Chapter 13
Categorical Data Analysis

13.1 Introduction

Chapter 13 introduces the topic of categorical data analysis to the reader. Both the one-way and two-way analyses are discussed in the text. **XLSTAT** allows the user to work with both of these analyses. We note that the *A First Course in Statistics* text covers this material in Chapter 8 of that text.

The following examples from *Statistics* are solved with **XLSTAT** in this chapter:

Excel Companion Exercise	Page	Statistics Example	Excel File Name
13.1	186	Example 13.2	MARIJUANA
13.2	189	Example 13.3	MARREL

13.2 Testing Categorical Probabilities: One-Way Table

The one-way analysis of categorical probabilities allows the user to compare observed counts of the levels of a qualitative variable against some hypothesized probabilities. The user has the flexibility of hypothesizing equal probabilities or individual probabilities, provided that they all sum to the value one. **XLSTAT** requires the user to create a data set that contains three columns – one that contains the levels of the variable being tested, a second that contains the observed counts for these levels, and a third that contains the hypothesized proportions (in decimal form) for these levels. We illustrate with the following example.

Exercise 13.1 Use Eexample 13.2 found in the *Statistics* text.

Problem: Suppose an educational television station has broadcast a series of programs on the physiological and psychological effects of smoking marijuana. Now that the series is finished, the station wants to see whether the citizens within the viewing area have changed their minds about how the possession of marihuana should be considered legally. Before the series was shown, it was determined that 7% of the citizens favored legalization, 18% favored decriminalization, 65% favored the existing law (an offender could be fined or imprisoned), and 10% had no opinion.

A summary of the opinions (after the series was shown) of a random sample of 500 people in the viewing area is given in Table 13.1 (and saved in the MARIJUANA data file). Test at the $\alpha = .01$ level to see whether these data indicate that the distribution of opinions differs significantly from the proportions that existed before the educational series was aired.

Table 13.1

Legalization	Decriminalization	Existing Laws	No Opinion
39	99	336	26

Solution:

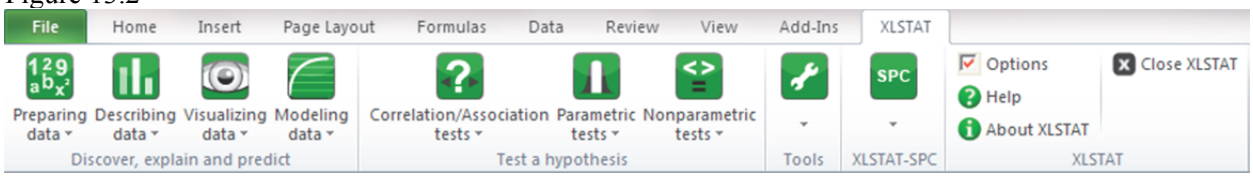
We solve Exercise 13.1 utilizing the **Multinomial goodness of fit test** presented in **XLSTAT**. **Open** the data file **MARIJUANA** by following the directions found in the preface of this manual. If done correctly, the data should appear in a workbook similar to that shown in Figure 13.1. Please note that we added the column, Proportion, and included the specified proportions stated in the previous example.

Figure 13.1

	A	B	C
1	COUNT	OPINION	Proportion
2	39	LEGAL	0.07
3	99	DECRIM	0.18
4	336	EXISTLAW	0.65
5	26	NONE	0.1

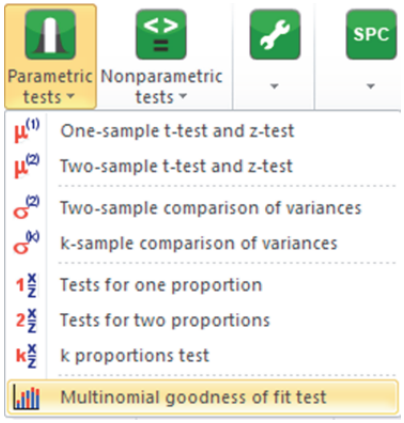
To conduct the desired analysis, we click on the **XLSTAT** tab at the top of the **Excel** workbook to access the **XLSTAT** menus shown in Figure 13.2.

Figure 13.2



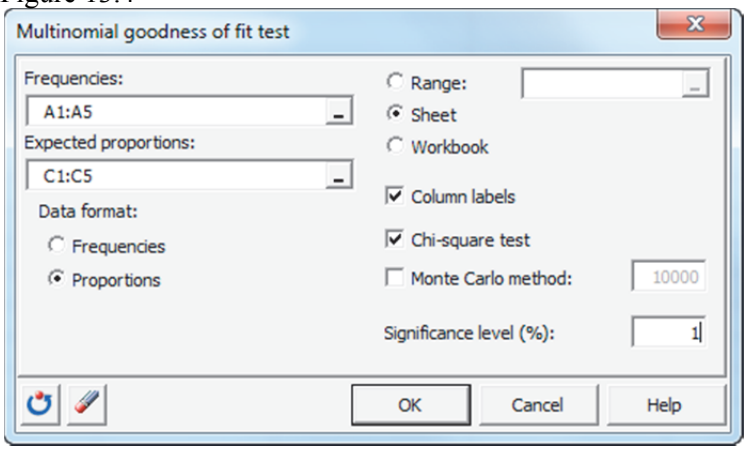
To conduct the desired test, we click on the **Parametric tests** menu and select the **Multinomial goodness of fit test** option shown in Figure 13.3.

Figure 13.3



This opens the **Multinomial goodness of fit test** menu shown in Figure 13.4. We need to first specify the location of the data that is to be analyzed. In our data set, the data counts are located in column A, rows 2 – 5, with row 1 being the variable label. The hypothesized proportions are located in column C, rows 2-5. We specify the column **A** data in the **Frequencies** box and the column **C** data in the **Expected proportions** box. We make sure that the **Proportions** button is selected and that the **Column labels** box is checked. We enter the **Significance Level (%)** of **1**%. To conduct the one-way analysis, we check the **Chi-square test** box and click **OK**.

Figure 13.4



The XLSTAT output is shown in Figure 13.5.

Figure 13.5

Chi-square test:	
Chi-square (Observed value)	13.2495
Chi-square (Critical value)	11.3449
DF	3
p-value	0.0041
alpha	0.01

Test interpretation:
H0: The distribution is not different from what is expected.
Ha: The distribution is different from what is expected.
As the computed p-value is lower than the significance level alpha=0.01, Ha.
one should reject the null hypothesis H0, and accept the alternative hypothesis
The risk to reject the null hypothesis H0 while it is true is lower than 0.41%.

We compare the test statistic of 13.2495 and the p-value of p = .0041 to the corresponding values listed in the text. We see they are identical.

13.3 Testing Categorical Probabilities: Two-Way Table

Chapter 2 introduced the reader to the idea of presenting descriptive results of collected data in a tabular form. In Chapter 13, we now take a look at a technique that allows us to determine if the outcomes of these two variables are dependent upon one another. The two-way analysis of data is available in **XLSTAT** through using the **Contingency table (descriptive statistics)** menu found in the **Describing data** option of the **XLSTAT** menu. We illustrate using the following exercise.

Exercise 13.2 Use Eexample 13.3 found in the *Statistics* text.

Problem: A social scientist wants to determine whether the marital status (divorced or not divorced) of U.S. men is independent of their religious affiliation (or lack there-of). A sample of 500 U.S. men is surveyed, and the results are tabulated and shown in table 13.2 (and saved in the MARREL data set.

Table 13.2

	Religion					
Status	**A**	**B**	**C**	**D**	**None**	**Total**
Divorced	39	19	12	28	18	116
Never	172	61	44	70	37	384
Total	211	80	56	98	55	500

a. Test to see whether there is sufficient evidence to indicate that the marital status of men who have been or are currently married is dependent on religious affiliation. Take $\alpha = .10$.

Solution:

We solve Exercise 13.2 utilizing the **Contingency table (descriptive statistics)** menu presented in **XLSTAT**. **Open** the data file **MARREL** by following the directions found in the preface of this manual. If done correctly, the data should appear in a workbook similar to that shown in Figure 13.6. Please note that we arranged this data into the contingency table shown in Figure 13.7.

Figure 13.6

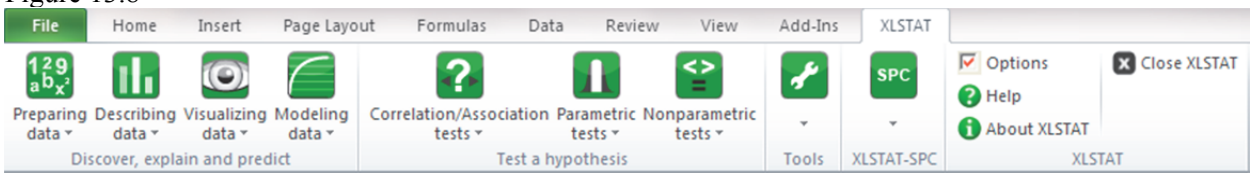
	A	B	C
1	MARITAL	RELIGION	NUMBER
2	DIVORCED	A	39
3	NEVER	A	172
4	DIVORCED	B	19
5	NEVER	B	61
6	DIVORCED	C	12
7	NEVER	C	44
8	DIVORCED	D	28
9	NEVER	D	70
10	DIVORCED	NONE	18
11	NEVER	NONE	37

Figure 13.7

	Religion					
Status	**A**	**B**	**C**	**D**	**None**	**Total**
Divorced	39	19	12	28	18	116
Never	172	61	44	70	37	384
Total	211	80	56	98	55	500

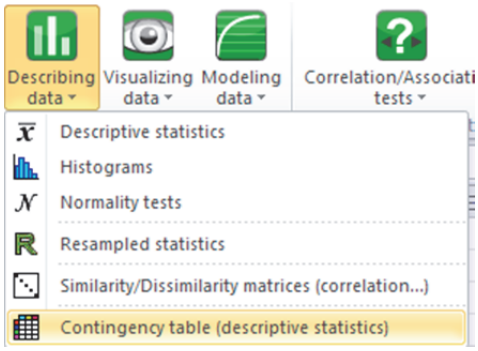
To conduct the desired analysis, we click on the **XLSTAT** tab at the top of the **Excel** workbook to access the **XLSTAT** menus shown in Figure 13.8.

Figure 13.8



We click on the **Describing data** menu and select the **Contingency table (descriptive statistics)** option shown in Figure 13.9.

Figure 13.9



This opens the **Contingency table** menu shown in Figures 13.10 – 13.12. We first specify the location of the data that is to be analyzed by clicking on the **General** tab at the top of the menu. In our data set, the contingency table data is located in rows 4 and 5 of columns H thru L. Labels are located in column G and in row 3 of the worksheet. We specify the **Contingency table** location and make sure that the **Labels included** box is checked.

Figure 13.10

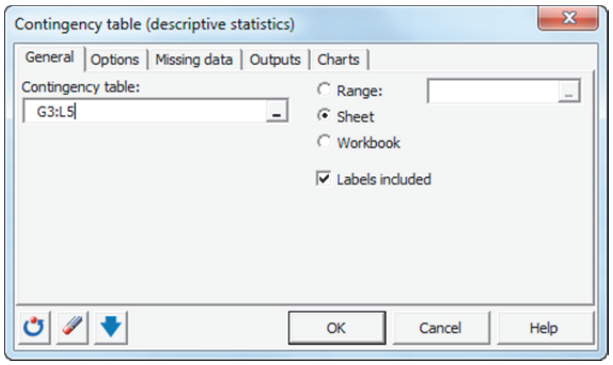
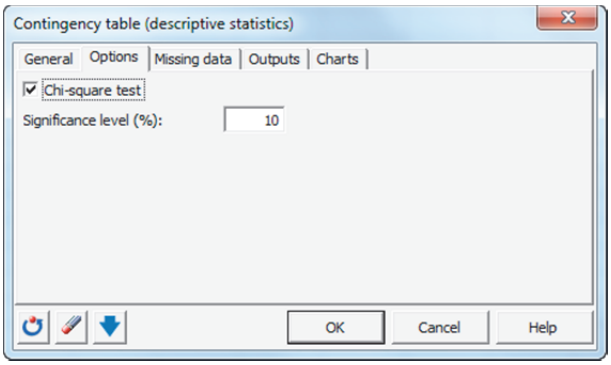
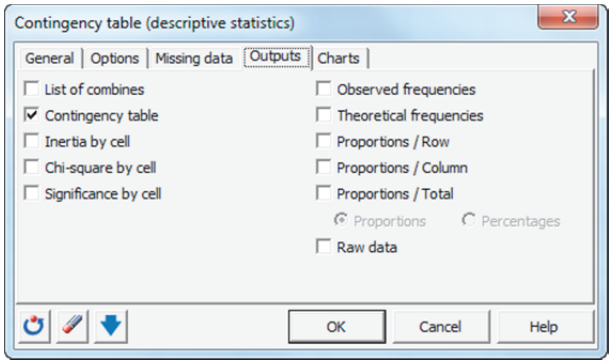


Figure 13.11



Click on the **Options** tab at the top of the menu to request the chi-square test and to specify the **Significance level (%)** of the test. We check the box and enter **10%** in the significance level box.

Figure 13.12



Click on the **Outputs** tab at the top of the menu to request a variety of output for the test. To generate the test statistic and p-value for the test, we **check** the **Contingency table** box and click **OK**. The **XLSTAT** output is shown in Figure 13.13.

Figure 13.13

Test of independence between the rows and the columns (Chi-square):

Chi-square (Observed value)	7.1355
Chi-square (Critical value)	7.7794
DF	4
p-value	0.1289
alpha	0.1

Test interpretation:
H0: The rows and the columns of the table are independent.
Ha: There is a link between the rows and the columns of the table.
As the computed p-value is greater than the significance level alpha=0.1,
one cannot reject the null hypothesis H0.
The risk to reject the null hypothesis H0 while it is true is 12.89%.

We compare the test statistic of 7.1355 and the p-value of p = .1289 to the corresponding values listed in the text. We see they are identical.

13.4 Technology Lab

The Technology Lab consists of problems for the student to practice the techniques presented in each lesson. Each problem is taken from the homework exercises within the *Statistics* text and includes an **Excel** data set (when applicable) that should be used to create the desired output. The completed output has been included with each problem so that the student can verify that he/she is generating the correct output.

1. **Detecting Alzheimer's disease at an early age.** Geneticists at Australian National University are studying whether the cognitive effects of Alzheimer's disease can be detected at an early age (*Neuropsychology*. Jan. 2007). One portion of the study focused on a particular strand of DNA extracted from each in a sample of 2,097 young adults between the ages of 20 and 24. The DNA strand was classified into one of three genotypes: $E4^+/E4^+$, $E4^+/E4^-$, and $E4^-/E4^-$. The number of young adults who are not afflicted with Alzheimer's disease, the distribution of genotypes for this strand of DNA is 2% with $E4^+/E4^+$, 25% with : $E4^+/E4^-$, and 73% with: $E4^-/E4^-$. If differences in this distribution are detected, then this strand of DNA could lead researchers to an early test for the onset of Alzheimer's. Conduct the test (at $\alpha = .05$) to determine if the distribution of E4/E4 genotypes for the population of young adults differs from the norm.

Genotype:	$E4^+/E4^+$	$E4^+/E4^-$	$E4^-/E4^-$
Number of young adults	56	517	1524

XLSTAT Output

Chi-square test:	
Chi-square (Observed value)	4.8440
Chi-square (Critical value)	5.9915
DF	2
p-value	0.0887
alpha	0.05

Test interpretation:
H0: The distribution is not different from what is expected.
Ha: The distribution is different from what is expected.
As the computed p-value is greater than the significance level alpha=0.05, one cannot reject the null hypothesis H0.
The risk to reject the null hypothesis H0 while it is true is 8.87%.

2. **Pig farmer study.** An article in *Sociological Methods & Research* (May, 2001) analyzed the data presented in the table. A sample of 262 Kansas pig farmers were classified according to their education level (college or not) and size of their pig farm (number of pigs). Conduct a test to determine whether a pig farmer's education level has an impact on the size of the pig farm. Use $\alpha = .05$ and support your answer with a graph.

	Education Level		
Farm Size	**No College**	**College**	**TOTALS**
<1,000 pigs	42	53	135
1,000-2,000 pigs	27	42	613
2,000-5,000 pigs	22	20	42
>5,000 pigs	27	29	256
TOTALS	118	134	262

XLSTAT Output

Test of independence between the rows and the columns (Chi-square):	
Chi-square (Observed value)	2.1422
Chi-square (Critical value)	7.8147
DF	3
p-value	0.5434
alpha	0.05

Test interpretation:
H0: The rows and the columns of the table are independent.
Ha: There is a link between the rows and the columns of the table.
As the computed p-value is greater than the significance level alpha=0.05, one cannot reject the null hypothesis H0.
The risk to reject the null hypothesis H0 while it is true is 54.34%.