Exhibiting the Central Limit Theorem on a Spreadsheet

Alan F Stent and Lynn G McAlevey - Dunedin, New Zealand

1. Introduction

Recent references in the literature outline the benefits of spreadsheets as teaching tools (Hsiao, 1985; Lee and Soper, 1985 and 1986). The present paper outlines a spreadsheet simulation model, suitable for teaching purposes, to simulate the operation of the Central Limit Theorem. The model is different from the usual stochastic model insofar as it relies on self-referencing formulas to create the dynamic simulation. It does *not* depend upon macro facilities, though dynamic simulations can be performed along these lines. While macros do offer greater capabilities, they make the application more akin to an exercise in traditional computer programming. They require advanced skills and detract from the spreadsheet's advantages of simplicity and minimal programming knowledge needed for its operation.

2. Design and construction of the spreadsheet

A good spreadsheet design will enhance the impact of the simulation. It is important that the results be displayed in a single screen and that the pace of the calculations be neither too slow and bore the student, nor too fast to comprehend what is happening. While a judicious use of graphics and colour will also add to the effect, the instructor will be constrained by the computer and software available. It is not necessary to use the very latest equipment to produce a satisfactory exercise, and for the purposes here a simple design has been adopted to explain the basic principles. This is shown in Figure 1. The spreadsheet used is the popular Lotus 1-2-3, Release 2, and will run on early IBM or compatible personal computers as well as later models. Such computers are to be found in many business schools. It relies on neither a graphic screen nor colour, though can be adapted for these.

Figure 1 is the result of a simulation of 100 runs, each producing an independent observation on the mean of a sample of size 8 taken from a bimodal population.

Further details are given below. A frequency distribution is contained in column I, headed F(j), and this is also presented as a histogram on the right of the display. Classes are aligned along the vertical axis (column K) with frequencies proportional to marks (the digit "1" has been used) displayed horizontally. An approximate normal shape is readily apparent.

```
        A        B C D      E    F  G      H    I      J      KLMNOPQRSTUVWXYZAAAA
   ---------------------------------------------------------------------------------
1  Restart     0     j=   100           Distribution of the Mean
2  --------- Obs'n SAMPLE  Class L(j)  F(j)   Scaled
3  POPULATION     1     8      0.5    0    0     0.0
4  P(X<Xo)  Xo    2     1      1.5    0    1     1.0    1
5    0.000   1    3     5      2.5    0    5     5.0    11111
6    0.400   5    4     5      3.5    0    2     2.0    11
7    0.500   8    5    15      4.5    0    7     7.0    1111111
8    0.600  15    6     8      5.5    0   18    18.0    111111111111111111
9                 7    15      6.5    0   12    12.0    111111111111
10                8    15      7.5    0   16    16.0    1111111111111111
11                             8.5    1   20    20.0    11111111111111111111
12                             9.5    0    7     7.0    1111111
13                            10.5    0    7     7.0    1111111
14                            11.5    0    3     3.0    111
15                            12.5    0    2     2.0    11
16                            13.5    0    0     0.0
17                            14.5    0    0     0.0
18 Parameters    Sample       Statistics for the distribution ...
19 Mean=7.700  Mean =   9               Mean =7.570
20 Std D=6.325                          Std D =2.296
21 ===============================================================================
22 j-1 =                 99   --WORKING SECTION--
23 (j-1)/j =           0.99   Max =       20
24 (j-2)/(j-1) =      0.989   Scale =      1
25 XBar(j) =           7.57
26 var(j) =           5.269
27 Last Obsn =            9
...
```

NOTES:

1. Calculation order is columnwise.
2. During the exercise the top 20 rows should be in the display. The formulas in cells beneath the row 20 contain intermediate results that are out of sight.

FIGURE 1

Spreadsheet to simulate sampling distribution of the mean

There are four components underlying the spreadsheet construction. These are addressed in turn below.

(i) *Obtaining a sample from the population:* When the population is infinite, sampling is simulated by repeated independent draws from a probability distribution. This is the case demonstrated in Figure 1 which uses the following discrete random variable X.

x	P(x)	P(X < x)
1	0.4	0.0
5	0.1	0.4
8	0.1	0.5
15	0.4	0.6
	1.0	

The cumulative distribution, P(X < x), is entered in the range of cells A5 to B8; then samples are drawn using the spreadsheet's random number (@RAND) and table look-up (@VLOOKUP) functions, and entered in the eight cells E3 to E10 in Figure 1. Upon each recalculation of the spreadsheet, a fresh set of eight values will be computed for these cells representing a sample of size n = 8. Their mean is calculated at the same time using the spreadsheet average function, @AVG(E3..E10). This formula is entered in cell E19. For the recalculation captured in Figure 1, the sample is 8, 1, 5, 5, 15, 8, 15 and 15 (cells E3 to E10) and its mean is 9 (cell E19). It is straightforward to adapt the spreadsheet for different sample sizes or populations.

(ii) *Accumulating results over repeated samples:* The next step is to accumulate the successive observations on the mean into a frequency distribution. For this purpose a selection of 15 class intervals is made,

$$[0.5,1.5),[1.5,2.5),\ldots,[14.5,15.5)$$

and entered in cells G3 to G17. The number 15 has been chosen as the maximum number of classes that can be included in the layout while keeping the display within the confines of one screen. The total interval spanned is roughly six standard errors with the population mean (7.7) falling in the middle class.

The accumulation of frequencies is carried out in two stages in Figure 1, the calculations being contained in columns H and I respectively. (The calculations in these two columns could be combined into one, but have not been for ease of exposition.) In the first stage, a vector of cells is created in the range H3 to H17, headed "L(j)" in Figure 1. This vector contains the value 1 beside the class that the current observation on the sample mean belongs to, and zeros beside all other classes. (For purposes below, successive recalculations of the spreadsheet, or "runs" for brevity, are numbered j = 1,2,...)

In the second stage, the following recursive scheme is used to aggregate the L_j.

$$(1) \qquad F_j = \begin{cases} L_j & \text{for} \quad j = 1 \\ F_{j-1} + L_j & \qquad j \geq 2 \end{cases}$$

The resulting vector, F_j, is the frequency distribution. It is labelled "F(j)" in Figure 1, and contained in the cells I3 to I17.

The run number, j, is calculated in cell E1 in Figure 1. The user must assign the value 1 to the restart flag in cell B1 for the first calculation, and then reassign it the value 0 for the ensuing runs.

(iii) *Presenting the output:* The form in which output is presented is important to maintain a high visual impact. An overriding principle is simplicity of use and purpose so that the student is not distracted from the central teaching objectives. Three objectives underlie the present simulation. The first is to convey the concept of a sampling distribution. The second and third are to exhibit the approach to normality and decreasing variation, respectively, as the sample size is increased. These lessons can in fact be imparted by simply displaying the build-up of the frequency distribution in column I (Figure 1) for different populations and sample sizes. However, the spreadsheet's graphic capabilities can be employed as a further visual dimension by plotting the frequency distribution. Modern spreadsheets, when run with graphic display units, enable graphs and portions of the spreadsheet to appear on the screen simultaneously, linking the pictorial output with the collection of the sample. The net effect is most effective, and even more so when the components are accentuated by colour. Many of the older spreadsheets do not permit this spontaneity and some of the effect must be sacrificed by switching to and from the graphical display.

An alternative that has been adopted for Figure 1 does not use the spreadsheet's graphics, but rather constructs a display in the cells of the spreadsheet itself, with the advantage that the total output is contained within the one screen. The graph is built up in a 15 by 20 rectangular block of cells from K3 to AD17. The 20 columns in the block are each set to a display width of one character. Formulas are entered to display the character "1" in a cell if the frequency of the class on the same row attains a minimum count, and to leave the cell's display blank otherwise. The formulas entered in column J select their neighbouring class counts in column I, but will express them as relative frequencies should one or more classes have counts exceeding 20. This ensures that the graph always fits within the 20 columns utilised for the display.

(iv) *The mean and standard deviation of the sampling distribution:* As well as accumulating the frequency distribution of the sample mean for the exercise, the distribution's own mean and standard deviation can be calculated concurrently, and are brought up to the visible portion in cells J19 and J20. The exercise gains a further component. The student will see the distribution's mean approach the population mean, becoming very close to the latter as the number of samples collected increases. Its standard deviation is of course an estimate of the standard error, $\sigma/\sqrt{n}$. It will be seen to approach the latter value, giving a practical insight to the formula.

3. Using the spreadsheet

There are three steps in using the spreadsheet: (1) keying it in; (2) setting up the population and sample size; and (3) performing the simulation. The spreadsheet must be keyed in accurately and tested for correct operation. In most cases this will be best done by the instructor who may wish to adapt it for the circumstances of the exercise. A template would be prepared that can be retrieved by students, into the spreadsheet.

The second step requires setting up the population and sample size. The initial template would in fact include a probability distribution for the population in cells A5 to B17, as well as the first sample size. Thus the student can proceed directly to the third step to simulate with this combination. The instructor could prepare in advance a series of templates for a selection of populations and sample sizes.

Step three, performing the simulation, is initiated by setting the restart flag in cell B1 to 1, and doing a recalculation. This gives the first observation on the sample mean. It is followed by setting the restart flag to 0, and recalculating the spreadsheet repeatedly - each recalculation gives a further observation on the mean. The recalculation time depends upon the capabilities of the computer. Using an original IBM PC with an 8088 chip running at 4.77Mz and no mathematical coprocessor, one recalculation takes about 2.7 seconds. With the iteration count set to 50 the time is 1 minute and 15 seconds. These times are reduced by a factor of ten when using an 80386 chip running at 20Mz with a mathematical coprocessor.

4. Summary and conclusions

The techniques described enable Monte Carlo simulations to be performed on a spreadsheet. They utilise self-referencing formulas to accumulate statistics and present these progressively in an informative manner. They do not rely on macros. The techniques described are general and are applicable in areas beyond education. For instance, they can be used to build simple queueing and inventory simulation models on spreadsheets, as well as adding a stochastic component to budgeting models as a means of treating uncertainty.

References

Hsiao, F S T (1985) *Int.J.Math.Educ.Sci.Technol.* **16**, 705.

Lee, M P and Soper, J B (1985) Spreadsheets in teaching statistics. *The Statistician* **34**, 317-321.

Lee, M P and Soper, J B (1986) Using spreadsheets to teach statistics in psychology. *Bulletin of The British Psychological Society* **39**, 365-367.

Lotus Development Corporation (1985) *123 Reference Manual.*