

A Microsoft[®] Excel Simulation Illustrating The Central Limit Theorem's Appropriateness For Comparing The Difference Between The Means Of Any Two Populations

David H. Moen, University of South Dakota, USA
John E. Powell, University of South Dakota, USA

ABSTRACT

Using Microsoft[®] Excel, several interactive, computerized learning modules are developed to illustrate the Central Limit Theorem's appropriateness for comparing the difference between the means of any two populations. These modules are used in the classroom to enhance the comprehension of this theorem as well as the concepts that provide the foundation for inferences involving the comparison of two population means.

Keywords: Central Limit Theorem; Interactive Excel Simulations; Difference between Two Means of Disparate Populations

INTRODUCTION

There are many instances where the comparison of two population means is desirable. One approach to these types of inferences is to select independent, random samples from each population and compute the sample mean for each sample. The difference between the two sample means ($\bar{x}_1 - \bar{x}_2$) is then used as a point estimator for the difference between the two population means ($\mu_1 - \mu_2$). Different samples result in various values for the two sample means, and it is the sampling distribution of ($\bar{x}_1 - \bar{x}_2$) that describes the characteristics of this point estimator. If both sample sizes are sufficiently large, the Central Limit Theorem leads to the conclusion that the sampling distribution of ($\bar{x}_1 - \bar{x}_2$) can be approximated by a normal probability distribution (a symmetrical bell-shaped distribution). Additional characteristics of the sampling distribution are that the mean is ($\mu_1 - \mu_2$), and the standard deviation is $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ (Anderson, 2008). These results are not intuitively obvious, and despite textbook illustrations and in-class discussions, the rationale for using the normal probability distribution often remains unclear. However, through the use of Microsoft[®] Excel simulations, it is possible for students to gain a clearer understanding and appreciation of both the Central Limit Theorem and the concepts that provide the foundation for inferences involving the comparison of two population means.

METHODOLOGY

The Central Limit 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 σ/\sqrt{n} (standard error of the mean) (McClave, 2005). This theorem can also be generalized, and in doing so it states that

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

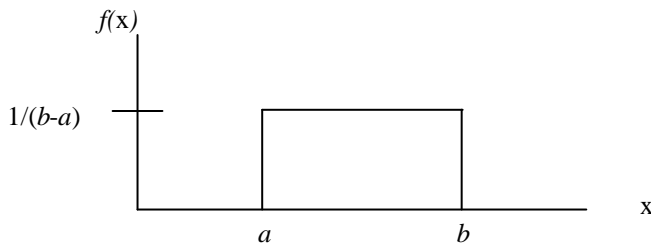
Consider the sampling distribution for the difference between two sample means. In the following discussion, several interactive Microsoft® Excel modules are created that illustrate the Central Limit Theorem and inferences for the difference between two population means. Sampling is done from two different populations. Specifically, Excel simulations are created using two different population distribution families: uniform and exponential. 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. The simulation techniques used below follow the procedures found in Moen and Powell, 2005. The actual Excel formulas are also found in that paper. These techniques provide the ability to simulate the selection of repeated random samples from uniform and exponential population distributions. The simulated sampling distribution can then be represented with a frequency distribution and histogram. The results also include calculations for the mean and standard deviation of the estimated sampling distribution. The creation of the frequency distribution, histogram, and descriptive statistics are actually dynamic. That is, each time function key F9 (Calculate) is depressed in Excel, new samples are simulated, the differences between the sample means are recalculated, and the accompanying frequency distribution, histogram, and descriptive statistics are recomputed. All of the illustrations below all based on the selection of 500 random samples each of size $n_1 = n_2 = 30$.

RESULTS WHEN BOTH POPULATIONS ARE UNIFORMLY DISTRIBUTED

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

$$f(x) = 1/(b - a), \text{ for } a \leq x \leq b, \text{ where } E(x) = \mu = (a + b)/2 \text{ and } \text{Var}(x) = \sigma^2 = (b - a)^2/12. \text{ (Anderson, 2008)}$$

$$= 0 \text{ elsewhere}$$



The Excel simulation module created for this population distribution allows the user to select values for parameters a and b . Consider the following case associated with estimating the difference between the population means when both populations are continuous uniform probability distributions. For illustration purposes, suppose the first population distribution has parameters $a = 20$ and $b = 80$, and the second population distribution has parameters $a = 10$ and $b = 50$. Then, $E(x) = \mu_1 = (20 + 80)/2 = 50$, $\text{Var}(x) = \sigma_1^2 = (80 - 20)^2/12 = 300$, and the standard deviation $\sigma_1 = \sqrt{\text{Var}(x)} = 17.321$ for the first distribution, while $E(x) = \mu_2 = (10 + 50)/2 = 30$, $\text{Var}(x) = \sigma_2^2 = (50 - 10)^2/12 = 133.33$, and the standard deviation $\sigma_2 = \sqrt{\text{Var}(x)} = 11.547$ for the second distribution. If independent random samples of size $n_1 = n_2 = 30$ are selected from these two populations, it follows that the mean of the

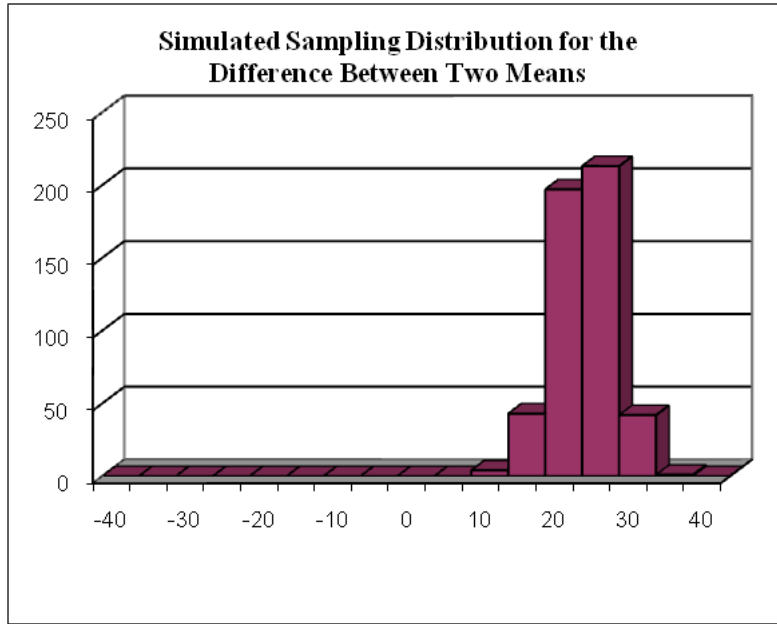
sampling distribution for $(\bar{x}_1 - \bar{x}_2)$ is $(\mu_1 - \mu_2) = (50 - 30) = 20$ and the standard deviation is $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} =$

$$\sqrt{\frac{300}{30} + \frac{133.33}{30}} = 3.8006.$$

Figure 1 provides the histogram and descriptive statistics for one iteration of this simulation example. Note that when samples of size 30 have been selected from two continuous uniform probability distributions, the

simulated sampling distribution's shape is approximately normal and the mean and standard deviation are close to $(\mu_1 - \mu_2)$ and $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ respectively.

Figure 1



Numerical Descriptive Measures

Population Mean = 20.

Population Std. Dev. = 3.8006

Simulated Sampling Distribution Mean = 20.0062

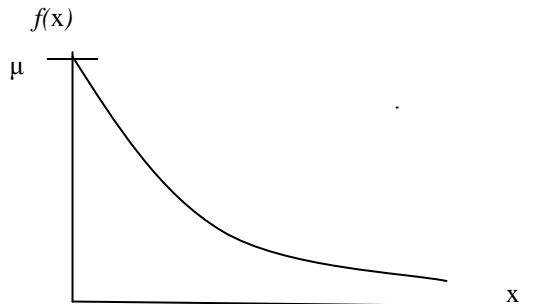
Simulated Sampling Distribution Std. Dev. = 3.8988

RESULTS WHEN BOTH POPULATIONS ARE EXPONENTIALLY DISTRIBUTED

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

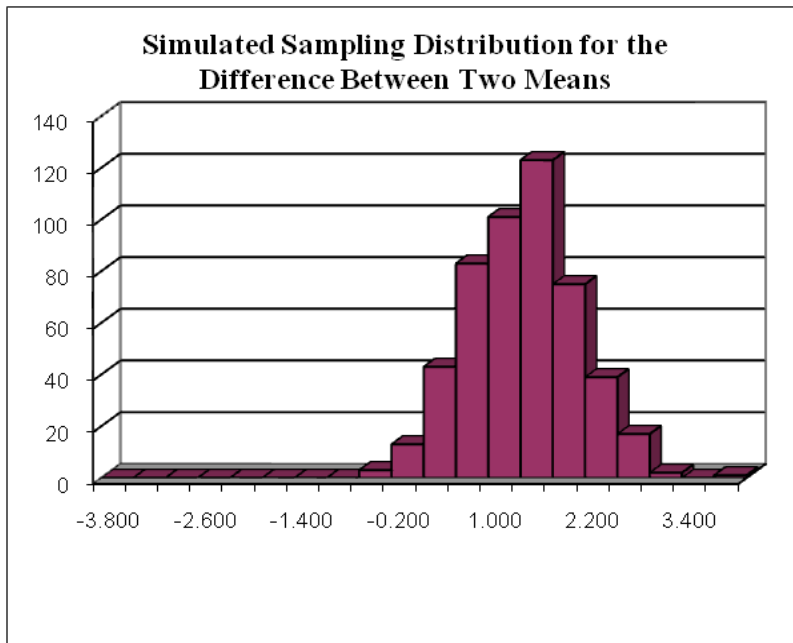
$$f(x) = \begin{cases} \mu e^{-\mu x}, & \text{for } x \geq 0, \mu > 0 \\ 0 & \text{elsewhere} \end{cases}, \text{ where } E(x) = 1/\mu \text{ and } \text{Var}(x) = \sigma^2 = 1/\mu^2. \text{ (Naylor, 1968)}$$



The Excel simulation module created with this population distribution allows the user to select values for the parameter μ . Consider the case associated with estimating the difference between the population means with both populations are exponential probability distributions. For illustration purposes, suppose the first population distribution has $\mu_1 = 1/3$ as a parameter. Then, $E(x) = 1/\mu_1 = 3.0$, $Var(x) = \sigma_1^2 = 1/\mu_1^2 = 9.0$ and the standard deviation $\sigma_1 = \sqrt{Var(x)} = 3.0$. Let $\mu_2 = 1/2$ for the second population distribution. It follows that $E(x) = 1/\mu_2 = 2.0$, $Var(x) = \sigma_2^2 = 1/\mu_2^2 = 4.0$ the standard deviation $\sigma_2 = \sqrt{Var(x)} = 2.0$. If independent random samples of size $n_1 = n_2 = 30$ are selected from these two populations, it follows that the mean of the sampling distribution for $(\bar{x}_1 - \bar{x}_2)$ is $(\mu_1 - \mu_2) = (3 - 2) = 1$ and the standard deviation is $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{9}{30} + \frac{4}{30}} = 0.6583$.

Figure 2 provides the histogram and descriptive statistics for one iteration of this simulation example. Just as with the two continuous uniform probability distribution example, when samples of size 30 have been selected from two exponential probability distributions, the simulated sampling distribution's shape is approximately normal and the mean and standard deviation are close to $(\mu_1 - \mu_2)$ and $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ respectively.

Figure 2



Numerical Descriptive Measures

Population Mean = 1.0000
 Population Std. Dev. = 0.6583

Simulated Sampling Distribution Mean = 1.0072
 Simulated Sampling Distribution Std. Dev. = 0.6614

RESULTS WHEN THE FIRST POPULATION IS UNIFORMLY DISTRIBUTED AND THE SECOND POPULATION IS EXPONENTIALLY DISTRIBUTED

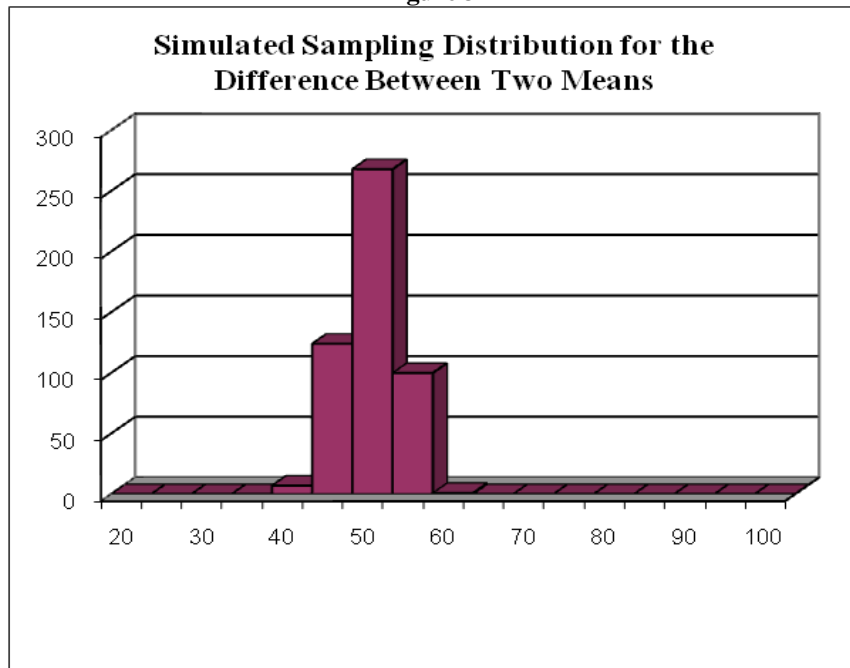
In the previous two examples, the population distributions have both been selected from the same family of distributions. This does not need to be the case, however, because the Central Limit Theorem applies to random

samples selected from any population with a mean of μ and a standard deviation of σ . Thus, consider the case where the first population has a continuous uniform probability distribution, while the second population is exponentially distributed. As before, the parameters associated both population distributions can be modified to allow for the simulation of a wide variety of populations within each family. For illustration purposes, one population distribution from each of the two earlier examples will be used. That is, suppose the first population distribution is uniformly distributed with parameters $a = 20$ and $b = 80$, and the second population distribution is exponentially distributed with parameter $\mu = 1/3$. Then, $E(x) = \mu_1 = (20 + 80)/2 = 50$, $\text{Var}(x) = \sigma_1^2 = (80 - 20)^2/12 = 300$, and the standard deviation $\sigma_1 = \sqrt{\text{Var}(x)} = 17.321$ for the first distribution, and $E(x) = 1/\mu_2 = 3.0$, $\text{Var}(x) = \sigma_2^2 = 1/\mu_2^2 = 9.0$ and the standard deviation $\sigma_2 = \sqrt{\text{Var}(x)} = 3.0$ for the second distribution. If independent random samples of size $n_1 = n_2 = 30$ are selected from these two populations, it follows that the mean of the sampling distribution for

$$(\bar{x}_1 - \bar{x}_2) \text{ is } (\mu_1 - \mu_2) = (50 - 3) = 47 \text{ and the standard deviation is } \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{300}{30} + \frac{9}{30}} = 3.2094.$$

Figure 3 provides the histogram and descriptive statistics for this simulation example. Note that even though the two population distributions were selected from different probability distribution families, the simulated sampling distribution's shape is still approximately normal and once again the mean and standard deviation are close to $(\mu_1 - \mu_2)$ and $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ respectively.

Figure 3



Numerical Descriptive Measures

Population Mean = 47.

Population Std. Dev. = 3.2094

Simulated Sampling Distribution Mean = 47.0412

Simulated Sampling Distribution Std. Dev. = 3.2249

CONCLUSION

The objective of this paper has been to develop a better understanding of the Central Limit Theorem's appropriateness when comparing the difference between the means of any two populations. Microsoft® Excel provides the opportunity to create simulations that demonstrate this non-intuitive theorem. It can be clearly observed that the simulated sampling distributions for the difference between two means follow a normal probability distribution fairly closely for samples of size 30. This approximation is not as good as the sample sizes drop farther and farther below 30; however, the approximation is even better for samples larger than 30 in size. The simulations also illustrate that the mean and standard deviation for the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$ are $(\mu_1 - \mu_2)$, and

$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ respectively. Only the continuous uniform probability distribution and the exponential probability distribution were considered as population distributions in this paper. However, these same results can be illustrated with the use of any two population distributions. By demonstrating these simulations in a statistics class, students will gain a clearer understanding and a better appreciation of the usefulness of the Central Limit Theorem in statistical analyses.

David H. Moen is a professor of decision sciences in the Beacom School of Business at the University of South Dakota. He also serves as the chair for the School's Division of Economics, Decision Sciences and Management Information Systems. He received his Ph.D. in Statistics from Oklahoma State University.

John E. Powell is a professor emeritus of decision sciences in the Beacom School of Business at the University of South Dakota. He previously served as the chair for the School's Division of Economics, Decision Sciences and Management Information Systems. He received his doctorate in Quantitative Business Analysis from Indiana University, Bloomington.

REFERENCES

1. Anderson, David R.; Sweeney, Dennis J.; Williams, Thomas A.; *Statistics for Business and Economics (10th edition)*, Thomson South-Western Publishing, 2008.
2. McClave, James T.; Benson, George P.; Sincich, Terry; *Statistics for Business and Economics (9th edition)*, Pearson Prentice Hall, 2005.
3. Microsoft® Office Excel® 2007, Copyright© Microsoft Corporation 2006.
4. Moen, David; Powell, John; "Illustrating the Central Limit Theorem through Microsoft Excel Simulations", *The International College Teaching Methods and Styles Journal*, Volume 1, Number 2, April 2005.
5. Naylor, Thomas; Balintfy, Joseph; Burdick, Donald; Chu, Kong; *Computer Simulation Techniques*, John Wiley & Sons, Inc., 1968.