RESAMPLING USING EXCEL IN TEACHING STATISTICS

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

Abstract

We will illustrate the use of Excel's standard functions for resampling from a population to demonstrate the Central Limit Theorem in introductory statistics courses. The Data Table facility of Excel will be utilized without writing macros. The concepts of confidence intervals and hypothesis testing are explained using resampling.

1. Introduction

Fundamental statistical concepts such as sampling distributions, central limit theorem, confidence intervals, hypothesis testing, and *p*-values are very important in an introductory statistics course. Many students struggle to understand these concepts. 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 the 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 and standard formulas in teaching these concepts in an introductory statistics class. We prefer Excel due to the availability to students, user friendliness, and ease of presenting the situation in multiple rows and columns.

The simulation and resampling have been used in teaching statistics for many years. Christie (2004) has used the Excel Data Tables for estimating the population mean and the 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).

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. 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:

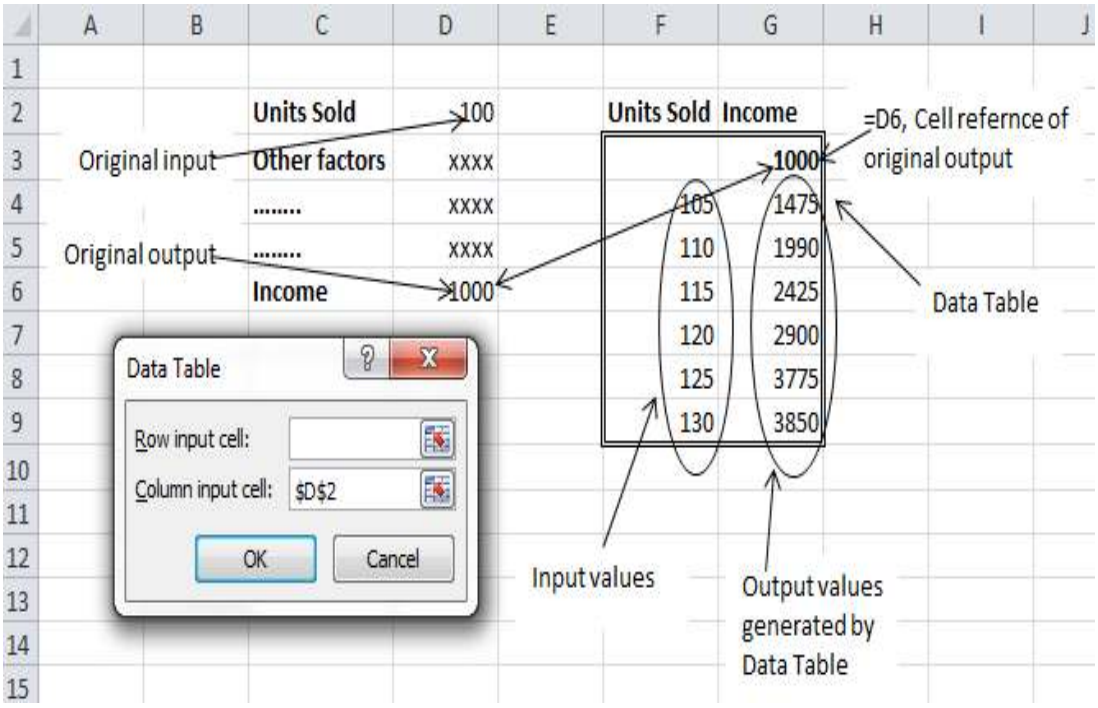

Figure 1: Data Table Setup

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. Sampling Distribution Using Excel Data Table

Now we use Excel Data Table to generate the sampling distribution of mean and introduce the concepts. Let $\bar{x}$ be the mean of a simple random sample (SRS) of size n from a population with mean μ and standard deviation σ. The mean and standard deviation of $\bar{x}$ are given by:

$$\mu_{\bar{x}} = \mu, \quad \sigma_{\bar{x}} = \sigma/\sqrt{n}$$

If the original population has N(μ,σ) distribution, then the sample mean $\bar{x}$ has the $N(\mu, \sigma/\sqrt{n})$ distribution. The Central Limit Theorem describes the shape of the distribution. If n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately normal. The larger the sample size, better the approximation. In other words, $\bar{x}$ is approximately $N(\mu, \sigma/\sqrt{n})$.

We generate random samples of sizes 10, 30, and 50 from following distributions:

- Uniform distribution: U(0, 100) (μ = 50, σ = 28.87), Excel formula: =rand()*100.
- Normal distribution: N(100, 15) (μ = 50, σ = 15), Excel formula: = (normsinv(rand())*15) + 100.

The *Figure 2* shows a portion of the spreadsheet that uses Data Table function to generate different sample means.

	A	B	C	D	E	F	G
1		**U(0, 100) Sample n= 10**			**Means of samples**		
2		86.60747178			68.25659		
3		76.65841508			47.64591	**Mean of X bar**	50.54535
4		84.33490381			43.43698	**Std Dev of X bar**	9.188041
5		43.56490315			61.26585		
6		86.00399853			56.26724		
7		98.31434904			48.11777		
8		35.73490052			62.8572		
9		97.36614357			61.93387	μ = 50	
10		72.40110249			48.52743	σ = 28.87	
11		1.579713836			46.69051	σ/√n = 9.12	
12					54.46102		
13	**X bar**	68.25659018			50.77352		
14					53.67147		
15					54.28063		
16		**Data Table**			79.71707		
17					60.66387		
18					39.9767		

Figure 2: Portion of the spreadsheet showing simulated sampling distribution

Notice that the mean and the standard deviation of $\overline{x}$ are approximately close to theoretical values, μ and $\sigma/\sqrt{n}$. The *Figure 3* shows histograms of the sample mean $\overline{x}$ for different values of n.

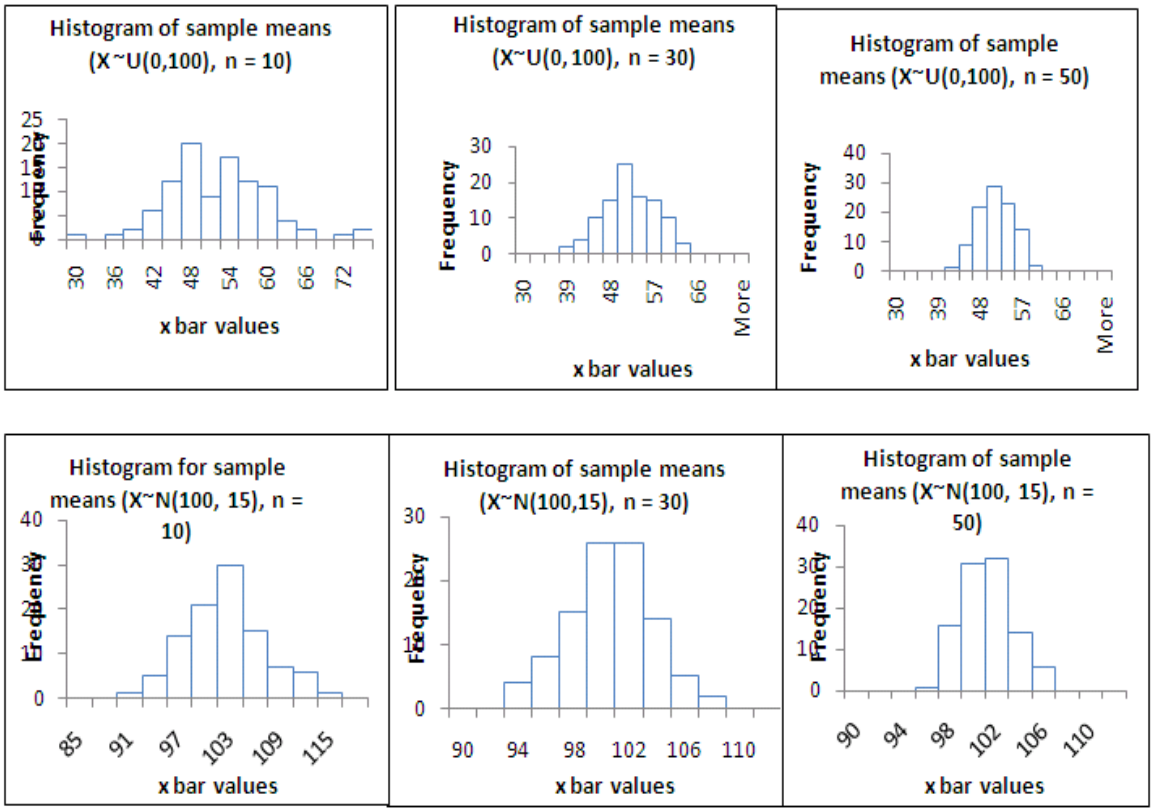

Figure 3: Sampling distribution of mean for different values of n

The histograms in *Figure 3* show the following properties of sampling distribution of the mean:

- The mean of the sampling distribution is equal to the population mean μ.
- The standard deviation (standard error) is equal to $\sigma/\sqrt{n}$ and as sample size n increases, the standard deviation gets smaller.
- The shape of the distribution gets closer to the normal distribution as the sample size n increases.

4. Confidence Intervals for Mean

The elementary statistics students have difficulties in understanding the meaning of confidence intervals. We can use simulation and Data Tables in Excel to generate confidence intervals and use them to understand the concepts. The level C confidence interval for the mean μ when the population standard deviation σ is known is given by $\overline{X} \pm Z^* \sigma/\sqrt{n}$ where Z* is the value of the standard normal curve with area C between critical points –Z* and Z* and n is the

sample size. This interval is exact when the population is normally distributed and is approximately correct when n is larger in other cases.

The confidence level C is the probability that the confidence interval actually does contain the population mean μ, assuming the estimation process is repeated a large number of times.

We can illustrate this concept to students using Excel Data Tables. We generate a large number of samples and construct the confidence interval for each sample, and compute the proportion of samples that do actually contain the mean. This proportion should approximately be closer to **C**. The *Figure 4* shows the portion of the spreadsheet that calculates the 95% confidence intervals for different samples and the proportion of them that contain the mean.

	A	B	C	D	E	F	G	H	I	J	K
1		N(100, 15) Sample of n = 30			Sample Means		95% CI Lower Bound	95% CI Upper Bound		CI contains the mean? (1-yes, 0-no)	Percent of samples captured the true mean
2		104.2839929			98.7834742		93.41579	104.1512		1	95
3		90.77831303			93.1199636		87.75228	98.48764		0	
4		106.942352			99.4106537		94.04297	104.7783		1	
5		113.5260259			104.382559		99.01488	109.7502		1	
6		75.84483671			104.284995		98.91731	109.6527		1	
7		121.8190126			101.705694		96.33801	107.0734		1	
8		101.6942447			101.429307		96.06163	106.797		1	
9		113.6176255			104.268969		98.90129	109.6366		1	
10	Data Table	115.2264053			98.7726596		93.40498	104.1403		1	
11		107.8495499			102.328763		96.96108	107.6964		1	
12		67.83280399			100.803594		95.43591	106.1713		1	
13		78.44055988			102.105308		96.73763	107.473		1	
14		98.86227308			98.7480305		93.38035	104.1157		1	
15		114.8702656			100.019148		94.65147	105.3868		1	
16		102.5517828			105.708207		100.3405	111.0759		0	
17		122.5290135			94.5695884		89.20191	99.93727		0	
18		122.1028138			103.052489		97.68481	108.4202		1	

Figure 4: Portion of the spreadsheet showing confidence intervals

Notice that 95% of the confidence intervals do contain the actual mean. If we generate another set of confidence intervals by pressing the **F9** key, we may not get 95%, but it will be closer to 95%.

5. Hypothesis Testing

Hypothesis testing is also one of the most difficult concepts for introductory statistics students. Many instructors use the *p*-value approach to make the decision of the test. Almost all software packages give the *p*-value of the test and the students need a good understanding of this concept to make the appropriate conclusion. 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).

We will show how to generate different random samples, calculate the value of the statistic, and the corresponding *p*-value. This simulation approach gives students a better understanding about the *p*-value. We use standard Excel formulas and Data Table function to calculate the value of the test statistic for different random samples from original population.

We consider the following problem from elementary statistics text "Essential Statistics" by David Moore (1996).

Problem 13.15: **Reading a computer Screen.** Does the use of fancy type fonts slow down the reading of text on a computer screen? Adults can read four paragraphs of text in an average time of 22 seconds in the common Times New Roman font. 25 adults were asked to read this text in the ornate font named Gigi. Here are their times:

23.2, 21.2, 28.9, 27.7, 23.4, 27.3, 16.1, 22.6, 25.6, 32.6, 23.9, 26.8, 18.9, 27.8, 21.4, 30.7, 21.5, 30.6, 31.5, 24.6, 23.0, 28.6, 24.4, 28.1, 18.4.

Suppose that reading times are normally distributed with σ = 6 seconds. Is there good evidence that the mean reading time for Gigi fonts is greater than 22 seconds? In other words, is μ greater than 22 seconds for Gigi fonts? We use z test procedure.

Null and alternative hypotheses:

H_0: μ = 22 seconds vs. H_1: μ> 22 seconds

Test statistic: $$Z = \frac{(\bar{x} - \mu_0)}{\sigma/\sqrt{n}}$$

where μ_0 is the value of the mean assuming null hypothesis.
Value of the test statistic based on the observed sample, assuming null hypothesis is given by z = (25.152-22)/(6/√25) = 2.627.

Is this value of the statistic, z = 2.627 too extreme? Do we have evidence against the null hypothesis? We calculate the *p*-value using simulation and Excel Data Tables. Five hundred samples (assuming H_0) from normal distribution with μ = 22 and σ = 6 are generated and the value of the test statistics is calculated for each sample. The *p*-value is the proportion of these samples given the test statistic value is extreme or more extreme than what we have observed (2.627). If this proportion is too small, the null hypothesis is unlikely.

Figure 5 shows the portion of the spreadsheet that generates different random samples, computes the values of the test statistic, and the corresponding p-value.

	A	B	C	D	E	F	G	H	I	J
1	**N(22,6) Sample of n= 25**					**Value of Test Statistic**				
2	22.42849946					1.101035		**p-value**	**0.004**	
3	20.94480037					0.397195				
4	26.77824616	**Mean**	23.3212419			0.197975				
5	33.99876437	**Value of the test Statistic**	1.10103489			1.407352				
6	25.72684404	**Actual value of statistic**	2.627			-0.40211				
7	26.38735477					0.011383				
8	33.13279105					0.232137				
9	17.67543878					0.773528				
10	11.97757585					1.88721				
11	25.47390478	**Data Table**				-1.44687				
12	7.212108948					0.392589				
13	23.3656068					-0.61523				
14	18.66161312					0.885123				
15	30.58873095					1.48786				
16	28.40138387					1.373605				
17	26.29319086					0.275735				
18	21.355453					-0.15012				
19	18.04176833					0.48726				
20	29.3601671					1.039901				

Figure 5: Portion of the spreadsheet showing values of statistic and p-value

Interpreting this small *p*-value, we can say that it is unlikely that we observe the value of the statistic as large as, or larger than the one we observed from actual sample data when the null hypothesis is true. We conclude in favor of the claim that Gigi fonts take more time to read.

The Exact *p*-value {P(z > 2.627) = 0.004} is equal to our simulated value.

Our goal of using this approach is to give students a better understanding of the *p*-value and concepts of hypothesis testing.

6. Conclusion

Many students have difficulties of understanding introductory statistics concepts. We have demonstrated the use of Excel simulation and Data Tables in teaching these concepts. This is a very useful way to visualize the sampling distributions and to get a better understanding of the *p*-values. Using Excel is quite easy and students have access to it. Our experience suggests that it is highly acceptable for students with varying backgrounds of mathematics.

References

1) Christie, D. (2004), "Resampling with Excel," *Teaching Statistics*, 36(1) 9-14.
2) 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.
3) Moore, David S. (1996), *Essential Statistics*, New York, NY: W. H. Freeman & Company.
4) Moore, D.S. and McCabe, G.P. (2003), *Introduction to Practice of Statistics* (4th ed.), New York, NY: W. H. Freeman & Company.
5) Winston, W. L. (2007), *Excel 2007, Data Analysis and Business Modeling*, Microsoft Press.