Exploration of the Central Limit Theorem

In this computer lab, we will explore the effect of the *Central Limit Theorem* with computer simulations. The following simulations could be constructed with many different pieces of software, but in this computer lab we will work with *Microsoft Excel*.

## Step 1

Open Microsoft Excel on your computer and start a new workbook. You may want to regularly save this workbook to your computer with an appropriate name in case your computer crashes at any stage in the simulations.

## Step 2

Re-format the worksheet so that all numbers will appear in the middle of the cells. To do this, first type “Ctrl+A”, which will select all cells, then make sure both of these middle buttons are selected.



## Step 3

Extend the width of column A and type in the column headings

normal

sample #

sample mean

value 1

into cells B1, C1, D1 and E1.



## Step 4

Click on E1 and hover the mouse over the bottom-right corner until a + symbol appears.



Click this corner and drag it to the right until you reach cell X1.



When you let go, this should create the column headings value 2, value 3, …, value 20. Your spreadsheet should now look like this.



## Step 5

Type the number 1 into C2 and then type =C2+1 into C3.



After pressing enter, C3 should now be the value 2.



This will be important for later.

## Step 6

We are now going to create our first sample of values. The Excel function RAND() will produce a random number between 0 and 1. We’re going to place one of these random number in every cell from E2 to X2. To do this most efficiently, type =RAND() into E2.



Press enter. The random value in your cell will be different to the one on this worksheet.



Now, click on E2 so that it is selected.



Then click on the bottom-right corner again and drag it to X2.



When you let go, it will have copied =RAND() into all 20 of these cells. Each of these cells now contains a random number between 0 and 1. Note also how E2 will look different to before. This is because each time Excel makes a calculation it will refresh *all* of the random numbers.



You can force Excel to refresh at any stage by pressing F9. Try it a few times and watch the cells refresh!

## Step 7

Now that we have created a random sample of 20 numbers, we’re going to calculate the average of these numbers, known as the *sample mean*. To do this type =AVERAGE(E2:X2) into D2. You will notice Excel highlight the range that you’re finding the average of.



Press enter to see your sample mean. You might ordinarily expect this sample mean to be $0.5$ (halfway between 0 and 1) but because this is a random sample there will always be some variation. Press F9 a few times and keep an eye on how high and low the sample mean can go.



This is our primary interest; observing these possible values of the sample of mean of a random sample.

## Step 8

Next, select cells D2 through to X2.



And by clicking again on the bottom-right corner, drag everything down one row.



After letting go, this will create a *second* sample of 20 random values, with their own sample mean stored in D3.



## Step 9

We currently have 2 independent sample means, but we want more. Many more! Select D3 through to X3, click again on the bottom-right corner, and drag down until you’re at row 1001 in the spreadsheet, and let go.









We have now created 1000 random sample, and calculate 1000 independent sample means. Also, every time you press F9, *everything* will refresh. Try it, but perhaps save your file first!

## Step 10

Scroll back to the top of the worksheet.

Our goal is to paint a picture of these 1000 sample means and this is best achieved with a histogram.

To do this, hover your mouse of the letter D at the top of column D until you see a downward pointing arrow.



Click, and this should select the entire column D.



In effect, we have selected all of the sample means. Now click on Insert.





Then look for this little symbol under charts.



Click on it and select the first type of Histogram chart.



This will generate a histogram of the 1000 sample mean values.



Change the title to Sample Means. You can do this by simply clicking on the chart title and editing it.





Next, right-click on the numbers under the histogram bars and select Format Axis.



Click on the third symbol on the right.



And choose for there to be 20 bins.



You can fiddle around with other options if you like, such as changing the colour, etc.

Finally, move the axis across to the top of column A. You might want to increase the width of column A so that it fits nicely.



Press F9 a few times and watch the histogram change! What do you notice?

## Step 11

Let’s also create a histogram of the original distribution. To do this, repeat Step 10 but with column E instead of column D (you could actually choose any of the 20 columns, it won’t matter). Name the histogram Original Distribution and place it under the histogram of the sample means.



Press F9 a few times.

## Step 12

The *Central Limit Theorem* predicts that as the sample size (in this case 20) gets larger, then the sample mean will follow a *Normal distribution*, no matter the distribution of the original sample. Furthermore, if $μ$ and $σ^{2}$ are the hypothetical mean and variance of the original distribution and $n$ is the sample size, the *Central Limit Theorem* states that the sample mean, known as $\overbar{X} \~ N\left(μ,{σ^{2}}/{n}\right)$.

To confirm this result, we are going to create a third histogram *just* for $N\left(μ,{σ^{2}}/{n}\right)$ and see if it looks similar to our histogram for the sample mean. This is where we will use column B.

In cell B2 type =NORM.INV(RAND(),0.5,sqrt(1/12)/sqrt(20)).





This will generate a random number from $N\left(μ,{σ^{2}}/{n}\right)$. In this example, $μ=0.5$, $σ^{2}={1}/{12}$ and $n=20$. In general you would type =NORM.INV(RAND(),$μ$,sqrt($σ^{2}$)/sqrt($n$)).

Drag this cell down to row 1001, create a histogram for column B and call it Normal from CLT. Press F9 a few times and confine yourself that you’re obtaining a similar distribution to that of the sample means.



Step 13 (optional)

In completing Steps 1 to 12 we have seen the Central Limit Theorem in action.

Having start from a Uniform distribution, where all values between 0 and 1 are equally likely,



we found that the sample mean from a sample of size $n=20$ will follow a Normal distribution, $N\left(μ,{σ^{2}}/{n}\right)$. For this Uniform distribution we used, $μ=0.5$ and $σ^{2}=1/12$.

Another aspect of the Central Limit Theorem is that it doesn’t matter what the original distribution is, the sample mean will always behave in this way, following $N\left(μ,{σ^{2}}/{n}\right)$.

Let’s explore two other initial distributions.

Name this tab uniform.



Right-click and select Move or Copy.



We’re going to duplicate this tab twice and name the new tabs exponential and quadratic. Simply select Create a copy.



Then rename the tab.



In the exponential tab, we will need to make very few adjustments. The new underlining distribution is called the Exponential distribution.



It can produces value from $0$ to $\infty $, but as can be seen lower values are much more likely. This distribution has the properties $μ=1$ and $σ^{2}=1$.

All we need to do in the exponential tab is to change two formulas.

Firstly, in cell E2 type -LN(1-RAND()). This formula will produce values from the Exponential distribution. Then drag this to X2, and then drag the whole sample down to row 1001.

The first two histograms will now be functioning automatically.

Secondly, in cell B2 type =NORM.INV(RAND(),1,SQRT(1)/SQRT(20)) and drag this down to row 1001.

Again, press F9 a few times and convince yourself that the sample means are following the expected Normal distribution predicted by the *Central Limit Theorem*.



In the quadratic tab, we will be using the following distribution.



This can produce numbers between 0 and 2, but values near 1 are very unlikely. This distribution has the properties $μ=1$ (even though 1 is almost impossible) and $σ^{2}=0.6$.

Again, we make the two slight changes. In cell E2 type =1+(2\*RAND()-1)^(1/3), which will always generate a random number from this distribution. Drag this across to X2, and then everything down to row 1001.

Then in B2 type =NORM.INV(RAND(),1,SQRT(0.6)/SQRT(20)) and drag everything down to row 1001.

Press F9 and compare.



Step 14 (optional)

It’s pretty amazing, isn’t it! If you’re still having fun playing around with the Central Limit Theorem, there are two things you could do.

1. Try some other distributions. You’ll need two things. Firstly, the inverse of its *cumulative distribution function* (CDF). This is needed to generate the original values in E2. The respective CDFs used so far in this worksheet were $F\left(x\right)=x$, $F\left(x\right)=1-e^{-x}$ and $F\left(x\right)=\frac{1}{2}\left[\left(x-1\right)^{3}-1\right]$ hence their inverses were $F^{-1}\left(x\right)=x$, $F^{-1}\left(x\right)=-ln⁡(1-x)$ and $F^{-1}\left(x\right)=1+\sqrt[3]{2x-1}$. Each time, replace $x$ with RAND() in Excel. Secondly, you’ll need to know the theoretical mean $μ$ and variance $σ^{2}$ of your distribution and simply use =NORM.INV(RAND(),$μ$,sqrt($σ^{2}$)/sqrt($n$)) in D2.
2. Try larger values for $n$. We’ve been using $20$, but the *Central Limit Theorem* states that everything becomes more accurate as $n$ gets larger