Illustrating the Central Limit Theorem Through Microsoft Excel Simulations

David H. Moen, (Email: dmoen@usd.edu) University of South Dakota
John E. Powell, (Email: jpowell@usd.edu) University of South Dakota

Abstract

Using Microsoft Excel, several interactive, computerized learning modules are developed to demonstrate the Central Limit Theorem. These modules are used in the classroom to enhance the comprehension of this theorem. The Central Limit Theorem is a very important theorem in statistics, and yet because it is not intuitively obvious, statistics students often have difficulty accepting it. Nevertheless, understanding this theorem is essential because of its importance in statistical inference.

Introduction

There are several statistical topics that students typically have difficulty understanding. Included in this list are concepts associated with measuring variation in data, sampling distributions, hypothesis testing, and regression analysis. Microsoft Excel (4) includes quite a few statistical analysis tools, including tools to analyze some of the topics just listed. Also, some statistics textbooks have Excel add-ins that provide additional analysis capabilities. However, since these tools simply present the output associated with a particular procedure, students must be able to correctly interpret the results. In addition, these tools do not provide an understanding of the concepts that underlie a procedure. Without this understanding, it is oftentimes much more difficult to know when the use of a particular technique is appropriate.

Statistics is a very valuable tool, and with today's technological capabilities, even more can be done to improve students' understanding of its importance in their future business careers. The specific intent of this paper is to discuss the development of several interactive, computerized learning modules that illustrate the validity of using the Central Limit Theorem in a variety of statistical inference procedures. Microsoft Excel is used to create these modules, since Excel is readily available and because many required undergraduate business statistics courses use Excel as the software package for statistical analysis.

Methodology

Many procedures in statistical inference are based on the use of the normal probability distribution (a symmetrical bell-shaped distribution). The normal probability distribution is frequently appropriate because of the Central Limit Theorem. This theorem states that when a random sample of n observations is selected from a population (any population) with a mean of μ and a standard deviation of σ, then when n is large, the sampling distribution of the mean is approximately a normal distribution with a mean of μ and a standard deviation of $\sigma/\sqrt{n}$ (standard error of the mean) (3, p. 332). This theorem can also be rewritten to apply to the sum of sample measurements. Thus, the Central Limit Theorem states

that under rather general conditions, sums and means of random measurements drawn from any population tend to possess, approximately, a bell-shaped distribution in repeated sampling.

Since many of the estimators that are used to make inferences about the characteristics of a population are sums or means of sample measurements, we can expect the estimator to be approximately normally distributed in repeated sampling, when n is sufficiently large. This is not an intuitive result, and despite textbook illustrations and in-class discussion, the rationale for using the normal probability distribution often remains unclear. And, of course, there is always the question, "What do you mean by n being sufficiently large?"

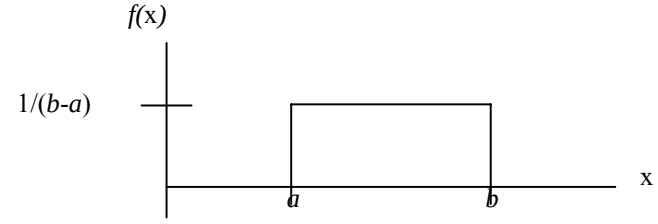
Consider the sampling distribution of the mean. In the following discussion, several interactive Microsoft Excel modules are created that illustrate the Central Limit Theorem. Sampling is done from three different populations, using different sample sizes, and the results also include calculations for the mean and standard deviation of the estimated sampling distribution. Specifically, Excel simulations are created using three different population distribution families: uniform, exponential and V-shaped. In each case, the parameters associated with a population distribution can be modified to allow for the simulation of a wide variety of populations within each family.

Uniform Probability Distribution Results

Consider the continuous uniform probability distribution with parameters *a* and *b,* where $a < b$. The probability density function for a random variable *x* is given by

$$f(x) = 1/(b-a), \text{ for } a \le x \le b$$
$$= 0 \text{ elsewhere}$$

, where $E(x) = \mu = (a + b)/2$ and $Var(x) = \sigma^2 = (b - a)^2/12$. (1, p. 225)



The Excel module created with the use of this population distribution allows the user to select values for parameters *a* and *b*. For illustration purposes, consider a continuous uniform probability distribution with parameters $a = 20$ and $b = 100$. Then, $E(x) = \mu = (20 + 100)/2 = 60$ and $Var(x) = \sigma^2 = (100 - 20)^2/12 = 533.33$. It follows that the standard deviation $\sigma = \sqrt{Var(x)} = 23.094$.

Microsoft Excel includes the RAND() function (2, p. 248) that returns a uniformly distributed random number greater than or equal to 0 and less than 1. Thus, a uniformly distributed random number in the interval [*a*, *b*) can be generated in Excel using the formula *a* + (*b*-*a*)*RAND(). It should be noted that Microsoft Excel does have a random number generation feature that can be accessed by selecting 'Tools', 'Data Analysis' and 'Random Number Generation', and the ability to generate uniformly distributed random numbers in the interval [a,b) is an available option. However, this approach does not provide the dynamic simulation capability that is described below when using the RAND() function.

Two scenarios were developed for this first situation, and in each case, the selection of 500 random samples was simulated. In the first instance, 500 random samples each of size n = 5 were selected. Once these 500 random samples had been generated, the simulated sampling distribution was created by computing the sample mean for each sample and then grouping these means to form a frequency distribution and histogram. Each time function key F9 (Calculate) is depressed, 500 new samples are simulated, the sample means are recalculated, and the accompanying frequency distribution, histogram, and descriptive statistics are recomputed. Figure 1 provides the Excel spreadsheet labeling and cell formulas used to create the first scenario, while Figure 2 displays the results from an example of one simulation.

The dynamic frequency distribution and histogram capability is accomplished through the use of the FREQUENCY function in Excel. The general format for this function is FREQUENCY(data_array, bins_array). The data range whose frequencies are to be counted is defined in the data_array field, while the bins_array is the array containing the upper class limits for the distribution. Using the same example to illustrate this process, it is observed in Figure 1 that the 500 sample means are located in cells G7 to G506. Suppose 26 upper class limits, ranging in our example from 20 up to 100, are stored in cells J6 to J31. These limits are defined so that the classes are of equal width. Next highlight the range where the frequencies are to be displayed, say K6 to K31. While this range is highlighted, enter the formula =FREQUENCY(G7:G506,J6:J31). Finally, while simultaneously holding down the CTRL and SHIFT keys, press the ENTER key. The result is that the formula entered into cell K6 will also be copied into all of the cells in this range, and further, when the sample means change (every time function key F9 to recalculate is depressed), the frequency counts will automatically be updated, which in turn will update the histogram. (The histogram can be created using the "Chart Wizard" feature in Excel.) Figure 3 displays the formulas used to create this frequency distribution. Note that Excel automatically places braces { } around the FREQUENCY formula once it has been entered into cell K6. (The user enters the formula without braces.)

Figure 1

	A	B	C	D	E	F	G
1	Population	Distr:Uniform(a,b)					
2	a =	20.000	(b – a) =	=B3–B2			
3	b =	100.000					
4							
5	Sample		Sample	Values			Sample
6	Number	1	2	3	4	5	Mean
7	1	=$B$2+$D$2*RAND()	…	…	…	=$B$2+$D$2*RAND()	=AVERAGE(B7:F7)
8	2	=$B$2+$D$2*RAND()	…	…	…	=$B$2+$D$2*RAND()	=AVERAGE(B8:F8)
…	…	…	…	…	…	…	…
506	500	=$B$2+$D$2*RAND()	…	…	…	=$B$2+$D$2*RAND()	=AVERAGE(B506:F506)

Figure 2

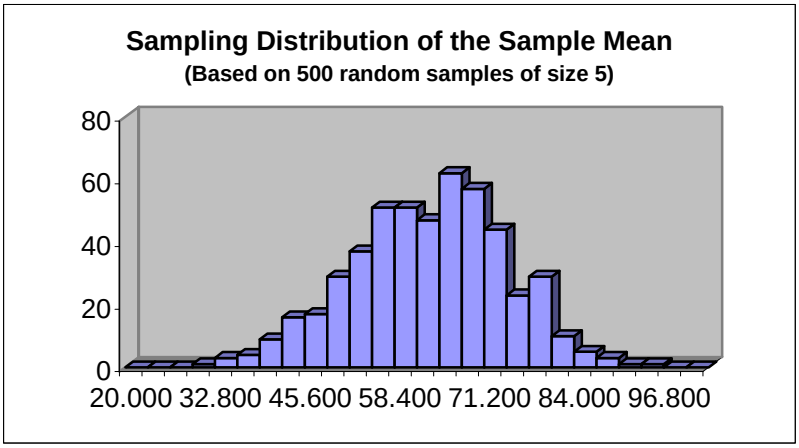
	A	B	C	D	E	F	G
1	Population	Distribution:Uniform(a,b)					
2	a =	20.000	(b – a) =	80.000			
3	b =	100.000					
4							
5	Sample		Sample	Values			Sample
6	Number	1	2	3	4	5	Mean
7	1	39.817	20.787	86.256	96.026	51.164	58.810
8	2	23.447	50.939	44.430	45.496	34.882	39.839
…	…	…	…	…	…	…	…
506	500	87.982	83.969	39.023	60.167	32.915	60.811

Figure 3

	J	K
5	**Bins**	**Frequency**
6	=$B$2	{=FREQUENCY(G7:G506,J6:J31)}
7	=J6+$D$2/25	{=FREQUENCY(G7:G506,J6:J31)}
…	…	…
31	=J30+$D$2/25	{=FREQUENCY(G7:G506,J6:J31)}

Figure 4 provides the histogram and descriptive statistics for this simulation example. Note that when sampling has been conducted from a continuous uniform probability distribution for a sample as small as n = 5, the simulated sampling distribution's shape is approximately normal and the mean and standard deviation are close to μ and σ/√n respectively.

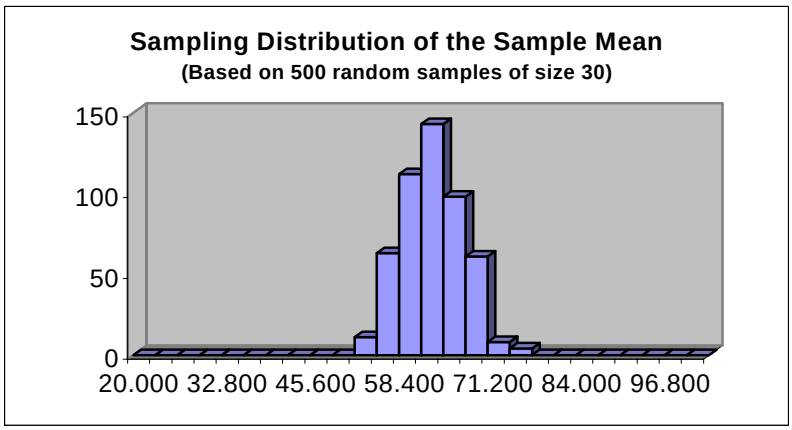
Figure 4



Numerical Measures	
Population Mean =	60.0000
Population Std. Dev. =	23.0940
Population Std. Dev./SQRT(n) =	10.3280
Simulated Sampling Distribution Mean =	59.8847
Simulated Sampling Distribution Std. Dev. =	10.8308

In the second case, 500 random samples each of size n = 30 were selected. Except for the fact that there are now 30 sample values for each sample, the same procedure was used to create this scenario as was described in the n = 5 case. Figure 5 provides one simulation example. Note that when the sample size increases to n = 30, the simulated sampling distribution's shape closely approximates a normal probability distribution, and once again, the mean and standard deviation are close to μ and σ/√n respectively. Of course, a larger sample size in this second scenario means that the standard deviation of the sampling distribution will be smaller, since now the standard error of the mean is σ/√30 instead of σ/√5.

Figure 5



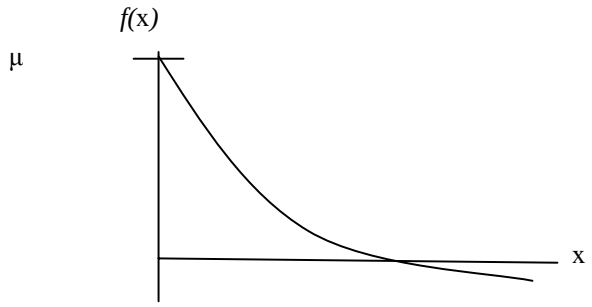
Numerical Measures	
Population Mean =	60.0000
Population Std. Dev. =	23.0940
Population Std. Dev./SQRT(n) =	4.2164
Simulated Sampling Distribution Mean =	59.8762
Simulated Sampling Distribution Std. Dev. =	4.3408

Exponential Probability Distribution Results

The exponential probability distribution is often used to describe the time between arrivals (IAT) at a service facility or the service time required at a facility.

Consider the continuous exponential probability distribution with parameter μ, where μ represents time. The probability density function for a random variable *x* is given by

$f(x) = \mu e^{-\mu x}$, *for* $x \geq 0, \mu > 0$
$= 0$ *elsewhere*
, where E(x) = 1/μ and Var(x) = $\sigma^2 = 1/\mu^2$. (5, p. 81)



The Excel module created with the use of this population distribution allows the user to select values for the parameter μ. For illustration purposes, consider a continuous exponential probability distribution with parameter μ = 0.5. Then, E(x) = 1/μ = 2.0 and Var(x) = $\sigma^2 = 1/\mu^2 = 4.0$. It follows that the standard deviation σ = √Var(x) = 2.0.

The exponential probability distribution is a special case of the gamma probability distribution, and Microsoft Excel includes the GAMMAINV function that returns the inverse of the gamma cumulative distribution. The general format for the GAMMAINV function is GAMMAINV(probability, alpha, beta). The probability associated with the gamma distribution is specified in the probability field. The alpha field is a parameter for the gamma distribution, and when set to the value 1, specializes to the exponential distribution. When alpha = 1, beta is the mean of the exponential distribution or 1/μ. This function can be used to generate values from an exponential distribution; however, it should be noted that the GAMMAINV function uses an iterative technique to converge to a value, so the recalculation process is much longer than another approach that will be described next.

For an expected value of 1/μ, exponential random variates can be generated with the formula –1/μ*LN(RAND()) where LN is the natural logarithm. (5, p. 82).

As in the case of the uniform probability distribution, two scenarios were developed for the exponential probability distribution. In the first instance, 500 random samples each of size n = 5 were selected. Figure 6 provides the Excel spreadsheet labeling and cell formulas used to create the first scenario, while Figure 7 displays the results from an example of one simulation.

Figure 6

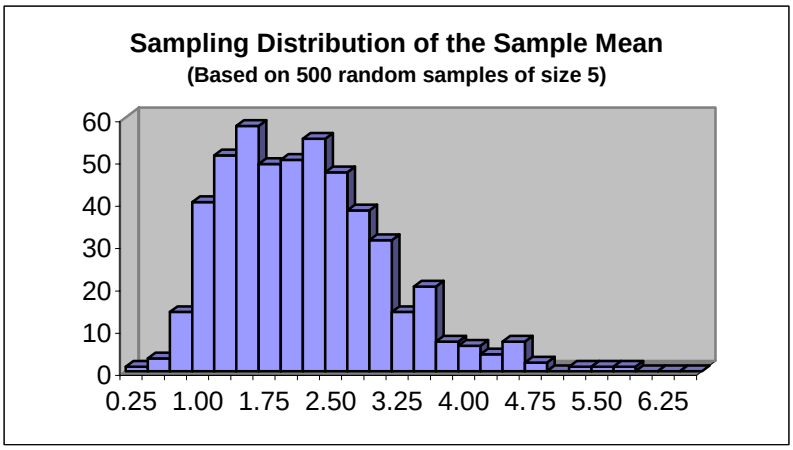
	A	B	…	F	G
1	Population	Distribution:Exponential			
2		1/μ = 2.000			
3					
4	Sample				Sample
5	Number	1	…	5	Mean
6	1	=-$B$2*LN(RAND())	…	=-$B$2*LN(RAND())	=AVERAGE(B6:F6)
7	2	=-$B$2*LN(RAND())	…	=-$B$2*LN(RAND())	=AVERAGE(B7:F7)
8	3	=-$B$2*LN(RAND())	…	=-$B$2*LN(RAND())	=AVERAGE(B8:F8)
…	…	…	…	…	…
505	500	=-$B$2*LN(RAND())	…	=-$B$2*LN(RAND())	=AVERAGE(B505:F505)

Figure 7

	A	B	C	D	E	F	G
1	Population	Distribution:Exponential	(μ),	where	μ >0		
2		1/μ = 2.000					
3							
4	Sample		Sample	Values			Sample
5	Number	1	2	3	4	5	Mean
6	1	0.647	2.513	5.061	0.088	2.880	2.238
7	2	1.979	1.918	0.697	0.391	0.022	1.002
8	3	0.977	0.438	0.644	6.650	1.304	2.003
…	…	…	…	…	…	…	…
505	500	4.758	4.596	0.143	1.690	2.010	2.639

Figure 8 provides the histogram and descriptive statistics for this simulation example. Note that when sampling has been conducted from a continuous exponential probability distribution for a sample of n = 5, the simulated sampling distribution's shape is somewhat normal and the mean and standard deviation are close to μ and σ/√n respectively.

Figure 8

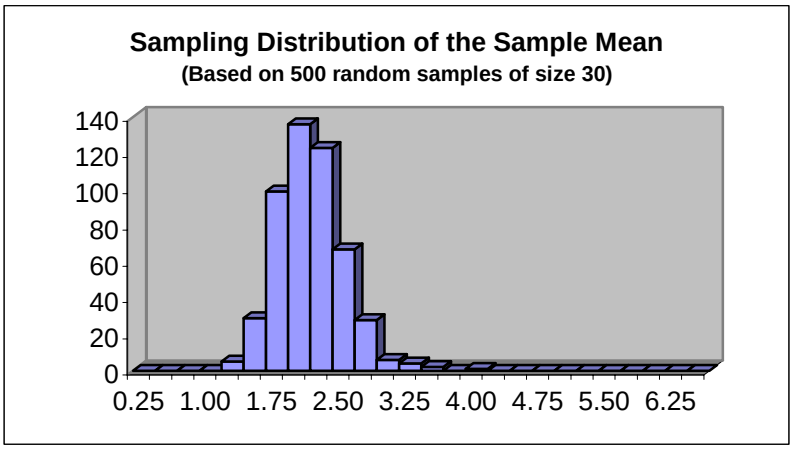


Numerical Measures

Population Mean =	2.0000
Population Std. Dev. =	2.0000
Population Std. Dev./SQRT(n) =	0.8944
Simulated Sampling Distribution Mean =	2.0145
Simulated Sampling Distribution Std. Dev. =	0.8973

In the second case, 500 random samples each of size n = 30 were selected. Figure 9 provides one simulation example. Note that when the sample size increases to n = 30, the simulated sampling distribution's shape more closely approximates a normal probability distribution, and once again, the mean and standard deviation are close to μ and σ/√n respectively. As with the uniform probability distribution, a larger sample size in this second scenario means that the standard deviation of the sampling distribution will be smaller, since now the standard error of the mean is σ/√30 instead of σ/√5. This smaller standard deviation is clearly observed when comparing the histograms in Figures 8 and 9.

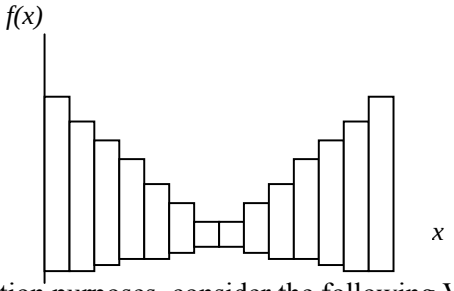
Figure 9



Numerical Measures	
Population Mean =	2.0000
Population Std. Dev. =	2.0000
Population Std. Dev./SQRT(n) =	0.3651
Simulated Sampling Distribution Mean =	1.9864
Simulated Sampling Distribution Std. Dev. =	0.3638

V-Shaped Discrete Probability Distribution Results

Consider a discrete probability distribution for a random variable x. Then $f(x) \geq 0$ and $\Sigma f(x) = 1$. The expected value for a discrete random variable is $E(x) = \mu = \Sigma x f(x)$, while the variance for a discrete random variable is $Var(x) = \sigma^2 = \Sigma(x-\mu)^2 f(x)$. (1, p. 194) A V-shaped discrete probability distribution would generally appear as observed below. Even with this very non-normal population distribution, the sampling distribution of the sample mean will still be approximately normal for larger sample sizes.



For illustration purposes, consider the following V-shaped discrete probability distribution in Figure 10, which is generated from Table 1.

Figure 10

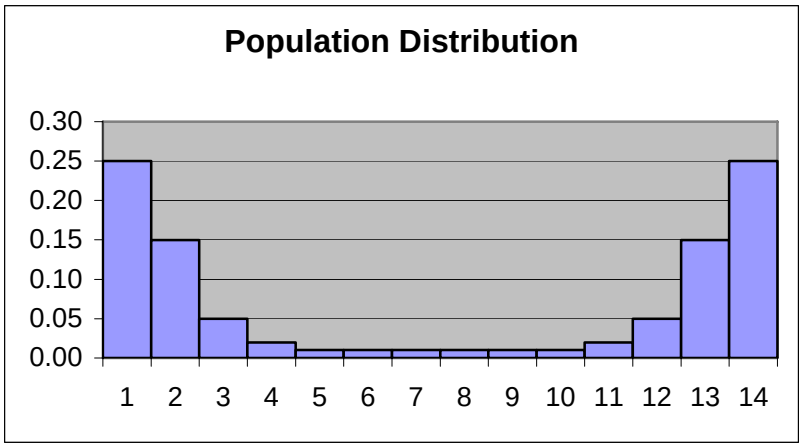


Table 1

X	F(X)
1	0.25
2	0.15
3	0.05
4	0.02
5	0.01
6	0.01
7	0.01
8	0.01
9	0.01
10	0.01
11	0.02
12	0.05
13	0.15
14	0.25

Sampling from a discrete probability distribution of this size cannot be accomplished by using nested IF statements in Excel, since the maximum allowable number of nested IF statements is seven. (4, IF function Microsoft Excel Help) However, Microsoft Excel has another function, VLOOKUP, which allows any number of groupings or classes. The VLOOKUP function has three arguments that will be used in this application. The general format for this function is VLOOKUP(lookup_value, table_array, column_index_number). The first argument is the number to be looked up. The second argument is the location of the table of information where the data is looked up, i.e., the vertical lookup table. The third argument indicates which column in the vertical lookup table contains the answer, that is, the value to be returned by the function. (2, p. 193)

As in the case of the previous probability distributions, two scenarios were developed for this discrete distribution. In the first instance, 500 random samples each of size n = 5 were selected. Figure 11 provides the Excel spreadsheet labeling and cell formulas used to create the first scenario, while Figure 12 displays the results from an example of one simulation. Note that Figure 11 assumes that the cumulative probability distribution shown in Table 2 is located in cells C2 through D15.

Table 2

	C	D
1	Cumulative Frequency	X
2	0.00	1
3	0.25	2
4	0.40	3
5	0.45	4
6	0.47	5
7	0.48	6
8	0.49	7
9	0.50	8
10	0.51	9
11	0.52	10
12	0.53	11
13	0.55	12
14	0.60	13
15	0.75	14

Figure 11

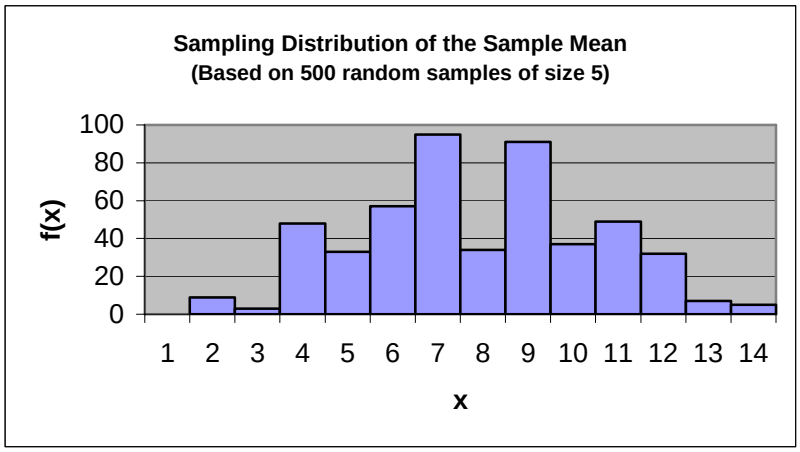
	A	B	…	F	G
19	Sample				Sample
20	Number	1	…	5	Mean
21	1	=VLOOKUP(RAND(),$C$2:$D$15,2)	…	=VLOOKUP(RAND(),$C$2:$D$15,2)	=AVERAGE(B21:F21)
22	2	=VLOOKUP(RAND(),$C$2:$D$15,2)	…	=VLOOKUP(RAND(),$C$2:$D$15,2)	=AVERAGE(B22:F22)
23	3	=VLOOKUP(RAND(),$C$2:$D$15,2)	…	=VLOOKUP(RAND(),$C$2:$D$15,2)	=AVERAGE(B23:F23)
…	…	…	…	…	…
520	500	=VLOOKUP(RAND(),$C$2:$D$15,2)	…	=VLOOKUP(RAND(),$C$2:$D$15,2)	=AVERAGE(B520:F520)

Figure 12

	A	B	C	D	E	F	G
19	Sample		Sample	Values			Sample
20	Number	1	2	3	4	5	Mean
21	1	4	3	7	14	13	8.20
22	2	13	14	4	1	3	7.00
23	3	14	14	1	13	13	11.00
…	…	…	…	…	…	…	…
520	500	3	1	2	2	14	4.40

Figure 13 provides the histogram and descriptive statistics for this simulation example. Note that when sampling has been conducted from this very non-normal discrete V-shaped probability distribution for a sample of n = 5, the simulated sampling distribution's shape is still somewhat mound-shaped and the mean and standard deviation are close to μ and $\sigma/\sqrt{n}$ respectively.

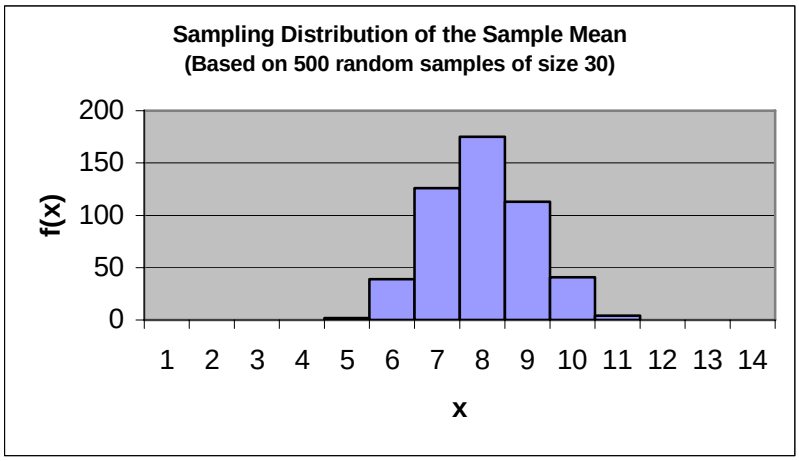
Figure 13



Numerical Measures	
Population Mean =	7.500
Population Std. Dev. =	5.735
Population Std. Dev./SQRT(n) =	2.565
Simulated Sampling Distribution Mean =	7.490
Simulated Sampling Distribution Std. Dev. =	2.437

In the second case, 500 random samples each of size n = 30 were selected. Figure 14 provides one simulation example. Note again that when the sample size increases to n = 30, the simulated sampling distribution's shape quite closely approximates a normal probability distribution, and once again, the mean and standard deviation are close to μ and $\sigma/\sqrt{n}$ respectively. As before, a larger sample size in this second scenario means that the standard deviation of the sampling distribution will be smaller, since now the standard error of the mean is $\sigma/\sqrt{30}$ instead of $\sigma/\sqrt{5}$.

Figure 14



Numerical Measures	
Population Mean =	7.500
Population Std. Dev. =	5.735
Population Std. Dev./SQRT(n) =	1.047
Simulated Sampling Distribution Mean =	7.452
Simulated Sampling Distribution Std. Dev. =	1.067

Conclusion

The objective of this paper has been to develop a better understanding of the Central Limit Theorem through the use of several widely different population distributions. Microsoft Excel provides the opportunity to create simulations that demonstrate this non-intuitive theorem. As the sample size increases from n = 5 to n = 30, it can be clearly observed that the simulated sampling distribution of the sample mean more closely represents a normal probability distribution. The simulations also illustrate that the mean and standard deviation for the sampling distribution are μ and $\sigma/\sqrt{n}$ respectively. The end result of demonstrating these simulations in a statistics class is that students will have a clearer understanding and a better appreciation of the usefulness of the Central Limit Theorem.

References

Anderson, David R.; Sweeney, Dennis J.; Williams, Thomas A.; (2005) *Statistics for Business and Economics (9th edition)*, South-Western Publishing.

Gips, James; (2003) *Mastering Excel: A Problem-Solving Approach (2nd edition)*, John Wiley & Sons, Inc..

McClave, James T.; Benson, George P.; Sincich, Terry; (2005) *Statistics for Business and Economics (9th edition)*, Pearson Prentice Hall.

Microsoft® Excel 2002, Copyright© Microsoft Corporation 1985-2001.

Naylor, Thomas; Balintfy, Joseph; Burdick, Donald; Chu, Kong; (1968) *Computer Simulation Techniques*, John Wiley & Sons, Inc.