# The Computer Spreadsheet : A Versatile Tool for the Teaching of Basic Statistical Concepts

Teck-Wong Soon - Singapore

## 1. Computer software for the teaching of statistics

Many students find statistics difficult and unattractive. There is therefore an urgent need for teachers to consider new approaches and tools for teaching the subject. Many have advocated the use of microcomputers as teaching tools.

When using microcomputers for teaching purposes, it is necessary to consider which software could be regarded as good teaching tools. A good teaching tool should, in my opinion, satisfy at least the following three criteria:

(i) It should be reasonably priced, and readily available.
(ii) It should be flexible thereby allowing teachers to introduce modifications in response to the special needs of their students.
(iii) It should be able to demonstrate the basic concepts of the subject.

While specialised statistical packages are now available for sophisticated statistical analyses, they may not be generally suitable for teaching purposes. Indeed, a common observation made by many statisticians is that such packages have increased the ease of statistical computation without raising the level of statistical awareness. This has given rise to numerous occasions whereby highly sophisticated analyses are undertaken without the ability of proper or meaningful interpretation. Similarly, software dedicated to computer-based learning of statistics usually has the disadvantages of being expensive and designed with highly specific objectives thereby making it difficult for teachers to adapt to the special needs of their students. Thus, these two categories of software fail to meet at least one of the three criteria stipulated above.

Computer spreadsheet programs such as Lotus 1-2-3, Microsoft Excel, Borland's Quattro, etc. have in recent years become popular and almost universal. They are readily available, reasonably priced and designed to be user-friendly. Moreover, they all incorporate the ability to use templates and macros, making it extremely easy for applications to be developed and modified. Thus, they easily satisfy the first two criteria of good teaching tools for statistics. It remains for us to show that although spreadsheets are usually thought of as a business tool, they satisfy also the third criterion, i.e. that they are able to demonstrate the basic concepts of the subject. Their ability to do so is enhanced by built-in function(s) for generating random numbers.

## 2. Examples

### 2.1 *A statistical sampling experiment*

Consider the simple coin tossing example which is discussed in virtually all basic statistical textbooks, for example, Hays (1988). The textbooks usually suggest that the teacher ask each student to toss a (fair) coin n times, and to record the number of heads obtained, say x. Each series of n tosses is considered a single trial. Each student performs a certain number of trials, and at the end of the exercise, the frequency distribution of the number or, equivalently, the proportion of heads, is plotted. This series of exercises, if undertaken as described, is tedious and time-consuming. Not surprisingly, it is rarely done. Thus, in general, students are not provided the opportunity to experience and appreciate the basic statistical concepts intended to be conveyed through the exercise. They have therefore to try to appreciate these concepts through intellectual efforts alone, i.e. on an abstract level. However, with the help of a computer spreadsheet, the exercise could be simulated easily and rapidly. While this is demonstrated below using Lotus 1-2-3, the procedures for carrying out such simulations on other spreadsheet programs are similar.

In setting up a worksheet for the simulation exercise, we begin by setting the re-calculation mode to 'manual'. Under this mode, the worksheet is re-calculated only when it is instructed to do so. In Lotus 1-2-3, we do so by pressing the [CALC] (F9) key. We then enter in cell A1 the formula "@rand", which returns a random number between 0 and 1. The value of the random number returned by @rand in cell A1 is then evaluated within the context of the statistical sampling experiment being simulated. Thus, if we were simulating the tossing of a fair coin, we could assign the values 1 (representing 'head') or 0 (representing 'tail') to an indicator variable depending on the value of the random number returned. Hence, we assign to cell B1 the formula "@if(A1>0.5,1,0)" which returns a value of 1 if the random number at A1 is greater than 0.5, and a value of 0 if the value of A1 is 0.5 or less. The value returned in cell B1 is therefore the outcome of a single toss of a fair coin. To simulate n tosses of a fair coin, all we have to do is to copy the formula in row 1 to the next (n-1) rows. The outcomes of n tosses of a fair coin are then displayed in the cells of column B. The total number of heads resulting from the n tosses can be easily obtained by counting the number 1's in column B, or similarly by adding all the values in column B. Further, it is relatively easy to set up column C to display

$$\text{Relative frequency} = \frac{\text{running total of column B}}{\text{total number of tosses}}.$$

Table 1 is the worksheet segment showing the outcomes of 10 tosses of a fair coin.

TABLE 1
Simulation of 10 tosses of a fair coin

| Toss | A | B | C |
|---|---|---|---|
| 1 | 0.511495 | 1 | 1.00 |
| 2 | 0.325439 | 0 | 0.50 |
| 3 | 0.085768 | 0 | 0.33 |
| 4 | 0.547836 | 1 | 0.50 |
| 5 | 0.124359 | 0 | 0.43 |
| 6 | 0.237757 | 0 | 0.33 |
| 7 | 0.675683 | 1 | 0.43 |
| 8 | 0.841214 | 1 | 0.50 |
| 9 | 0.883926 | 1 | 0.56 |
| 10 | 0.722461 | 1 | 0.60 |
| Total | | 6 | |

Note: The value 1 in column B represents 'head' while the value 0 represents 'tail'.

## 2.2 *Sampling distributions*

The simple sampling experiment described above represents a single trial. Its outcome may be regarded as a single random sample. Much of statistical inference is based on repeated sampling. With a spreadsheet, the sampling experiment can be repeated simply by instructing the worksheet to re-calculate itself. This is because each time the [CALC] key is pressed, new values are assumed by each cell assigned the "@rand" formula. The number of heads, (or equivalently, the proportion of heads) of each experiment (or simulation) is recorded. This need not be done manually, as a simple macro can be written to copy the required value in a different segment of the worksheet. The results of all the simulations can then be examined, summarised in a frequency distribution, and graphed; all these operations can be undertaken by the spreadsheet program, which is extremely versatile. Thus, a very large number of simulations of n tosses of the coin (for various values of n) can be performed easily and quickly.

The ability of the spreadsheet to easily and quickly simulate sampling experiments allows us to demonstrate the various properties of a sampling distribution. In the present example, we can see that the proportion of heads varies from sample to sample. Furthermore, we can also see that the sample proportion is not generally equal to the population (or theoretical) proportion, which for a fair coin is 0.5. The difference between the sample proportion and the population proportion is sometimes referred to as

the sampling error. Our interest, insofar as statistical inference is concerned, is the *sampling distribution* of the sample proportion which reflects its variability. This variability is measured by the sample variance, which is simply the mean of the square of the sampling error. Table 2 shows the sampling distribution of the sample proportion of heads arising from 500 simulations of 10 tosses of a fair coin.

TABLE 2
Sampling distribution of 500 simulations (n = 10 and p = 0.5)

| Sample Proportion (1) | Relative Frequency (2) | (2) × (1) | Square of Sampling Error (3) | (3) × (1) |
|---|---|---|---|---|
| 0.0 | 0.000 | 0.0000 | 0.25 | 0.00000 |
| 0.1 | 0.006 | 0.0006 | 0.16 | 0.00096 |
| 0.2 | 0.034 | 0.0068 | 0.09 | 0.00306 |
| 0.3 | 0.146 | 0.0438 | 0.04 | 0.00584 |
| 0.4 | 0.200 | 0.0800 | 0.01 | 0.00200 |
| 0.5 | 0.232 | 0.1160 | 0.00 | 0.00000 |
| 0.6 | 0.198 | 0.1188 | 0.01 | 0.00198 |
| 0.7 | 0.126 | 0.0882 | 0.04 | 0.00504 |
| 0.8 | 0.050 | 0.0400 | 0.09 | 0.00450 |
| 0.9 | 0.006 | 0.0054 | 0.16 | 0.00096 |
| 1.0 | 0.002 | 0.0020 | 0.25 | 0.00050 |
| | 1.000 | 0.5016 | | 0.02484 |

The last row of Table 2 also shows the mean and variance of the sampling distribution. Since the population (theoretical) proportion is 0.5, Table 2 confirms that the mean of the sampling distribution of the sample proportion is the population proportion, i.e. the sample proportion is an *unbiased* estimator of the sample mean. Also, it can be readily seen that the variance of the sample proportion which is known to be 0.25/n, i.e. 0.025 when n = 10, is simply the variance of the sampling distribution. Thus, basic concepts concerning the sampling distribution can be illustrated in a concrete manner through the use of the versatile spreadsheet.

### 2.3 *Central Limit Theorem*

Besides being able to rapidly simulate a large number of sampling experiments, spreadsheets are also able to produce theoretical frequency distributions and graph them. For example, the density function of the normal distribution can be tabulated and graphed. This allows us to demonstrate easily the Central Limit Theorem, or the Weak Law of Large Numbers, which states simply that the sample mean is distributed approximately as a normal distribution. For the coin tossing experiment, the sample proportion is simply the mean number of heads, i.e. it is a sample mean. Thus, to demonstrate the Central Limit Theorem, we need only repeat the sampling experiments

for different values of n, and superimpose the resulting frequency distributions of the sample proportions on a normal distribution with the same mean and variance as the sampling distribution. Indeed, as shown in Figure 1, the normal distribution is a good approximation even for a sample size as small as 10, which agrees with the conventional wisdom stated in textbooks that when $np \geq 5$, the binomial distribution may be approximated by the normal distribution.

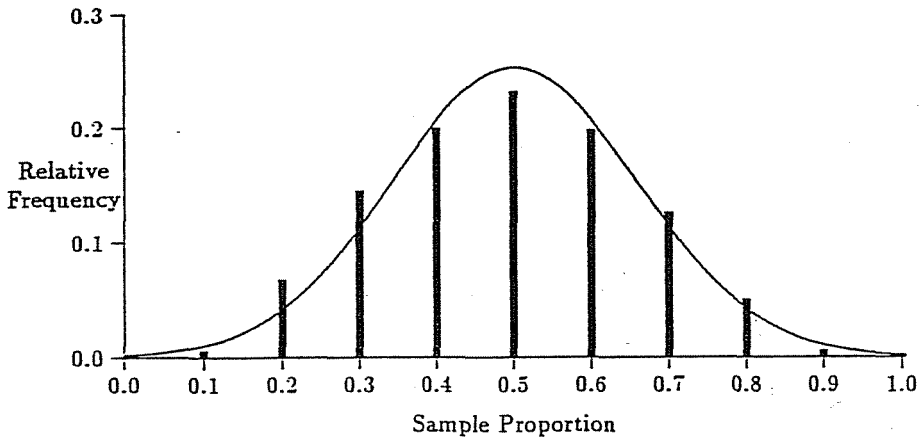


FIGURE 1
Distribution of sample proportion based on 500 simulations (n = 10)

## 2.4 *Data analysis*

While we have so far concentrated on the ability of the spreadsheet to demonstrate basic concepts, we should not forget its versatility in data entry, manipulation and analysis. Moreover, the spreadsheet is a convenient form for storing example datasets for analysis by students. Each row of the spreadsheet can be regarded as representing an observation while each column represents a variable. Most of the basic statistical procedures presented in the usual statistical textbooks can be undertaken on the spreadsheet. Indeed, with the graphical capability of most spreadsheets, exploratory data analysis can be undertaken easily and quickly. For example, in teaching correlation and regression, students are often told that "a correlation coefficient of zero implies only that there is no *linear* relationship" and that "a non-linear relationship" can exist. The classical example is given by the dataset:

| X | 0 | 1 | -1 | 2 | -2 | 3 | -3 | 4 | 4 |
|---|---|---|---|---|---|---|---|---|---|
| Y | -4 | -3 | -3 | 0 | 0 | 5 | 5 | 5 | 12 |

Students, when presented with this dataset (or one similar to it), usually compute immediately the correlation coefficient of X with Y and finding it to be zero, conclude that there is no relationship between X and Y. In fact, on closer examination, it can be

seen that far from being totally unrelated, X and Y have a perfect quadratic relationship, viz $Y = X^2$. This relationship would be revealed clearly if a scatter diagram were plotted. The ease with which the spreadsheet can sketch scatter diagrams makes it much more likely that such diagrams are indeed sketched and examined. When students have to draw the scatter diagrams by hand, the usual tendency is *not* to do it because of the tediousness involved in plotting the points on a piece of graph paper.

## 3. Conclusion

The ease in which spreadsheets can be used to convey basic statistical concepts has been demonstrated. Thus, spreadsheets satisfy all the three criteria stipulated for a good teaching tool. However, there are at least four other reasons (or advantages) for using spreadsheets as a teaching tool.

First, almost all statistical procedures taught at the undergraduate level could be performed with spreadsheets. The integrated graphic capability of spreadsheets allows the use of exploratory statistical techniques. While spreadsheets may be less efficient than specialised statistical software, using spreadsheets does have the advantage of forcing students to specify the steps needed for statistical analysis. Specialised statistical software has sometimes been characterised as a black box whereby the results are generated from the data without users having to be aware of the methodology applied in processing the data.

Second, data entry and manipulation with spreadsheets are extremely easy to do. Moreover, if necessary, data maintained by spreadsheets can be easily exported for analysis with sophisticated statistical software. Thus, even where sophisticated statistical analysis is required, it may be worthwhile to use spreadsheets as intermediaries for data entry and preliminary exploration of data.

Third, given the widespread applications of spreadsheets in many areas, notably in business, the skills acquired by students through learning statistics with spreadsheets will prove invaluable. Conversely, many students may already be familiar with spreadsheets before learning statistics, which would suggest that spreadsheets might be a useful media for introducing probabilistic and statistical concepts to them.

Finally, there are immense possibilities in using spreadsheets for sophisticated data analysis. Kaciak and Koczkodaj (1989) demonstrate an application of spreadsheets in such diverse tasks as multidimensional data analysis using the principal components method. While spreadsheets may be less efficient (or slower) than specialised statistical software, they do have the advantage of transparently exposing their memory contents and revealing all the steps of the analysis.

## References

Harvey, Greg (1987) *Lotus 1-2-3 Desktop Companion.* Sybex.

Hays, William L (1988) *Statistics* (4th ed). Holt, Rinehart & Winston Inc.

Kaciak, Eugene and Koczkodaj, W W (1989) A spreadsheet approach to principal components analysis. *Journal of Microcomputer Applications* **12**, 281-291.