Simulating Chi-Square Test Using Excel

Leslie Chandrakantha
John Jay College of Criminal Justice of CUNY
Mathematics and Computer Science Department
524 West 59th Street, New York, NY 10019
lchandra@jjay.cuny.edu

Abstract

Students in introductory statistics classes struggle to grasp the basic concepts. We illustrate the use of Excel's Data Table function and standard formulas to perform the Chi-square test for independence. The Chi-square test is an integral part in introductory statistics. Simulation using Excel is used to generate many random samples and calculate the *p*-value of the test. The empirical distribution of the statistic is also tabulated. This approach will improve the students' ability to understand the meaning of the *p*-value to interpret the results of hypothesis testing.

1. Introduction

Many college students are required to take at least one statistics course depending on their major field. The Introductory statistics course is the only statistics course they take or the first of a sequence of courses. Fundamental statistical concepts such as sampling distributions, central limit theorem, confidence intervals, hypothesis testing, and *p*-values are very important in an introductory statistics courses. Many students have difficulties in understanding these topics. The use of computers to mimic the real life sampling or repeated sampling from a population helps to understand these concepts. Almost all software packages offer ways to perform the simulation. Many introductory statistics students do not have necessary skills to write macros to perform these tasks. Excel provides ways to accomplish the same task without writing macros. We show how to use Excel Data Table facility, standard formulas, and repeated simulated sampling to teach the Chi-square test for independence in an introductory statistics class. Excel's availability to students and the ease of presenting the situation in multiple rows and columns are advantageous.

Use of computer simulation methods is becoming very popular in teaching more difficult concepts in introductory statistics courses. Decades ago, computer simulation is used in upper level statistics courses. Computer simulation gives clear and visible justification of

the concept and that can be more convincing to students. Cobb (1994) noted that incorporating computer simulation methods to illustrate key concepts and allow students to discover important principles themselves will enhance their knowledge. Use of Excel Data Tables for simulation is very straightforward for students. In this paper, we describe how to use the Excel Data Tables to generate many different random samples, calculate the value of the test statistics, and the corresponding *p*-value in making the correct conclusion.

In next section, we give an introduction to Excel Data Table facility, and how use it to generate many different samples. The following section gives an overview of the Chi-square test for independence using an example. The next ext section shows the simulation of the test and the *p*-value. We finish the paper with tabulating the empirical distribution of the Chi-square statistic, some future work, and concluding remarks.

2. Excel Data Tables

Excel Data Table function allows a table of "what if" questions to be posed and answered simply in sensitivity analysis, and is useful in simulation. Christie (2004) has used the Excel Data Tables for estimating the population mean and correlation. Winston [2007] explained how to use Excel Data tables to simulate stock prices in asset allocation models. A valuable introduction to Excel Data Tables is given by Ecklund [2009].

The Data Table function can be accessed from menu bar **Data > What IF Analysis > Data Tables** in Excel 2007 and 2010. The *Figure 1* shows a simple Excel Data Table setup:

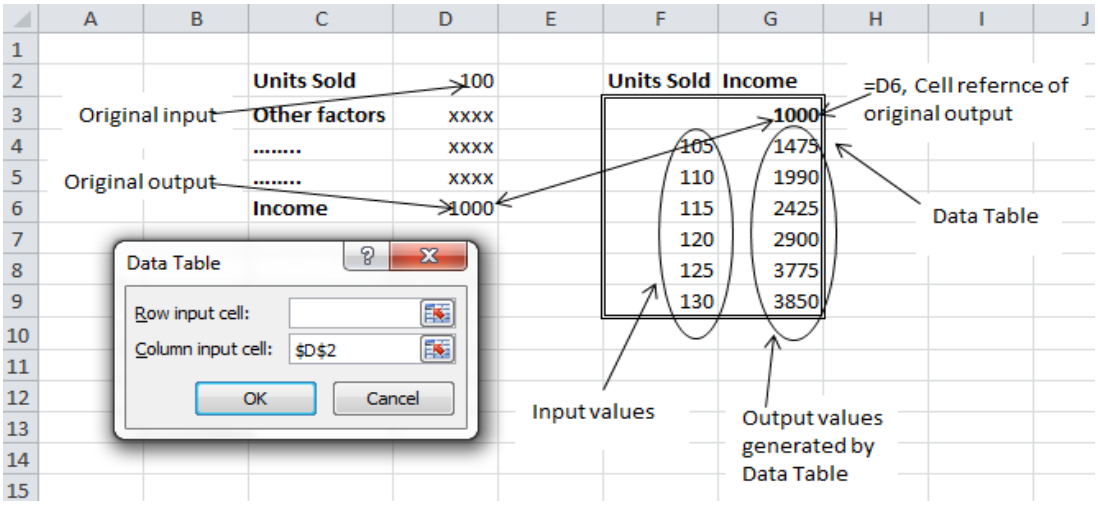

Figure 1: Data Table Setup

In this setup, it calculates values of income (output) for different units sold values (input). This way is more convenient than typing the formula or copying the formula in each output cell.

To generate the values of a statistic for different samples using Data Table, first we calculate the value of the statistic using a random sample. This can be done using Excel random number generating functions for the appropriate population and other standard functions. This statistic value will be our original input value. This value (formula) will be put in the top cell of the right column of the Data Table. We set up our Data Table by selecting two columns and a certain number of rows depending on the number of values of the statistic we need to calculate. Leave the left column blank. The menu bar **Data > What IF Analysis > Data Tables** gives the Data Table dialog box. In this dialog box, leave Row input cell blank and type an empty cell reference that has no part of this Data Table setup for the Column input cell. Excel generates a new sample and computes the value of the statistic for each substitution of this empty input cell and fills the table. Copying the formula down in the output cells does not work in this case. If we do this manually, we need to recalculate the statistic by repeated sampling by pressing the **F9** key and recording these values in a column. The Data Table function provides a convenient way of generating values of the statistic for different samples.

3. Overview of Chi-square Test using an Example

The Chi-square test is used to determine whether there is an association between two categorical variables. Each subject in the sample is classified on this two variables and that is presented in a contingency table where rows correspond to one categorical variable and columns correspond to second categorical variable. The null hypothesis is that the variables are not associated, in other words, they are independent. The alternative hypothesis is that the variables are associated, or dependent.

We consider the following problem from elementary statistics text "Essentials of Statistics for the Behavioral Sciences" by Gravetter & Wallnau (2011). I have used this book as my official text book for my introductory statistics course in the past.

Example: Researcher has demonstrated strong gender differences in teenagers' approaches to dealing with mental health issues. In a typical study, eight-grade students are asked to report their willingness to use mental health services in the event they were experiencing emotional or other mental health problems. The data for a sample of 150 students are shown in the following contingency table. Do the data show a

significant relationship between gender and willingness to seek mental health assistance?

In the following contingency table, gender is the row variable and the willingness to use the mental health services is the column variable.

Willingness to Use Mental Health Services

	Probably No	Maybe	Probably Yes
Males	17	32	11
Females	13	43	34

The null and alternative hypotheses are defined as follows:
H_o: In general population, gender and willingness to use mental health services are independent (no relationship).
H_1: In general population, gender and willingness to use mental health services are dependent (relationship).
The Chi-square test for independence uses the following test statistic:

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

where *O* represents the observed counts from the sample and *E* represents the expected counts for each cell assuming the null hypothesis is true. This statistic has approximate Chi-square distribution with degrees of freedom, *df* = (*r*-1)(*c*-1), where *r* is the number of rows and *c* is the number of columns in the contingency table.

The observed counts, *O*, are the counts recorded for each cell from the sample data. If the gender and willingness to use the health services are **independent**, we would expect 40% of the **yes** group to be males, 40% of **may be** group to be males, and 40% of **no** group to be males.

The expected counts, *E*, are the counts expected in each cell assuming null hypothesis. They are calculated using the formula, *E = (row total × column total)/grand total,*

where the row total is the total for each row, the column total is the total for each column, and the grand total is the total counts for the entire table. Assuming null hypothesis, two variables are independent, the expected counts for the table are calculated and shown below.

Willingness to use Mental Health Services

	Probably No	Maybe	Probably Yes
Males	12	30	18
Females	18	45	27

The value of the test statistic is calculated to be **8.23**. Now we use the Excel Data Table facility to generate different random samples to compute the proportion of the test statistic values that are as extreme or more extreme than the observed value **8.23.**

This simulation approach gives students better understanding about the *p*-value. The definition of the *p*-value is "*The probability, assuming the null hypothesis is true, that the test statistics would take a value as extreme or more extreme than that actually observed*" (Moore & McCabe 2003).

4. Simulation using Excel Data Tables

Now we show how to generate different random samples for the contingency table assuming the null hypothesis is true. We generate independent Binomial random variables for each column variable with row proportions for males and females as 0.4 and 0.6 respectively. These are the proportions we had in the original contingency table. These proportions result from the null hypothesis assumption.

The Excel function **BINOM.INV** is used to generate independent Binomial random variables. The earlier versions of Excel (2007 and prior) used **CRITBINOM** function for this purpose. The syntax of the function is **BINOM.INV(n, p, s)** where n is the number of trials, p is the success probability and s is the criterion value. This function returns the smallest value for which the cumulative Binomial distribution is greater than or equal to a criterion value which is between 0 and 1. The *Figure 2* shows the original table and a simulated table.

The *Figure 3* shows the formula version of the table. For the first column of the table, generate Binomial random variables (for males) with 30 trials and 0.40 success rate, using the formula = BINOM.INV(30, 0.40, rand()). To generate the number of females in that column, use the formula = 30 - BINOM.INV(30, 0.40, rand()). Same way we generate counts for the other two columns.

	A	B	C	D	E	F	G
1							
2							
3			Original Sample Data				
4			Probably		Probably	Column	
5			No	May Be	Yes	Total	p
6		Male	17	32	11	60	0.4
7		Female	13	43	34	90	0.6
8		Row Total	30	75	45	150	
9							
10			Simulated Table assuming independence				
11			Probably		Probably		
12			No	May Be	Yes		
13		Male	10	34	14		
14		Female	20	41	31		
15		Row Total	30	75	45		
16							

Figure 2: Original and Simulated Tables

The *Figure 3* shows the Excel formula version of the simulated table.

	A	B	C	D	E	F	G
1							
2							
3			Original Sample Data				
4			Probably		Probably	Column	
5			No	May Be	Yes	Total	p
6		Male	17	32	11	60	0.4
7		Female	13	43	34	90	0.6
8		Row Total	30	75	45	150	
9							
10			Simulated Table assuming independence				
11			Probably		Probably		
12			No	May Be	Yes		
13		Male	=BINOM.INV(30, 0.4, RAND())	=BINOM.INV(75, 0.4, RAND())	=BINOM.INV(45, 0.4, RAND())		
14		Female	=C15-C13	=D15-D13	=E15-E13		
15		Row Total	30	75	45		
16							

Figure 3: Formula Version of the Table

Now we use Excel Data Table facility to generate **one thousand** random samples and compute the value of the test statistic for each. The *p*-value is the proportion of these test statistic values

that are as extreme or more extreme than what we have observed (8.23). If this proportion is too small (generally less than 0.05), the null hypothesis is unlikely. The *Figure 4* shows a portion of the spreadsheet implementation of this calculation.

			Original Sample Data							Data Table
			Probably		Probably	Column				1.88888889
			No	May Be	Yes	Total	p			2.37037037
		Male	17	32	11	60	0.4			7.39814815
		Female	13	43	34	90	0.6			12.2592593
		Row Total	30	75	45	150				7.12037037
										1.39814815
			Simulated Table assuming independence							4.52777778
			Probably		Probably					6.44444444
			No	May Be	Yes					1.52777778
		Male	10	33	21					1.34259259
		Female	20	42	24					3.09259259
		Row Total	30	75	45					6.42592593
										1.34259259
										0.138888889
			Expected Counts assuming H_0							1.342592593
			Probably		Probably					1.148148148
			No	May Be	Yes					8.064814815
		Male	12	30	18					4.064814815
		Female	18	45	27					2.731481481
										4.731481481
										0.287037037
			Actual Value of the Statistic		8.231481481					4.037037037
			Simulated value of the Statistic		1.888888889					3.527777778
										2.509259259
			p - value		0.016					1.842592593
										3.388888889
										0.87037037
										1.861111111

Figure 4: Portion of the spreadsheet showing Data Table and p-value

The simulated *p*-value is **0.016**, which is less than the assumed significance level of 0.05. This leads to believe that the null hypothesis is not supported and there is a significant relationship between gender and willingness use the mental health services. The Exact *p*-value {P(χ^2 (df = 2) > 8.23) = 0.016} is equal up to three decimal places to our simulated value. Our goal of using this approach is to give students a better understanding of the *p*-value in concepts of hypothesis testing. The empirical distribution of the test statistic is also tabulated. It is shown in *Figure 5*.

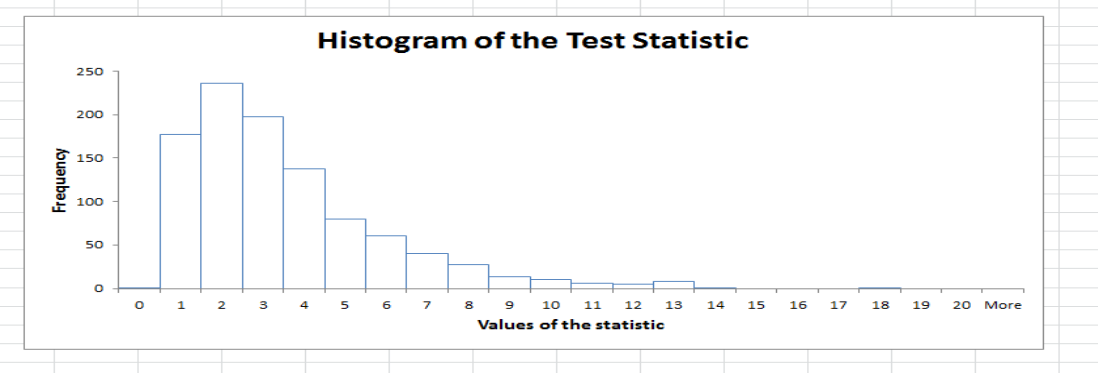

Figure 5: Empirical distribution of the test statistic

The empirical distribution of the simulated values of the test statistic is approximately closer to the Chi-square distribution with 2 degrees of freedom which is the theoretical distribution of the statistic for this contingency table.

5. Future work

a) Even though the *p*-value calculated is very accurate in this case, there is a concern about the accuracy of the BINOM.INV function in Excel. We plan to use R software which is freely available, to generate Binomial random variables in simulating the contingency table and compare the results.

b) We cannot use this approach of using independent Binomial random variables to generate a contingency table when both row variable and column variable have more than two levels. We plan to find a method to simulate such tables using Excel or some other software.

c) We want to compare the two methods of teaching, traditional way of teaching introductory statistics using formulas and calculations to the method of using computer simulation in classroom. Another interesting project would be to compare Excel based simulation to those from other statistical software packages designed to use in teaching.

6. Conclusion

Many students have difficulties of understanding introductory statistics concepts such as hypothesis testing. Statistics instructors are always searching for new and efficient teaching methods to improve statistics instruction in hopes of enhancing student learning. Computer simulation methods as teaching tools are considered to be effective methods. We have demonstrated the use of simulation using Excel and Data Tables in teaching Chi-square test. This is a very useful way to visualize the sampling distribution and to comprehend the *p*-value. Excel is easy to use software and students have access to it. Our experience suggests that this approach is highly acceptable to students with varying backgrounds of mathematics.

References

1. Chandrakantha, L. (2012). "Resampling using Excel in Teaching Statistics", *Electronic Proceedings of ICTCM* , 24, 13 – 20.
2. Christie, D. (2004), "Resampling with Excel," *Teaching Statistics*, 36(1) 9-14.

3. Cobb, P. (1994), "Where is the Mind? Constructivist and Sociocultural Perspectives on Mathematical Development", *Educational Researcher*, 23, 13-20.
4. Gravetter, Frederick and Wallnau, Larry (2011), *Essentials of Statistics for the Behavioral Sciences* (7th edition), WADSWORTH Cengage Learning.
5. Ecklund, P. (2009), "Introduction to Excel 2007 Data tables and Data Table Exercises," Available at http://faculty.fuqua.duke.edu/~pecklund/ExcelReview/Excel%202007%20Data%20Table%20Notes.pdf.
6. Moore, D.S. and McCabe, G.P. (2003), *Introduction to Practice of Statistics* (4th edition), New York, NY. W. H. Freeman & Company.
7. Winston, W. L. (2007), Excel 2007, *Data Analysis and Business Modeling*, Microsoft Press.