Excel Exercise: Demonstration of the Central Limit Theorem

In this exercise, you will …
1. Create tables of random numbers on multiple sheets in an Excel workbook;
2. Calculate means of subsets of these random numbers;
3. Calculate the descriptive statistics of the random numbers and of the means;
4. Compare the frequency distribution of the means with the *normal distribution*;
5. Graph the frequency distributions of the random numbers, means, and normal distribution;
6. Observe the behavior of these distributions as the set of random number changes.
7. Interpret your observations in terms of the Central Limit Theorem

In completing this exercise, you will probably learn a few new tricks in working with Excel. The primary text reference for this exercise is Triola, Ch. 5, esp. Section 5-5.

Before starting this exercise, complete the worksheet on the Uniform Distribution.

IMPORTANT: Read each pair of paragraphs, the corresponding "Operation" and "Comments," COMPLETELY before actually carrying out the operation described. Do not proceed to the next step until you understand what you have done and what has happened as a result.

Operation What you do	Comments What you see and why
1. Open Excel from the Windows Start command.	1. You should see an empty spreadsheet grid, with the cursor in cell *A1* of *Sheet 1*. The worksheet should will probably have 3 sheets as indicated on the lower tabs.
2. Open the file *16Sheets.xls*. Check to see if you have a total of 16 sheets by clicking on the ▶| symbol in the lower left-hand corner next to the tab for *Sheet 1*.	2. The tab for *Sheet 16* should become visible after you click on the ▶| symbol. Note that you can jump from one end of the worksheet tabs to the other by pressing the ▶| and |◀ symbols. Clicking on the ▶ and ◀ symbols will move the row of tabs one sheet tab to the right or left, respectively. To select a specific sheet, however, you need to click on the tab for that sheet.
3. Click on the |◀ symbol to return to the beginning of the worksheet tabs if necessary. Then click the tab for *Sheet 1* with the <u>right</u> mouse button (🖰), and then click on "**Select All Sheets**."	3. Whenever you click on the tab for a *Sheet 1*, it becomes highlighted (white rather than gray). When you "select all sheets," the sheets are "*grouped*" together, all their tabs are highlighted, and anything you do on one sheet will also be done on all the other sheets in the group. Note that the label for ***Sheet 1*** appears in bold type. This indicates that this is the sheet you are actually viewing.
4. Before entering any data, save this workbook <u>with a new name</u>, **CENTLIM.XLS** on your disk. (Use **Save As** under the **File** menu.)	4. This is your working file. It is a good idea to save your file periodically by pressing $\boxed{\text{Ctrl}}+\boxed{\text{S}}$ or clicking on the disk symbol (🖫) on the upper toolbar.

The next series of steps creates a large quantity of "random numbers," 1,000 on each of the sixteen sheets. These numbers will serve as the population to be analyzed. These should approximate the behavior of a uniform probability distribution where the occurrence of each value of *x* is equally probable for $0 \leq x \leq 1$.

Operation	Comments

5. Click on cell *A1* on *Sheet 1*. Type in the formula **=rand()** and press the [Enter] key. A decimal number between *0* and *1* should appear in cell. Click on the tabs of the other sheets. A different number will appear in cell *A1* of these sheets.

5. The function *RAND()* generates a random number between *0* and *1*. Every time Excel does a calculation, a new random number is generated, and a <u>different</u> number is produced for each occurrence of the function, which is why the numbers are different on every sheet. You can confirm that the numbers change with each calculation by pressing the [F9] key, which causes the Excel to recalculate all the functions in the workbook.

6. Return to *Sheet 1* by clicking on its tab. Regroup all 16 sheets as described in step (3) above. Click on cell *A1* to make it the active cell. Then move the mouse pointer over the small square in the lower right-hand corner of the cell (the ***Fill Handle***); the cursor pointer should become a small plus sign (+). When this occurs, hold down the left mouse button and drag across <u>10 columns</u> to cell *J1* so that all ten cells are highlighted. Release the mouse button. A different number appears in each of the ten cells and the entire row is highlighted.

6. You have just generated 160 different random numbers, ten on each of the sixteen sheets. The square in the lower right-hand corner of the active cell is called the *fill handle*. It can be used to copy numbers, formulae, or series of numbers across a row of cells or up or down a column of cells. Do you understand why each of the ten numbers on *Sheet 1* is different from the others, even though they each contain the same function?

Include your answer this question in your response to this exercise..

7. With the entire row still highlighted (do NOT click on any of the cells) drag the *fill handle* (the small square in the lower left hand corner of cell *J1*) down to cell *J100*. Release the mouse button.

7. When you released the mouse button, all the cells from *A1* to *J100* on each of the sixteen worksheets should have filled, each with a different random number, for a total of 16,000 random numbers, presumably uniformly distributed between 0 and 1.

8. Press the function key [F9] several times. Note the values all change each time. Look at other sheets. They will have different values as well and these also change when you press [F9]. Before moving to the next step, <u>be sure all sheets are grouped and all filled cells are still highlighted.</u>

8. Each value of *rand()* is recalculated each time you press [F9], a key you can use to recalculate the spreadsheet at any time, without entering new data. Normally all cells are also recalculated each time you make a change or enter something into any cell.

9. While all 16 sheets are grouped, press [Ctrl]+[C] to copy the cells; a flickering dashed line will appear around them. On the **Edit** menu, select **Paste Special** and then the buttons for **Paste: Values** and **Operation: None**; press **OK**. Press [Ctrl]+[Home] to return to cell *A1* and then the [Esc] key to end the copy selection.

9. Note that the *=rand()* in A1 has become a number, and there are different values in each cell on every sheet. When you copy a formula and then Paste Special: Values instead of the normal Paste, the formula is evaluated. Each occurrence of the rand() function is separately evaluated in this operation. Pressing [Ctrl]+[Home] always takes you to cell *A1* at the top of a page.

10. Click on the tab for *Sheet 1* with the <u>right</u> mouse button (🖱), and select "**Ungroup sheets**." Save your file by pressing [Ctrl]+[S] or clicking on the disk icon (💾) on the upper toolbar.

10. This is the last reminder to save your file. But remember to do it periodically, in case of a power failure or other gremlin.

You are ready to add a two more sheets and analyze the data you have created, first by taking the mean of a large number (1,000) of samples, each sample consisting of sixteen units of the population of 16,000 uniformly distributed random numbers.

Operation	Comments

11. With *Sheet 1* selected as the <u>only</u> active sheet, press [Shift]+[F11] <u>two</u> (2) times. Click the <u>right</u> mouse button on the tab of each of the two new sheets that appear, select "**Rename**," and name the very first sheet (originally *Sheet 18*) "**Analysis**," and the second (originally *Sheet 17*) "**Means**."

11. The key combination [Shift]+[F11] is a "keyboard shortcut" which enables you to do a routine process quickly. You can find all the keyboard shortcuts by double clicking on the <u>Help</u> button or selecting <u>Contents</u> from the Help menu and entering the word **keyboard**.

12. Select cell *A1* of the *Means* sheet. Be sure <u>none</u> of the sheets are grouped together. Enter the formula, =**average(sheet1:sheet16!a1:a1)** in the cell.

12. A number, probably between 0.4 and 0.6, should appear in the cell. This is the mean (average) of the sixteen random numbers in cell *A1* on each of the sixteen sheets underneath it. The "sheet1:sheet16!" tells Excel you want to include the relevant cells on *Sheets 1-16*, but the cell range on each sheet only includes cell *A1.*.

13 Be sure cell *A1* of the *Means* sheet is selected. Then, just as you did in steps and , drag the *fill handle* of the cell to cell *J1* and then, after releasing the mouse button briefly, drag the *fill handle* from cell *J1* down to cell *J100* and release the mouse button. When all the cells from *A1* to *J100* are filled with numbers, press [Ctrl]+[Home] to return to cell *A1* and unselect the remaining cells. Look at the range of values in the cells and compare them with the values in the cells on *Sheets 1-16*.

13. When it stops filling the cells and does its calculation, Excel presents you with 1000 numbers; each number is the mean of the 16 different random numbers in the cells in the corresponding row and column in the sheets below it. Before proceeding and doing the actual calculations, think about how these means compare to the mean of the entire set of 16,000 random numbers. Why are they all not the same? What is the mean of the means? What do you think is the approximate value of the standard deviations of the entire set of 16,000 random numbers and of the set of 1000 means (Hint: Remember that it's a uniform random distribution)?

In the next few steps, you will calculate the descriptive statistics of both the population and the sample. To save you entering a large number of individual formulae, the instructions will guide you through several labor-saving tricks available in Excel. After you complete each step, you should pause briefly and analyze exactly what you did and what it accomplished.

14 Select *Analysis* as the active sheet. In cell *A1*, enter your name. Click on cell *I1* and type today's date in the form *mm/dd*. Then starting in row 23, type in the following table:

14. Since it is not preceded by an = sign, Excel interprets the "*mm/dd*" as a date; note that the date may appear differently in the cell, depending on the default format and appears as "*m/dd/yyyy*" in the formula bar.* The remainder of this step sets up the table for the first set of descriptive statistics. Format cells *A25-A30* as right justified and *B23-C24* as centered.

* Another way of assuring that "*mm/dd*" is recognized as a date is to press [Ctrl]+[D] (you will not notice any effect) before entering the date.

	A	B	C
23			Sample
24		Population	Means
25	Count		
26	Minimum		
27	Median		
28	Mean		
29	Maximum		
30	StDev		

| | Operation | | Comments |

| | **Operation** | | **Comments** |

15. In cell *B25*, enter the formula
=COUNT(Sheet1:Sheet16!a1:j100) Be sure
to include the "!" and "$" as indicated.

15. The number *16000* should appear in cell *B25*, if
everything up to this point has been done correctly.
As in step , the "sheet1:sheet16!" indicates you want
to include data on all of those sheets, but this time,
the "a1:j100" indicates you want to include all
the data on each sheet using <u>absolute</u> references, *i.e.*,
when the formula is copied, the cell references stay
constant. Note the *a* and *j* become capitalized.

16 Then select cell *B25* and drag its *fill handle*
down to cell *B30*. The formula in *B25* is copied
exactly to cells *B26-B30*. Click on each of these
cells and in the formula bar, edit the formula in each
of these cells so that it contains the function
indicated, with the same cell range, as indicated
below:

	A	**B**	**C**
23			Sample
24		Population	Means
25	Count	=count(sheet…	
26	Minimum	=min(sheet…	
27	Median	=median(sheet…	
28	Mean	=average(sheet…	
29	Maximum	=max(sheet…	
30	StDev	=stdev(sheet…	

16. By copying the formula down the column, you
retain the references to the same range of cells, the
16,000 numbers in cells *A1-J100* on *Sheets 1-16*.
Thus you only need to edit the name of the function.
The easiest way to edit the function name is to
highlight just the name with the mouse, type in the
new function name, and then press ⌁Enter⌁ or click
with the mouse on the green check mark (√) on the
formula bar. Examine the values that appear in cells
B26-B30. Are they consistent with your
expectations? If not, or you think there is an error,
you should probably consult with your faculty
before proceeding. If necessary, format the value in
B26 (the Minimum) to standard decimal notation,
e.g., 0.0003 instead of the Scientific notation form of
3.2E-4 that may appear. To do this, on the menu bar
select <u>Format</u> and <u>Cells</u> (or press ⌁Ctrl⌁+⌁1⌁ for a
keyboard shortcut) and select <u>Category</u> *Number*.
Press <u>OK</u>. Then use the <u>Increase Decimal</u> (🔘)
and/or <u>Decrease Decimal</u> (🔘) buttons in the formula
bar to get an appropriate number of digits..

17. To get the descriptive statistics for the sample
means, first highlight cells *B25-B30* and drag the *fill
handle* across to column *C*. Then edit the function
in each cell by replacing the sheet reference
Sheet1:Sheet16! with the reference **means!**, so that
the range reference for each formula reads
**means!$A$1:$J$100**.

17. By this time you should have a good
understanding of what you have done and what the
numbers in cells *C25-C30* mean. Are they
consistent with what you predicted in step 13? Are
there any that surprise you?

18. Now calculate two more numbers, which should
help characterize the means: In cell *B31* type
Sample Size and in *B32* type **Exp. StDev of Means**;
format both cells as <u>right justified</u>. Then in cell *C31*,
enter the formula, **=count(Sheet1:Sheet16!a1:a1)**,
and in cell *C32* the formula, **=B30/sqrt(C31)**.

18. According to the Central Limit Theorem, the
standard deviation of the sample means of a
population is approximately equal to the standard
deviation for the population divided by the square
root of the size of the samples. Is your result
consistent with this?

Now that you have calculated the descriptive statistics, you will determine the frequency distributions of both the
populations and the means and compare the distribution of the sample means with the normal distribution. This
uses the "array formula" for frequency, which requires that you enter the formula in an unusual way, as described
in Step 21.

	Operation		Comments

Operation

19. Start by entering the headings shown in the opposite column in the cell range *E23-I25* and center them. Then select the range *G23-I23*, and click on the Center Across Columns (or Merge and Center) button (⊞) on the toolbar to center the word "Frequency" across those three cells.

Comments

19. Type in the labels as shown below; the final form will have "Frequency" centered across cells *G23-I23*.

	E	F	G	H	I
23			Frequency		
24	Top	Midpoint		Sample	Normal
25	of Range	of Range	Population	Means	Distribution

20. In cell *E36*, (note: *E36*, not *E26*), type the formula =**$B$28**; be sure to include both $ signs. In *E35* enter =**E36-0.05** (no $ signs this time), and with *E35* as the active cell, drag its *fill handle* up to *E27*. Enter the number **0** in cell *E26*. Then, in cell *E37*, enter the formula =**E36+0.05** (again, no $ signs) and using a *fill handle* copy this formula down to fill cells *E38-E45*. Enter the number **1** in cell *E46*. Finally, enter the formula =**(E26**+**E27)/2** in cell *F27* and copy this formula to fill cells *F28-F46*. Select all the cells *E26-F46* and use the Increase Decimal button to set them to three decimal places, *e.g.*, 0.498.

20. By this time you should be familiar with the techniques of entering and copying formulae. Column *E* should end up with a series of numbers, with the mean of the population data as the midpoint and with a difference of 0.050 between adjacent numbers, except that the endpoints are fixed at 0.000 and 1.000. Similarly column *F* contains a similar set of numbers, each halfway between the adjacent number in column *E* and the one just above it. These two columns contain the upper limits and midpoints, respectively, of the "bins" of frequency distribution.

21. Starting with cell *G26* select the empty cells *G26-G46*. Without pressing ⌨Enter⌨, type the formula =**frequency(sheet1:sheet16!a1:j100,e26:e46)**; then hold down both the ⌨Ctrl⌨ and ⌨Shift⌨ keys and press the ⌨Enter⌨ key.

Click on cell *G47*, then on the Σ symbol in the upper toolbar; confirm that the formula =*SUM(G26:G46)* appears in the toolbar and press ⌨Enter⌨. The total should be 16,000.

21. Excel's *frequency* function has the form FREQUENCY(data-array,bins-array), and creates an array of numbers showing the frequency of the numbers in the full set of cells identified by the data-array, according to the bins set in the bins-array. When entering any array function such as frequency, Excel requires that you press ⌨Ctrl⌨+⌨Shift⌨+⌨Enter⌨ rather than ⌨Enter⌨ alone. Column *G* should now contain the frequency distribution of the population of 16,000 random numbers, centered around the mean of those numbers

22. Now determine the frequency distribution of the 1000 means by highlighting cells *H26-H46*, typing in the function =**frequency(means!a1:j100,e26:e46)**, and pressing ⌨Ctrl⌨+⌨Shift⌨+⌨Enter⌨. Confirm that the total number of means is 1000 by calculating the sum of H26:H46 in H47.

22. Compare the frequency distributions of the population and the sample means. How do they differ and why? Are they consistent with your predictions of the standard deviations in step 13? Take a few minutes to review what you have done to this point and make sure you understand the population data, the sample means, their descriptive statistics and their frequency distributions.

The next step calculates the frequency distribution of the for 1000 numbers with the same mean and standard deviation as those for the sample means.

Operation	Comments
23. In cell *I27* type the following formula: **=(NORMDIST(E27,$C$28,$C$30,TRUE)-NORMDIST(E26,$C$28,$C$30,TRUE))*1000** and copy it down to fill cells *I28-I46*. Format the numbers in these cells as integers, *i.e.*, no digits after the decimal point, by pressing Ctrl+1, and selecting Category *number* and Decimal Places *0* (Code *0* in Excel 5) as the format rather than the *scientific* notation category the numbers appear in initially. Compare the values in column *I* with those in column *H*. Also confirm that the sum of the values in column *I* is also 1000.	23. The *NORMDIST* function calculates the value of the *normal distribution* for a given *x* value. The formula in Excel is *NORMDIST(x,mean,stdev, cumulative)*, where the variable *cumulative* is a switch that tells it whether to calculate the value of the distribution curve at the point *x* (*cumulative*=FALSE) or to calculate the total area under the curve to the value *x* (*cumulative*=TRUE). By subtracting the value of the function for *x-0.05* from its value at *x*, we obtain the area under the curve between the two values. By multiplying it by 1000, the total number of sample means, for values of *x* between 0 and 1, we obtain the expected frequency distribution of means if they were normally distributed.

The penultimate step in this analysis is to display the three frequency distributions graphically. We will use the *midpoints* of the ranges rather than the *tops* for easier interpretation of the graph. To save space, the "Operation" instructions will be presented across the entire page, since at least the initial steps should be relatively familiar to you from other charting exercises and the remainder should be largely self-explanatory otherwise. There are a number of new "tricks" of chart formatting that you will be shown, however, that you may be useful to you in the future.

24. Create your chart by highlighting the cells *G27-I46*. Then click on the Chart Wizard button. The ChartWizard Step 1 of 4 box should appear. Select Line-Column on 2 axes as the *Custom chart* type, click on Next >. In the Step 2 of 4 window, confirm the selected data range *$G$27:$I$46* and under the *Series* tab, enter **=Analysis!$F$27:$F$46** as the *Category (X) axis labels*. Note in the window under series, one of the labels, "Series 1," "Series 2," or "Series 3" (they may not be in sequential order) is highlighted and a data range in a single column is listed in the Values box; the Name box is blank. When you click on different "Series" labels, the corresponding column of data appears in the Values box. Click on each label in turn and in the Name box, enter the following names to replace the default series labels: **Population** for the **Analysis!$G$27:$G$46** values, **Means** for the **Analysis!$H$27:$H$46** values, and **Norm. Dist.** for the **Analysis!$I$27:$I$46** values. Click Next >

25. In the Window for Step 3 of 4 – Chart Options, first, under "Titles," type **Central Limit Theorem Demonstration** as the *Chart Title*, **Midpoint of Range** for the *Category (X) Axis* title, **Frequency in Population** for the *Value (Y) Axis* title, **Frequency of Means** for the *Second Value (Y) Axis* title. Position the Legend at the bottom. Click Next>, be sure the button to *Place chart* as object in [the sheet] **Analysis** is selected and click Finish. Your chart should appear in the middle of your screen, presenting the three frequency distributions as lines or columns. Click on a blank area of the chart and move the chart so that the upper left corner of the chart is in the upper left corner of cell *A3*; you will probably need to scroll the screen upward to complete the move. Drag the lower right corner of the chart to the lower right corner of cell *I21*. It will be useful to do a some editing of the chart to make the data presentation clearer.

26. ***Right click*** with the mouse on one of the histogram bars of the Means display, and select Format Data Series. Under Axis select Secondary; and, if you wish, on the Patterns tab, you can change the border color and/or area color and/or pattern. Click OK when done.

27. Right click on one of the tall columns of the population distribution data (labelled series 2 in the legend) and select the *Chart Type* option; choose line for the chart type for these data and press OK. The columns have become a line and the Primary (left) axis has no longer has *0* as its lowest value. Right click on the left axis, select Format Axis, set the Scale Minimum to 0, Maximum to 900 and Major Unit to 150. Under the Patterns tab you can change the characteristics of the line and/or markers, if you wish. Press OK and note the change in the chart.

28. Right Click on one of the triangles on the line indicating <u>Normal Distribution</u> and <u>Format Data Series…</u>. On the <u>Patterns</u> tab, change the <u>line</u> <u>color</u> to something darker, with no <u>markers</u>; and under <u>Axis</u> select <u>Secondary</u>. Click <u>OK</u> when done. After saving the file (so if something goes wrong you can recover your work), explore other options for formatting other aspects of the chart by right-clicking on them. For example, can you eliminate the background shading in the plot area?

29. Study the chart and the results of the various calculations on the "Analysis" sheet; if you have thought about what you have done in setting up the calculations and constructing the chart, the data shown there should make sense. Compare the results with your calculations on the worksheet, "The Uniform Distribution," which you were asked to complete before starting this exercise. Print out one copy of the sheet (use <u>Print Preview</u> and <u>Setup</u> to be sure that the chart, descriptive statistics, and frequency distribution tables are all printed to fit on a single page; be sure the Chart is not selected before you start printing, or you will only get a printout of the chart) and submit it with this Topic's work.

30. In this step, you will restore the *RAND()* function to the cells on Sheets 1-16 in place of the fixed set of random numbers and see what happens as the numbers change. To do this, click on the tab for Sheet 1, then click on the ▶| symbol to reveal the tab for Sheet16, and <u>holding down the ⑤hift key</u>, click on that tab to select Sheets 1-16. The "Analysis" and "Means" sheets should NOT be in the selection. Now repeat steps 5-7 from the beginning of the exercise. When all the cells are converted to the *RAND()* function, click on cell *A1* and then on the tab for the "Analysis" Sheet (click on the |◀ symbol to display that tab).

31. Press the ⑤9 key, wait for Excel to complete the recalculation and note the changes, especially in the shape of the frequency distribution of the sample means. Repeat this several times. Consider what changes and what doesn't and the magnitude of the changes. Print out <u>one</u> copy of this worksheet. Press ⑤9, wait for the recalculation, and print a copy of these new data.

32. Save the worksheet, and explore another variation as follows. Under the data summary in cells A23-B32 on the "Analysis" sheet, you can calculate the means of the 1000 random numbers on each of the 16 data sheets, *i.e.*, enter =**average(sheet__!a1:j100)** in each of the 16 cells, B35-B50, where each cell has a different number 1-16 in place of the underlined blank. Then determine the mean and standard deviation of those 16 numbers and compare them with the earlier results and the expected values – be sure and calculate the <u>Exp. St. Dev of Means</u> for this series.

Finishing up: Write a description and analysis of results in terms of the relationships among the descriptive statistics for the population and sample means, including their means, standard deviations, and frequency distributions, and <u>comparing those distributions to the standard Uniform and Normal Distributions</u> and to the calculations on the worksheet, "The Uniform Distribution." Consider what happens when you vary the number of units in a sample. Also describe the results in terms of the Central Limit Theorem as discussed in your text.

Submit your responses to the various questions on paper, and in a Word file if you wish. The final Excel file, **CENTLIM.XLS**, is pretty large by the time you have completed the exercise. Therefore I would not recommend trying to e-mail it; just copy it to a disk and submit it with the paper work for this Topic's assignments.

Last Revision 3/1/2002; minor correction made 2/15/2011.