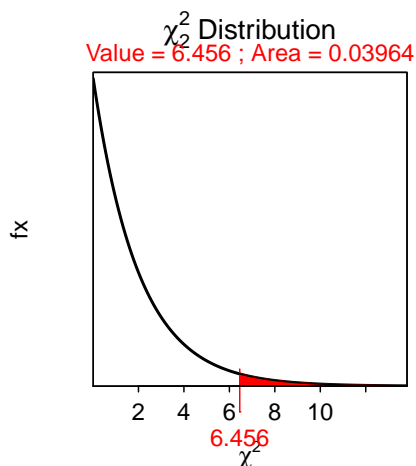


Figure 18.2. Depiction of the area to the right of  $\chi^2 = 6.456$  on a  $\chi^2$  distribution with 2 df.

## Review Exercises

- 18.1** What is the p-value if  $\chi^2 = 10.25$  and  $df = 3$ ? [Answer](#)
- 18.2** What is the p-value if  $\chi^2 = 10.25$  and  $df = 4$ ? [Answer](#)
- 18.3** What is the p-value if  $\chi^2 = 10.25$  and  $df = 6$ ? [Answer](#)

## 18.2 Chi-Square Test Specifics

Researchers commonly want to compare the distribution of individuals into the levels of a categorical variable among two or more populations. For example, researchers may want to determine if the distribution of failing students differs between males and females, if the distribution of kids playing sports differs between kids from high- or low-income families, if the distribution of four major plant species differs between two locations, or if the distribution of responses to a five-choice question differs between respondents from neighboring counties. All of these questions have a categorical response variable (fail or no, play sport or not, plant species, answer to five-choice question) compared among two or more populations (gender, income category, two locations, neighboring counties). The Chi-Square Test, the subject of this module, can be used for each of these situations.<sup>2</sup>

♦ A Chi-Square test is used to compare the distribution of responses among two or more populations.

<sup>2</sup>The Chi-Square Test presented here is quite flexible and can be derived from different types of hypotheses than those described here.

### 18.2.1 Hypotheses

The statistical hypotheses for a Chi-Square Test are “wordy.” To explore this, let’s first assume that a two-way frequency table (see Module 9) will summarize the data where the rows correspond to separate populations and the columns correspond to levels of the response variable. In this organization, the Chi-Square Test null hypothesis is that the row percentages (or proportions) are equal – i.e., “the percentage or proportional distribution of individuals into the levels of the response variable is the same for all populations.” The alternative hypothesis states that there is some difference among the row percentages – i.e., “the percentage or proportional distribution of individuals into the levels of the response variable is NOT the same for all populations.”

As one example (more are shown below), consider the following situation,

*An association of Christmas tree growers in Indiana sponsored a survey of Indiana households to help improve the marketing of Christmas trees. In telephone surveys of 421 households they found 160 households in rural areas and 261 households in urban areas. Of the rural households, 64 had a natural tree (as compared to an artificial tree). Of the urban households, 89 had a natural tree. Use these results to determine, at the 10% level, if the distribution of households with a natural tree differed between rural and urban households.*

The hypotheses for this situation are,

$H_0$  : “the distribution of households into the tree types is the same for urban and rural households”

$H_A$  : “the distribution of households into the tree types is NOT the same for urban and rural households”

### 18.2.2 Tables

As noted above, all two-way frequency tables used for a Chi-Square Test will be organized such that the response variable forms the columns and the populations form the rows. With this organization, the row-percentage table becomes the table of primary interest because it relates directly to the hypotheses described above. The question of a Chi-Square Test then becomes one of determining whether each row of the row-percentage table is equal given sampling variability.

◊ For Chi-Square Tests, the populations form the rows of the two-way frequency table such that the Chi-Square Test becomes a test of whether or not each row in the row-percentage table is equivalent given sampling variability.

The observed raw data must be organized into a two-way frequency table as described in Module 9. For example, the Christmas tree data is summarized as in Table 18.1. The actual calculations for a Chi-Square Test are performed on this observed table. However, the hypothesis test, as described above, is best viewed as a method to determine if each row of the row-percentage table is statistically equivalent or not. Thus, the row-percentage table computed from the frequency table is useful when interpreting the results of a Chi-Square Test (Table 18.2).

The Chi-Square Test requires constructing a table of expected values that is derived from the null hypothesis. Specifically, the “expected” table contains the expected frequency of individuals in each level of the response variable for each population assuming that the distribution of responses does not differ among populations. These expected table are computed from the margins of the observed table, but are best explained with an illustrative example.

In the Christmas tree example, the null hypothesis states that there is no difference in the distribution of households with a natural tree between the rural and urban areas. Thus, under this null hypothesis, one

Table 18.1. Frequency of individuals in urban and rural households that have a natural or an artificial Christmas tree.

Household	Tree Type		
	Natural	Artificial	
Urban	89	172	<b>261</b>
Rural	64	96	<b>160</b>
	<b>153</b>	<b>268</b>	<b>421</b>

Table 18.2. Percentage of individuals within urban and rural households that have a natural or an artificial Christmas tree.

Household	Tree Type		
	Natural	Artificial	
Urban	34.1	65.9	<b>100.0</b>
Rural	40.0	60.0	<b>100.0</b>
	<b>36.3</b>	<b>63.7</b>	<b>100.0</b>

would expect the proportion of households with a natural tree to be the same in both groups. The proportion of households with a natural tree, regardless of location, is  $\frac{153}{421}=0.363$ . Thus, under the null hypothesis, the proportion of rural AND the proportion of urban households with a natural tree is 0.363. Because there is a different number of urban and rural households in the study, the actual NUMBER (rather than proportion) of households expected to have a natural tree will differ. The NUMBER of urban households expected to HAVE a natural tree is found by multiplying the number of urban households by the common proportion computed above – i.e.,  $261 * 0.363=94.743$ . The remaining urban households would be expected to NOT have a natural tree – i.e.,  $261 - 94.743=261(1 - 0.363)=166.257$ . Similar calculations are made for the rural households (i.e.,  $160 * 0.363 = 58.08$  expected to have a natural tree and  $160 * (1 - 0.363) = 101.92$  expected to NOT have a natural tree.

These expected frequencies are computed directly and easily from the marginal totals of the observed frequency table (Table 18.1). For example, substituting the fractional representation of the decimal proportions into the calculation of the expected number of urban households with a natural tree gives  $261 * \frac{153}{421} = \frac{261*153}{421}$ . A close examination of this formula and the marginal totals in Table 18.1 shows that this value is equal to the product of the corresponding row and column marginal totals in the observed table divided by the total number of individuals. The other expected values follow a similar pattern as follows,

- $261 * \frac{268}{421} = \frac{261*268}{421} = 166.147$  urban households to NOT have a natural tree.
- $160 * \frac{153}{421} = \frac{160*153}{421} = 58.147$  rural households to have a natural tree.
- $160 * \frac{268}{421} = \frac{160*268}{421} = 101.853$  rural households to NOT have a natural tree.

Thus, all expected values in a Chi-Square test can be calculated multiplying the row and column totals of the frequency table and dividing by the total number of individuals. These expected values are summarized in a two-way table, called the expected frequencies table (Table 18.3).

Table 18.3. The expected frequency of individuals in urban and rural households that have a natural or an artificial Christmas tree.

Household	Tree Type		
	Natural	Artificial	
Urban	94.853	166.147	<b>261</b>
Rural	58.147	101.853	<b>160</b>
	<b>153</b>	<b>268</b>	<b>421</b>

◊ Expected frequencies are computed by multiplying observed row and column marginal totals and dividing by the total number of individuals.

### 18.2.3 Specifics

The Chi-Square Test is characterized by a categorical response variable recorded for two or more populations. The specifics of the Chi-Square Test are identified in Table 18.4.

Table 18.4. Characteristics of a Chi-Square Test.

- **Null Hypothesis:** “The distribution of individuals into the levels of the response variable is the same for all populations”
- **Alternative Hypothesis:** “The distribution of individuals into the levels of the response variable is NOT the same for all populations.”
- **Statistic:** Observed frequency table.
- **Test Statistic:**  $\chi^2 = \sum_{\text{cells}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$
- **df:**  $(r - 1)(c - 1)$  where  $r$  = number of rows and  $c$  = number of columns
- **Assumptions:** Expected value for each category is  $\geq 5$ .
- **When to Use:** Categorical response, two or more populations.

In general, confidence intervals are not constructed with a Chi-Square Test because of the complexity of the parameter (i.e., same size as the observed table). Thus, in this course, Step 11 for a hypothesis test will not be computed for a Chi-Square Test.

◊ Step 11 is not be computed for a Chi-Square Test.

### 18.2.4 Example – Christmas Trees

The 11-steps (Section 15.1) for a hypothesis test for the Christmas tree example are as follows:

1.  $\alpha=0.10$ .
2.  $H_0$ : “distribution of households with a natural tree is the same for urban and rural households” vs.  $H_A$ : “distribution of households with a natural tree is NOT the same for urban and rural households.”
3. A Chi-Square Test is required because (i) a categorical response variable was recorded (type of tree) and (ii) two populations were sampled (urban and rural households).
4. The data appear to be part of an observational study with no clear indication of randomization.
5. The expected frequency in each of the four cells is greater than five (Table 18.1).
6. The observed frequency table is in Table 18.1.
7.  $\chi^2 = \frac{(89-94.853)^2}{94.853} + \frac{(172-166.147)^2}{166.147} + \frac{(64-58.147)^2}{58.147} + \frac{(96-101.853)^2}{101.853} = 0.3611 + 0.2062 + 0.5891 + 0.3363 = 1.4927$  with 1 df.
8. p-value=0.0264.
9.  $H_0$  is not rejected because the p-value is  $> \alpha$ .
10. There does not appear to be a significant difference between the distribution of rural and the distribution of urban households that have a natural Christmas tree.
11. Not performed for Chi-Square Test.

**R Appendix:**

```
( distrib(4.927,distrib="chisq",df=1,lower.tail=FALSE) )
```

## Review Exercises

- 18.4** Researchers in Asia (Roberts, 2000) wanted to describe the distribution of the fish genera Cyprinidae in Asian rivers. They collected 228 fish from the Brahmaputra, Irrawaddy, and Salween rivers and recorded whether the fish was a member of the Cyprinidae family or not. Because the rivers were relatively equal in size, they expected the same proportions of Cyprinidae in each of the rivers. Using the data in the table below, test to see if there was a difference in the proportion of Cyprinidae among the rivers at the 5% level.

[Answer](#)

River	Cyprinidae	
	Yes	No
Brahmaputra	22	51
Irrawaddy	25	53
Salween	30	47

- 18.5** The American Nurses Credentialing Center (ANCC) has created guidelines for nursing administration. Some research findings have suggested that ANCC-recognized hospitals also have favorable practice environments for nurses. To study this further and in relation to oncology units, Fries (2005) examined the practice environments and outcomes of nurses working in and out of oncology units in hospitals that adhere and don't adhere to the ANCC guidelines. As part of his study, he determined, through surveys, whether nurses were experiencing high emotional exhaustion (HEE) or not. The results of his study are shown in the table below (note "onc" represent oncology units). Use these results to determine, at the 5% level, if the proportion of nurses experiencing HEE differs among the four categories of hospitals. [Answer](#)

Clinic Type	HEE	not HEE	total
non-ANCC, non-Onc	362	534	896
non-ANCC, Onc	58	92	150
ANCC, non-Onc	197	558	755
ANCC, Onc	30	125	155
total	647	1309	1956

- 18.6** Fiebach *et al.* (1990) examined the immediate survival of 790 males and 332 females who were hospitalized following a myocardial infarction (i.e., a "heart attack"). During hospitalization, 70 men and 47 women died. Is there a difference, at the 5% level, in mortality rate (proportion of patients that died) between men and women during hospitalization? [Answer](#)

- 18.7** Eight American undergraduate women were part of a study to determine if whether or not a response is received depends on the size of group addressed (Jones and Foshay 1984). Each student was instructed to say "Hello" to strangers or groups of strangers that they encountered around campus, on the streets in town, in stores, etc. They were told to not make direct eye contact with anyone in the group but to look in the general direction of the group focusing on the shoulders or hair of individuals or the general middle of

a group. The students recorded a variety of information for each encounter including how many individuals were in the group and whether at least one person responded to the greeting. The study included 119 people greeted individually, 94 groups of two or three, and 27 groups of four, five or six. They found that 92 of the individuals, 65 of the groups of two or three, and 13 of the groups of four, five, or six responded to the greeting. Determine, at the 5% level, if there is a significant difference in the frequency of responses among the three different sizes of groups (i.e., individuals; two or three; or four, five, or six). [Answer](#)

## 18.3 Chi-Square Test in R

### 18.3.1 Data Format

The data for a Chi-Square Test may be entered from summarized data or computed from raw data on individuals. Raw data must be in stacked format where one column in the data.frame represents the response variable and another column represents the populations (see Sections 4.3.2 and 17.3). Raw data must be summarized into a two-way frequency table with `xtabs()` as described in Module 9. The two-way table must contain frequencies, not proportions or percentages (don't use `percTable()`), without marginal totals (don't use `addMargins()`).

Summarized data must be entered into a two-dimensional **matrix** with `matrix()`. The frequencies must first be entered into a vector with the first row of values followed by second row and so on. This vector is then the first argument to `matrix()`, which will also include the number of rows in the frequency table in `nrow=` and `byrow=TRUE` (indicate that the values in the vector should be entered into the matrix in a row-wise manner). The process of entering summarized data into a matrix is better explained by example.

Suppose that you are given this observed frequency table.

Location	Species						
	A	B	C	D	E	F	
DI	34	22	14	13	12	5	100
BP	62	12	8	7	6	5	100
	96	34	22	20	18	10	200

The observed frequencies, ignoring the marginal sums, are first entered into a vector called `freq` below, which is then transformed into a matrix called `obstbl`.

```
> ( freq <- c(34,22,14,13,12,5,62,12,8,7,6,5) )
[1] 34 22 14 13 12 5 62 12 8 7 6 5
> ( obstbl <- matrix(freq,nrow=2,byrow=TRUE) )
      [,1] [,2] [,3] [,4] [,5] [,6]
[1,]   34   22   14   13   12    5
[2,]   62   12    8    7    6    5
```

The matrix is more informative if the rows and columns are named. The following use of `rownames()` and `colnames()` illustrates how to label the rows and columns of the `obstbl` matrix, respectively.



```

> rownames(obstbl) <- c("DI","BP")
> colnames(obstbl) <- c("A","B","C","D","E","F")
> obstbl
  A B C D E F
DI 34 22 14 13 12 5
BP 62 12 8 7 6 5

```

### 18.3.2 Chi-Square Test

The Chi-Square Test is performed with `chisq.test()`, which takes an observed frequency table either entered through `matrix()` or summarized with `xtabs()` as the first argument. The only other argument needed is `correct=FALSE` so that the continuity correction is not used.<sup>3</sup> The results of `chisq.test()` should be assigned to an object. The Chi-Square test statistic and p-value are extracted by simply printing the saved object. The expected frequency table is returned by appending `$expected` to the saved object.

### 18.3.3 Post-Hoc Analysis

Rejecting the null hypothesis in a Chi-Square Test indicates that there is some difference in the distribution of individuals into the levels of the response variable among some of the populations. However, rejecting the null hypothesis does not indicate which populations are different. In addition, as mentioned previously, confidence intervals are generally not performed with a Chi-Square Test. A post-hoc method for helping determine which populations differ is obtained by observing the Pearson residuals.

A Pearson residual is computed for each cell in the table as,

$$\frac{\text{Observed} - \text{Expected}}{\sqrt{\text{Expected}}}$$

which is the appropriately signed square root of the parts in the  $\chi^2$  test statistic calculation. Therefore, cells that have Pearson residuals far from zero contributed substantially to the large  $\chi^2$  test statistic that resulted in a small p-value and the ultimate rejection of  $H_0$ . Patterns in where the large Pearson residuals are found may allow one to qualitatively determine which populations differ and, thus, which levels of the response differ the most. This process will be illustrated more fully in the examples and review exercises. The Pearson residuals are obtained from the saved `chisq.test()` object by appending `$residuals`.

### 18.3.4 Example - Father Present at Birth

Consider this situation,

*Daniel Weiss (in “100% American”) reported the results of a survey of 300 first-time fathers from four different hospitals (labeled as A, B, C, and D). Each father was asked if he was present (or not) in the delivery room when his child was born. The results of the survey are in [FatherPresent.csv](#). Use these data to determine if there is a difference, at the 5% level, in the proportion of fathers present in the delivery room among the four hospitals.*

The 11-steps (Section 15.1) for a hypothesis test for this situation are as follows:

<sup>3</sup>The continuity correction is not used here simply so that the results using R will match hand-calculations. The continuity correction should usually be used.

1.  $\alpha=0.05$ .
2.  $H_0$ : “distribution of fathers presence (or not) during the birth of their child is the same for all four hospitals” vs.  $H_A$  : “the distribution of fathers presence during the birth of their child is NOT the same for all four hospitals.”
3. A Chi-Square Test is required because (i) a categorical variable (present or absent) was recorded and (ii) four populations were sampled (the hospitals).
4. The data appear to be part of an observational study with no clear indication of randomization (likely a voluntary response survey).
5. There are at least five individuals in each cell of the expected table (Table 18.5).
6. The statistic is the observed frequency table (Table 18.6).
7.  $\chi^2=5$  with 3 df (Table 18.7).
8. p-value=0.1718 (Table 18.7).
9.  $H_0$  is not rejected because the p-value is  $> \alpha$ .
10. There does not appear to be a significant difference between the distribution of father’s presence at their child’s birth and the hospital where that birth occurred. For comparative purposes, the row-percentage table is in Table 18.8.

### R Appendix:

```
fp <- read.csv("data/FatherPresent.csv")
( fp.obs <- xtabs(~hospital+father,data=fp) )
( fp.chi <- chisq.test(fp.obs) )
fp.chi$expected
percTable(fp.obs,margin=1,digits=1)
```

Table 18.5. Expected frequency table the Chi-Square Test for differences in father’s presence during child birth among four hospitals.

hospital	father	
	Absent	Present
A	15.25	59.75
B	15.25	59.75
C	15.25	59.75
D	15.25	59.75

Table 18.6. Observed frequency table the Chi-Square Test for differences in father’s presence during child birth among four hospitals.

hospital	father	
	Absent	Present
A	9	66
B	15	60
C	18	57
D	19	56

Table 18.7. Results from the Chi-Square Test for differences in father’s presence during child birth among four hospitals.

X-squared = 5.0003, df = 3, p-value = 0.1718

Table 18.8. Percentage of father's presence during child birth among four hospitals.

hospital	father		Sum
	Absent	Present	
A	12.0	88.0	100.0
B	20.0	80.0	100.0
C	24.0	76.0	100.0
D	25.3	74.7	100.0

### 18.3.5 Example - Apostle Islands Plants

Consider this situation,

*In her Senior Capstone project a Northland College student recorded the dominant (i.e., most abundant) plant species in 100 randomly selected plots on both Devil's Island and the Bayfield Peninsula (i.e., the mainland). There were a total of six "species" (one group was called "other") recorded (labeled as A, B, C, D, E, and F). The results are shown in the table below. Determine, at the 5% level, if the frequency of dominant species differs between the two locations.*

Location	Species						
	A	B	C	D	E	F	
DI	34	22	14	13	12	5	<b>100</b>
BP	62	12	8	7	6	5	<b>100</b>
	<b>96</b>	<b>34</b>	<b>22</b>	<b>20</b>	<b>18</b>	<b>10</b>	<b>200</b>

The 11-steps (Section 15.1) for a hypothesis test for the Northland College student's Senior Capstone example above is as follows

1.  $\alpha=0.05$ .
2.  $H_0$ : "the distribution of dominant plants species is the same between Devil's Island and the Bayfield Peninsula" vs.  $H_A$ : "the distribution of dominant plants species is NOT the same between Devil's Island and the Bayfield Peninsula."
3. A Chi-Square Test is required because (i) a categorical variable with six levels (plant species) was recorded and (ii) two populations were sampled (Devil's Island and Bayfield Peninsula).
4. The data appear to be part of an observational study where the plots were randomly selected.
5. There are more than five individuals in each cell of the expected table (Table 18.9).
6. The statistic is the observed frequency table given in the background.
7.  $\chi^2=16.54$  with 5 df (Table 18.10).
8. p-value=0.0055 (Table 18.10).
9.  $H_0$  is rejected because the p-value is  $< \alpha$ .
10. There does appear to be a significant difference in the distribution of the dominant plants between the two sites. A look at the Pearson residuals (Table 18.11) and the row-percentage table (Table 18.12) both suggest that the biggest difference between the two locations is due to "plant A."<sup>4</sup>

<sup>4</sup>When "Plant A" is removed from the observed table, the Chi-Square Test performed on the remaining plant species showed no difference in the distribution of the remaining plants between the two locations ( $p = 0.9239$ ). Thus, most of the difference in plant distributions between Devil's Island and the Bayfield Peninsula appears to be due primarily to "plant A" with more of "plant A" found on the Bayfield Peninsula than on Devil's Island.

**R Appendix:**

```
freq <- c(34,22,14,13,12,5,62,12,8,7,6,5)
ai.obs <- matrix(freq,nrow=2,byrow=TRUE)
rownames(ai.obs) <- c("DI","BP")
colnames(ai.obs) <- c("A","B","C","D","E","F")
( ai.chi <- chisq.test(ai.obs) )
ai.chi$expected
ai.chi$residuals
percTable(ai.obs,margin=1,digits=1)
ai.obs1 <- ai.obs[,-1]
( ai.chi1 <- chisq.test(ai.obs1) )
```

Table 18.9. Expected frequency table for the Chi-Square Test for differences in dominant species Devil's Island and Bayfield Peninsula.

	A	B	C	D	E	F
DI	48	17	11	10	9	5
BP	48	17	11	10	9	5

Table 18.10. Results from the Chi-Square Test for differences in dominant species Devil's Island and Bayfield Peninsula.

X-squared = 16.5442, df = 5, p-value = 0.00545





Table 18.11. Pearson residuals from the Chi-Square Test for differences in dominant species Devil's Island and Bayfield Peninsula.

	A	B	C	D	E	F
DI	-2.020726	1.212678	0.904534	0.9486833	1	0
BP	2.020726	-1.212678	-0.904534	-0.9486833	-1	0


Table 18.12. Percentage of dominant species within each location (Devil's Island and Bayfield Peninsula).


	A	B	C	D	E	F	Sum
DI	34	22	14	13	12	5	100
BP	62	12	8	7	6	5	100


## Review Exercises


- 18.8**  Saenz *et al.* (1998) examined the effectiveness of “restrictor plates” (a metal plate designed to reduce “pecking” by pileated woodpeckers (*Dryocopus pileatus*) in reducing damage by pileated woodpeckers) on cavity trees for red-cockaded woodpeckers (*Picoides borealis*) in Eastern Texas. For each red-cockaded woodpecker cavity hole they recorded whether the hole was fit with a restrictor plate or not and, ultimately, whether the cavity hole was damaged or not. The results of their study are recorded in [RestrictorPlates.csv](#). Examine these data to determine, at the 5% level, if restrictor plates reduced the damage done by pileated woodpeckers. [Answer](#)
- 18.9**  On the eastern slopes of the Rocky Mountains in Colorado, Wyoming, and Montana, whitetail deer (*Odocoileus virginianus*), mule deer (*Odocoileus hemionus*), and elk (*Cervus canadensis*) habitats overlap. It has been observed that in these areas where these species interact, diseases common to each species tend to infect more animals than in other areas. To examine this phenomenon, infection information on all three species was observed from individuals killed during the hunting seasons in areas where the habitats overlapped. In particular, it was recorded whether the animal was infected with one of the diseases common to each species or not. These data are recorded in [CervidDisease.csv](#). Test at the 1% significance level if there is a difference in the infection rate among the three species. [Answer](#)
- 18.10**  Ashland High School conducted a survey to determine if parents or students favored the idea of uniforms being required apparel for attending school (December 5, 1999, Ashland Daily Press). The surveys were administered to 223 parents at parent-teacher conferences and to 572 students by the Student Council. No other information about the surveys was given in the report. From these surveys it was learned that 70 parents and 101 students FAVORED the wearing of uniforms. Determine, at the 5% level, if there is a difference in the level of support for wearing uniforms between parents and students. [Answer](#)
- 18.11**  Five hundred patients participated in a comparison of the effectiveness of three arthritic pain relievers (175 received medication A, 150 received medication B, and 175 received medication C). Each patient used one of the three medications for one month and then was asked if the product was effective. The results showed 115 patients using medication A, 78 patients using medication B, and 140 patients using medication C said their medication was effective. Test, at the 10% level, if there is a difference in effectiveness among the three medications. [Answer](#)
- 18.12**  USA Today presented two sets of data on why Americans don’t exercise. One set was for 1000 randomly selected men. The other was for 1000 randomly selected women. The results of the surveys are recorded in [Exercise.csv](#). Determine, at the 1% level, if the distribution of men and women differs among the six responses given. [Answer](#)
- 18.13**  Fairley *et al.* (1994) gave the results in the table below concerning the age and the number who were positive for human papillomavirus infection among the 290 participants in their study. Test the hypothesis, at the 5% level, that the same proportion for each age-group is HPV-positive. [Answer](#)

Age	n	HP+
under 20	27	11
21-25	81	30
26-30	108	34
31-35	74	18

- 18.14**  Passengers aboard the RMS Titanic were classified as to their “class” (first, second, third, or crew) and whether or not they survived the wreck (yes or no). Use the data found in [Titanic.csv](#) to determine if there was a difference, at the 1% level, in the survival rate among the classes of passengers. [Answer](#)


- 18.15**  Meliker *et al.* (2004) examined the records of 773 motor-vehicle crashes in southeastern Michigan. Of these, 139 had a driver with a blood alcohol level greater than 0.10% and were defined as alcohol-related crashes. Of these alcohol-related drivers, 79% were male, while 56% of the non-alcohol-related drivers were male. Use this information to determine, at the 1%, if males are more or less likely to be involved in a an alcohol-related crash than females. [HINT: I'd construct a 2x2 contingency table (Section ??) of these results with the response variable as columns. Note that the results as presented above are in column-percentage format and the results needed to answer the question are row-percentage format. Also, note that the column totals are given indirectly in the information above but the row totals need to be determined.] [Answer](#)

- 18.16**  Shrimp trawlers are required to have turtle exclusion device (TED), that allows most loggerhead sea turtles (*Caretta caretta*) to escape the net, thus reducing turtle mortality due to by-catch. In the Gulf of Mexico, the TEDs were originally required to be 32" x10" but a new law now requires to TEDs to be 71" x26" with the thought that turtle mortality would be further reduced by the larger opening. This thought was examined by recording the number of trawl tows that had at least one turtle mortality. In 75 tows with the original smaller opening there were 16 tows with at least one turtle mortality. In contrast, in 88 tows with the the newer larger opening there were 8 tows with at least one turtle mortality. Test at the 10% level if there is a significant difference in the proportion of trawl tows with at least one turtle mortality between trawls with the different sized openings. [Answer](#)

- 18.17**  Researchers observed groups of dolphins off the coast of Iceland near Keflavik in 1998<sup>5</sup>. The researchers recorded the time of the day (“Morning”, “Noon”, “Afternoon”, and “Evenings”) and the main activity of the group, whether travelling quickly (“Travel”), feeding (“Feed”), or socializing (“Social”). The number of dolphin groups observed by each time of day and activity is shown in the table below. Use this information to determine, at the 5% level if the proportion of groups exhibiting each activity differs by time of day.

[Answer](#)

Time of Day	Activity		
	Travel	Feed	Social
Morning	6	28	38
Noon	6	4	5
Afternoon	14	0	9
Evening	13	56	10

- 18.18**  The data in Zoo1.csv contains a list of animals found in several different zoos. In addition, each animal was classified into broad “type” categories (“mammal”, “bird”, and “amph/rep”). The researchers that collected these data wanted to examine if the distribution of broad animal types differed among zoos. Test the researcher’s question at the 5% level [Answer](#)

<sup>5</sup>Data was originally from [here](#).