Chapter 8
Chi-Square Test

Previous chapters have presented information on sampling distributions, Central Limit Theorem, confidence intervals, TYPE I error, TYPE II error, and hypothesis testing. This information is useful in understanding how sample statistics are used to test differences between population parameters. The statistical tests presented in this and subsequent chapters depend upon the level of measurement and type of research design.

A popular statistic for testing research questions involving categorical data is the chi-square test statistic. The chi-square statistic was developed by Karl Pearson to test whether two categorical variables were independent of each other. A typical research question involving two categorical variables can be stated as, "Is drinking alcoholic beverages independent of smoking cigarettes?" A researcher would gather data on both variables in a "yes-no" format, then cross tabulate the data. The cross-tabulation of the data for this research question would look like the following:

	Do you drink alcoholic beverages?	
Do you smoke cigarettes?	Yes	No
Yes		
No		

Individuals would be asked both questions and their separate responses recorded. The cross-tabulation of the data would permit an indication of the number of people who *did* smoke cigarettes and *did* drink alcoholic beverages, the number of people who *did* smoke cigarettes and *did not* drink alcoholic beverages, the number of people who *did not* smoke cigarettes and *did* drink alcoholic beverages, and the number of people who *did not* smoke cigarettes and *did not* drink alcohol. Consequently, four possible outcomes are represented by the cross-tabulation of the yes/no responses to the two questions.

R. Schumacker and S. Tomek, *Understanding Statistics Using R*,
DOI 10.1007/978-1-4614-6227-9_8, © Springer Science+Business Media New York 2013

The **chi-square statistic** is computed by taking the sum of the observed frequency minus the expected frequency squared divided by the expected frequency in each of the four cells. The chi-square formula is expressed as:

$$\chi^2 = \sum \left(\frac{(O-E)^2}{E} \right)$$

Multiplying the respective row and column sums and dividing by the total number of individuals yields the expected frequencies in each of the four cells. The calculation of the difference between what is observed and what is expected by chance alone forms the basis for the test of independence between two categorical variables. The *expected cell frequencies* are based on the two categorical variables being independent. An example will help to illustrate how to calculate the expected frequencies and the chi-square statistic.

A school district is interested in having parents pass a bond referendum to build a new high school. The superintendent decides to conduct a preliminary poll of the voters to see if they might favor passing the bond. The superintendent is also concerned about whether men and women would vote differently on the bond referendum. Consequently, 200 parents (100 men and 100 women) in the district were randomly selected and telephoned to collect the data. Each parent was asked their gender and whether they favor or oppose a bond to build a new high school. The responses are cross-tabulated below.

Gender	Favor	Oppose	Total
Men	40 (60)	60 (40)	100
Women	80 (60)	20 (40)	100
Totals	120	80	200

The observed values indicate 40 out of 100 (40%) men and 80 out of 100 (80%) women are in favor of the bond, while 60 out of 100 (60%) men and 20 out of 100 (20%) women are opposed to the bond. If the null hypothesis is true (no difference in the percent between men and women in favor of the bond), then we would expect the percentages to be the same for both men and women, i.e., of the 120 observed in favor, one-half or 60 individuals would be expected in each gender cell. The expected cell frequencies are what would be expected if gender and voting were independent. The most convenient way to calculate the expected cell values is to multiply the corresponding row and column sums and divide by the total sample size. For men, the first expected cell value is: $(100 \times 120)/200 = 60$. The other expected cell value is: $(100 \times 80)/200 = 40$. For women, the first expected cell value is: $(100 \times 120)/200 = 60$. The other expected cell value is: $(100 \times 80)/200 = 40$. The expected cell values are in parentheses in the table. The expected cell values should always add up to the total for each row and/or column, respectively.

The chi-square statistic compares the corresponding observed and expected values in the cells of the table under the assumption that the categorical variables are independent, i.e., the null hypothesis is true. If the row and the column variables are independent, then the proportion observed in each cell should be similar to the

proportion expected in each cell. Is the difference between what we observed and expected in the four cells statistically different or due to random chance (expected values)? A decision about the null hypothesis is made on the basis of the chi-square statistic, which is computed as follows:

$$\chi^2 = \sum\left(\frac{(40-60)^2}{60} + \frac{(60-40)^2}{40} + \frac{(80-60)^2}{60} + \frac{(20-40)^2}{40}\right)$$

$$\chi^2 = \sum(6.67 + 10 + 6.67 + 10) = 33.34$$

The computed chi-square value is compared to a tabled chi-square value in the appendix for a given degree of freedom. The degrees of freedom are always determined by the number of rows minus one $(r-1)$ times the number of columns minus one $(c-1)$. This can be expressed as: $df=(r-1)(c-1)$. Since there are two rows and two columns, the degree of freedom is: $df=(2-1)(2-1)=1$. The tabled chi-square value for $df=1$ and a .05 level of significance is 3.84. Since the computed chi-square value of 33.34 is greater than the tabled chi-square value of 3.84, we reject the null hypothesis in favor of the alternative hypothesis that men and women differ in the percent favoring a bond for building a new school.

The chi-square value is computed over all the cells and therefore a significant chi-square doesn't specify which cells may have contributed to the significance of the overall chi-square value. Our interpretation of the overall chi-square result is greatly enhanced by realizing that each cell value is itself a chi-square value! Consequently, we can interpret each cell value individually and compare it to the tabled chi-square value of 3.84 with $df=1$. Since each cell value is greater than 3.84 (6.67, 10, 6.67, and 10), we would conclude that each cross-tabulated cell significantly contributed to the overall chi-square. Also, each expected cell frequency should be greater than five to meet the assumption for computing the chi-square statistic.

Another helpful approach to interpreting chi-square results is to take each individual chi-square value (cell value) as a percent of the overall chi-square value. This provides a variance accounted for interpretation. In our example, $6.67/33.34=20\%$, $10/33.34=30\%$, $6.67/33.34=20\%$, and $10/33.34=30\%$. The sum of these cell percents must always equal 100%. Our interpretation would then be based on which cell or cells contributed the most to the overall chi-square.

The chi-square statistic will be small if there are small differences between the observed and the expected values, and it will be large if there large differences. The chi-square statistic for a two-by-two table is distributed as a theoretical chi-square sampling distribution with 1 degree of freedom. Therefore, the theoretical chi-square distribution can be used to determine the region of rejection for the null hypothesis. The region of rejection includes any chi-square statistic greater than the 1-α percentile of the theoretical chi-square distribution with degree of freedom $=(r-1)(c-1)$. Consequently, the chi-square test of independence can be performed on tables of any dimension, i.e., varying numbers of rows and columns for categorical variables.

The chi-square statistic for two-by-two tables only are discussed. You will determine if the row and column values are independent or dependent. If the rows and columns are independent, you will be able to observe the variability in the observed values that occur because of random sampling. A TYPE I error will occur when the rows and columns are independent, i.e., the null hypothesis will be rejected when it is true. If the rows and columns are dependent, you will be able to observe TYPE II errors, i.e., the null hypothesis is retained. As in previous programs, the probability of a TYPE II error will depend on how much the true situation differs from the null hypothesis. If the rows and columns are almost independent, then the probability of a TYPE II error will be high.

A chi-square statistic can be used to test research questions involving cross-tabulated categorical variables. An overall chi-square statistic is computed by summing the individual cell values (chi-squares) in a cross-tabulated table. The degrees of freedom for a cross-tabulated table are row minus one times column minus one, i.e., $df = (r-1)(c-1)$. The chi-square test of independence can be used for any number of rows and columns, as long as the expected cell frequency is greater than five. A chi-square test of independence is used to determine whether or not the rows and columns are independent (null hypothesis). If the null hypothesis is true, it is still possible that the chi-square test could lead to a rejection of the null hypothesis (TYPE I error). If the null hypothesis is false, it is still possible that the chi-square test could lead to retaining the null hypothesis (TYPE II error). The ratio of each cell value to the overall chi-square value provides a variance accounted for interpretation of how much each cell contributed to the overall chi-square value. The chi-square table of expected values in the Appendix permits testing whether the computed chi-square value occurs beyond a chance level of probability.

CROSSTAB R Program

The CROSSTAB program inputs four percents for the true population and the sample size. The program will then select a random sample of size N for this population. The percents will be printed in a table along with the observed and expected values. A chi-square statistic will be computed and printed along with the degrees of freedom and probability value. You will make a decision about whether to retain or reject the null hypothesis based on the probability value being $p < .05$. A second example uses large percent differences between the observed and expected values to better understand the magnitude of the chi-square value. You may want to examine each individual cell chi-square value for significance and hand calculate the percent contribution to the overall chi-square value.

The CROSSTAB program uses matrices to represent cross-tabulated categorical variable tables. The program first defines the true proportions within each cell (the numbers are arranged by the first row followed by the second row) and then the sample size. Next, random data are generated from a discrete population of data from 1 to 4, simulating categorization into a cell of the table. The vector of data is factored

and tabled in order to get counts for each outcome and then placed into a matrix. The table of expected values is built from the sample data by taking the sum of the values for column 1 and dividing it by the sample size to determine the proportion of total outcomes that fall within the first column, then multiplying it by the total number of responses for row 1. This gives the expected value for cell (1,1). The process is repeated for the other three cells and the results are fitted into a matrix and rounded to three digits. Three matrices are then printed: True Population Proportions, Observed Proportions, and Expected Proportions. Finally, the function **chisq.test** is performed on the *testMatrix* object in order to do the actual chi-square test. The chi-square value, degrees of freedom, and p-value are printed by default. The **chisq.test** function uses the Yates correction, which uses real limits around numbers, for example, real limits of 5 is 4.5–5.5.

CROSSTAB Program Output

Example 1

```
Cell Probabilities=0.1 0.15 0.3 0.45
Sample Size=100

Population Proportions
      X       Y
A  0.10    0.15
B  0.30    0.45

Observed Proportions
      X       Y
A  0.10    0.17
B  0.25    0.48

Expected Proportions
       X      Y
A  0.094  0.176
B  0.255  0.475

Pearson chi-square test with Yates' continuity correction
Chi-square = 6e-04 df = 1 p-value = 0.98116
```

Example 2

```
Cell Probabilities=0.7 0.1 0.1 0.1
Sample Size=100

Population Proportions
      X       Y
A  0.70    0.10
B  0.10    0.10
```

```
Observed Proportions
      X       Y
A    0.7     0.1
B    0.1     0.1

Expected Proportions
      X       Y
A   0.64    0.16
B   0.16    0.04

Pearson chi-square test with Yates' continuity correction
Chi-square=11.8164 df = 1 p-value = 0.00059
```

Chi-Square Exercises

1. Run CROSSTAB using the Population proportions below with a sample size of 100 using .05 level of significance.
   Population Proportions

.50	.20
.10	.20

   Observed Proportions

   Expected Proportions

   a. Are the expected values in the cells computed correctly? YES____ NO____
   b. Is the chi-square statistic computed correctly? YES____ NO____
   c. What is the chi-square value? _____
   d. What is the statistical decision? Retain Null ____ Reject Null____
   e. Does the decision agree with the true percent? YES____ NO____
   f. What percent of the time would you expect the null hypothesis to be rejected by mistake? _____
   g. What is the name of this type of error? _____

2. Run CROSSTAB using the Population proportions below with a sample size of 100 using a .01 level of significance.
   Population Proportions

.16	.32
.28	.24

   Observed Proportions

   Expected Proportions

a. Are the expected values in the cells computed correctly? YES____ NO____
b. Is the chi-square statistic computed correctly? YES____ NO____
c. What is the chi-square value? _____
d. What is the statistical decision? Retain Null ____ Reject Null_____
e. Does the decision agree with the true percent? YES____ NO____
f. What percent of the time would you expect the null hypothesis to be rejected by mistake? _____
g. What is the name of this type of error? _____

True or False Questions

Chi-Square

T F a. A chi-square statistic is used with ordinal data.
T F b. The null hypothesis in the chi-square test corresponds to no difference in the row and column categories.
T F c. The chi-square statistic will be large if there is a large difference between the observed and the expected values in the cells.
T F d. If the true population has independent rows and column, then a TYPE I error may occur.
T F e. The chi-square statistic can be negative.
T F f. The overall chi-square value indicates which cells contributed to a statistically significant finding.
T F g. The tabled chi-square value is 3.84 with one degree of freedom at the .05 level of significance.