# AW4 Use SPSS to demonstrate the Central Limit Theorem

We can demonstrate the Central Limit Theorem by using SPSS to create a series of random samples of size n, calculating each sample mean, and using descriptive statistics (mean, histograms) to show the effect of increasing the number of samples (or sample size) on these descriptive statistics. To demonstrate the Central Limit Theorem, we will sample from a uniform probability distribution. The uniform distribution is very different to the normal distribution as illustrated in Figure 1.



Figure 1

**Example 1**

Let us decide that the value of the minimum and maximum value in the uniform distribution are given as minimum, a = 0, and maximum, b = 10. The uniform distribution has the property that all subintervals within the range ‘a’ to ‘b’ have the same probability of occurrence no matter where they are located, i.e. the value of x can take any real number value between ‘a’ and ‘b’. Theoretically, we can calculate the mean and standard deviation for any uniform distribution if we know the value of the minimum and maximum values:

Minimum a = 0

Maximum b = 10

Mean = (b + a)/2

Mean = (b + a)/2 = (0 + 10)/2 = 5

Variance = (b – a)2/12

Variance = (b – a)2/12 = (10 – 0)2/12 = 8.3333333333’

Standard deviation = $\sqrt{variance}$ = 2.886751346 …..

Now we can use SPSS to create 40 random samples from this uniform distribution, where each sample is of size 100. From these samples, we can then calculate the mean value for each of the 40 samples and then plot a histogram to see how these sample means vary in value. Furthermore, we will calculate the mean and standard deviation of the 40 sample means and compare with what the Central Limit Theorem predicts.

The Central Limit Theorem predicts that as we increase the size of the random sample from the uniform distribution, then:

1. the distribution of the sample means from the uniform distribution should become more normally distributed,
2. the overall mean of all sample means should become closer to the mean of the normal distribution, and
3. the corresponding standard error of the sample means should become closer in value to what the normal distribution predicts:

Summary statistics

Mean of the sample means = 5

Standard error of the sample means = $^{σ}/\_{\sqrt{n}}= ^{2.886751346…..}/\_{\sqrt{40}}=0.456435464…$

To illustrate the use of SPSS we will undertake the calculations for 5 samples and then 40 samples where each sample is of size 100. This will result in 100 sample means of size 5 and then 100 sample means of size 40. We can then compare the results and see that the Central Limit Theorem applies – please note that we are demonstrating the Central Limit Theorem using SPSS but to prove this theoretically requires a great deal of advanced mathematics.

SPSS procedure – initial set-up:

1. Name the first column X1
2. Enter a number in the 100th cell of this first column that will define the number of values in sample X1. Hint – you may find that in the SPSS Data View menu that only 40 rows are visible. If you click on this cell you will find that another 40 rows will appear. Use this method to be able to access the 100th cell (row 100) and place any number there. You should have no other numbers after row 100.
3. Use Transform > Compute Variable to generate the sample values for X1.
4. Repeat for X2, X3, X4, X5 (each sample is in a row of size 5)
5. Use Transform > Compute Variable to calculate the 100 sample means and store in X6.
6. Use Analyze > Descriptive Statistics > Frequencies menu to calculate the mean of all 100 sample means and plot the histogram. These results will be stored in the SPSS Output file.

(a) Generate 5 samples of size 100

Step 1 – 2 completed.

Step 3 Transform > Compute Variable to generate the sample values for X1.

Step 4 Repeat for X2, X3, X4, X5.



Figure 2



Figure 3

Step 5 Use Transform > Compute Variable to calculate the 100 sample means and store in X6.



Figure 4



Figure 5

Step 6 Use Analyze > Descriptive Statistics > Frequencies menu to calculate the mean of all 100 sample means and plot the histogram. These results will be stored in the SPSS Output file.

* Transfer variable xbar into Variables box.
* Uncheck ‘Display frequency table’.
* Click on Statistics and select Mean and Std deviation.



Figure 6

Click on Charts and select Histogram (also chosen Show normal curve on histogram)



Figure 7

The final menu is as illustrated in Figure 8.



Figure 8

Click OK and the results will appear in the SPSS Output file as illustrated in Figure 9.

Save the SPSS data and output files (Part 2 Chapter 5 Sampling\_5.sav, Part 2 Chapter 5 Sampling\_5.spv).



Figure 9

From SPSS, we observe that the histogram for the sample means does not look like the uniform distribution that we sampled from but looks more like a normal distribution. Furthermore, the mean = 4.91 and standard deviation for the sample means = 1.343. This can be compared to what we would expect if this was a normal distribution where the mean = 5 and standard error = 0.456. We conclude, the two means are quite close, but the standard errors are quite different. What would happen if we increased the sample size from 5 to 40?

(b) Generate 40 samples of size 100

Repeating the steps above so that we have X1, X2, X3,……….., X40, where each column has 100 data values. The results are presented in Figure 10



Figure 10

From SPSS, we observe the histogram is a better fit to the normal distribution, is more centred at xbar = 5, and is less spread out around the value of the mean of the sample mean when compared to a sample size of 5 illustrated in Figure 9. Furthermore, the mean = 4.99 and standard deviation for the sample means = 0.496. This can be compared to what we would expect if this was a normal distribution where the mean = 5 and standard error = 0.456. We conclude, the two means and standard errors are quite close, and the values are as predicted by the Central Limit Theorem.

Test the sample data for normality

Finally, you could undertake a Shapiro-Wilks or Kolmogorov-Smirnov test to confirm if the sample means presented can be considered to significantly follow a normal distribution (see Figure 11).

In SPSS, Analyze > Descriptive Statistics > Explore > Plots and select Histogram and choose Normality plots and tests.



Figure 11

The Shapiro-Wilk test gives a p-value equal to 0.854.

This value is greater than the significance level of 0.05 and therefore we conclude that the sample means are normally distributed.