Teaching Elementary Inference with Computer-based Simulations

W Douglas Stirling - Palmerston North, New Zealand

1. Introduction : why do simulations?

The topics in most introductory statistics courses can be split into two categories: exploratory and confirmatory data analysis. Teachers usually introduce each data structure with graphical and tabular methods for examining and presenting a batch of such data (Exploratory Data Analysis). These EDA topics are followed by the idea that batches of values can be considered to be random samples from some population; the remaining treatment is based on the properties of such random samples (Confirmatory Data Analysis). The importance of EDA is gradually being recognised and emphasised, though it often is poorly integrated with the CDA component of courses. However, in this paper we shall only be dealing with the teaching of CDA in courses.

The central theme of CDA is that a set of data being analysed is a random sample from some population whose exact characteristics are unknown. The basic concept that must be understood is that repeating the sampling would give a different sample similar to, but not exactly the same as, the observed data.

Most of CDA is based on the sample-to-sample variability of quantities that are calculated from such samples. In particular, we teach about the distributions of sample means, standard deviations, proportions, correlation coefficients, least squares lines, and other sample statistics. When inference is introduced, the concept of a confidence interval and its confidence level are explained in terms of the sample-to-sample variability of the intervals and the proportion of them that would include the true parameter value. Significance levels (p-values) are described as the proportion of repeated samples that would give rise to a more "extreme" value of a test statistic when the null hypothesis is true.

Students often find such properties hard to understand, largely because they concern the sample-to-sample variability of the relevant summaries - students must visualise a random sample of random samples in order to appreciate that the sample mean, for example, itself has a variance.

All such properties can be explained mathematically, but this is a poor way to teach such essential concepts, especially to mathematically weak students. Even students with stronger mathematical backgrounds often have difficulty relating the theory to real sets of data when taught in this way.

Since our explanations of the main CDA techniques are based on the variability of results from sample to sample, all such concepts can be illustrated by simulating this sample-to-sample variability with samples generated from a suitable population. The resulting empirical distribution gives a concrete illustration of the underlying theory and helps explain the more abstract concepts.

2. How should simulations be performed?

Until recently, any simulations in introductory courses were performed with hand calculations. For example, students might be asked to roll a single die several times and display the distribution of the value on top; this could be repeated with the means of pairs of dice and the means of four dice at a time, in order to illustrate the Central Limit Theorem. Unfortunately, such simulations have several major drawbacks. They take a long time and are extremely tedious, even for the simplest types of simulations. They are prone to numerical errors when used for topics like least squares lines or hypothesis tests; such errors are likely to ruin the effect of the simulation. For some types of simulation (such as showing the properties of 95% confidence intervals), large numbers of samples must be analysed for a reliable illustration. Even if the simulation is done during a class, with each student contributing one or two values to an empirical distribution, numerical mistakes and the logistics of collecting and presenting the empirical distribution limit its use to once or twice in a course.

To perform regular simulations during a course, computers must be used for the calculations and for the display of the resulting empirical distributions. However, simulations using programs such as MINITAB essentially require the student to write a program - difficult for weak students. If the student is provided with a macro which generates the relevant empirical distribution, then the mechanism of the simulation is lost. Most statistical programs are designed for analysis of data, not for teaching; they do not clearly distinguish between the three basic components of a simulation - populations, samples, and empirical distributions - and do not clearly show the relationships between these. Also, some types of empirical distribution, such as those of confidence intervals and least squares lines, cannot be displayed pictorially in most general-purpose statistical programs.

Effective software for teaching with simulations must have various features that have not been present in statistical programs until recently.

(i) Populations, samples, and empirical distributions must be clearly distinguished.
(ii) The basic display of all such entities should be graphical.
(iii) Taking a sample from a population should automatically add the appropriate sample summaries to whatever empirical distributions are being accumulated. The user should get the visual feedback of seeing the sample and seeing the summary being added to the empirical distribution.
(iv) The user interface must be simple and menu-driven. Graphical user-interfaces are

the easiest to use in this respect. Students in introductory courses should not be required to master a complex command language to run the program.

(v) The design of the program should be as orthogonal as possible. For example, the empirical distribution of sample means from a normal distribution should be created in a similar way to that of sample correlation coefficients from a bivariate normal.

It is only recently that such software has become available. For example, StatLab (Stirling, 1987a,b), has a display with two windows - one for the active population and sample, and the other for the empirical distribution which is currently being built up. Each window has various ways to display its contents graphically and these are selected by mouse clicks. An on-screen button takes a sample from the current population, displays it and adds a summary statistic to the empirical distribution's display. The program also has various ways to examine interactively the various displays, such as reading off probabilities. A sample screen is shown in Figure 1.

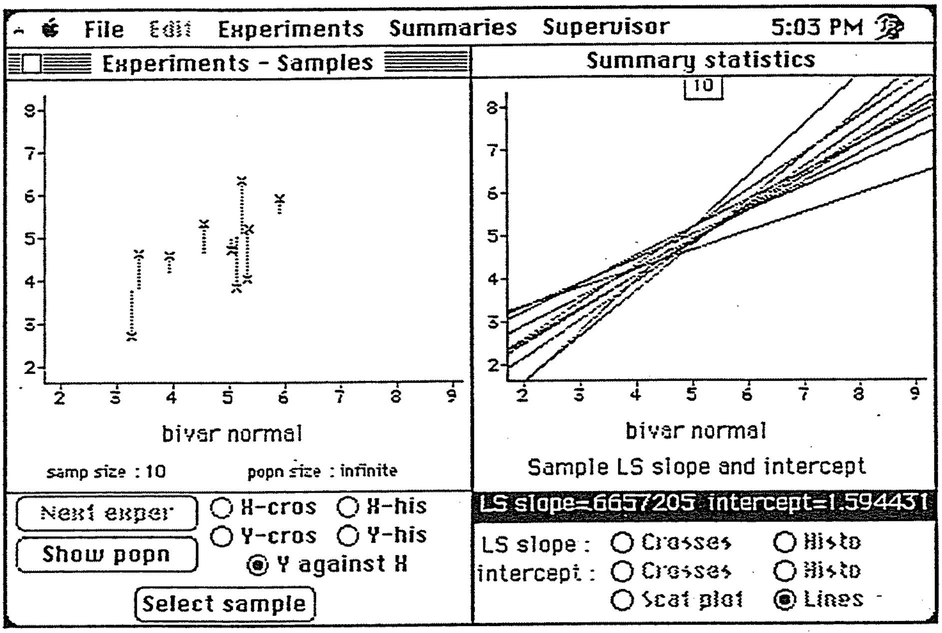

FIGURE 1

StatLab screen showing empirical distribution of least squares lines based on samples from a bivariate normal population with correlation 0.5; the last of the 10 samples is shown on the left with dotted lines indicating the residuals.

3. What can we teach with simulations?

The most basic CDA concept is the variability of random samples and their graphical displays. A good feel for the range of "shapes" of such displays from standard models is *essential* before students can be taught to recognise data sets for which the standard models are inappropriate. For example, 12 random scatter plots were generated from a bivariate normal distribution with correlation coefficient 0.5; Figure 2 shows the four most unusual of these.

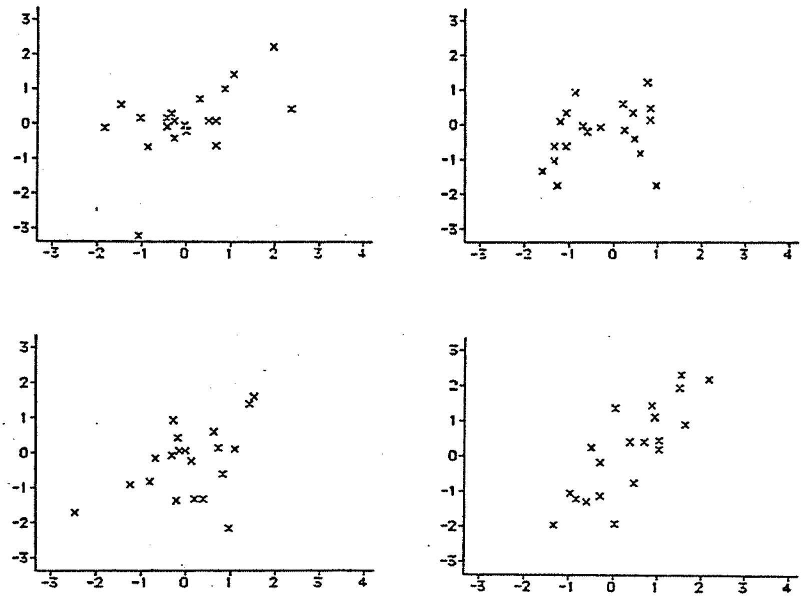

FIGURE 2
The four most unusual scatter plots chosen from 12 random samples from a bivariate normal distribution with correlation 0.5

Similarly, students need to examine histograms of samples from a normal distribution before they can assess skewness, outliers, the existence of two or more modes, etc. in real data sets. They need to examine probability plots of normal samples before they can visually assess what should be interpreted as non-normality.

Few statistics courses give students this feel for likely patterns in random samples and in fact students are often *taught* to interpret patterns such as those in Figure 2 as positive proof of incorrect assumptions, such as non-linearity, outliers, non-constant variance or clusters.

Much of the statistical theory in introductory courses is based on the properties of summary statistics calculated from random samples. Once the variability of random samples has been demonstrated, the randomness of quantities calculated from these samples follows immediately. However, the empirical distributions of the summaries are necessary to demonstrate their properties.

The most basic summary statistic is the sample mean and its behaviour is described by the Central Limit Theorem. The essential properties of the mean are its approach towards normality as the sample size increases and the halving of spread as sample size quadruples. These are easily illustrated with simulations. For example,

Figure 3(a) shows the empirical distributions of means of samples of size 1, 4 and 16 from an exponential distribution, based on 50 samples of each size.

Similar demonstrations can be used to examine the properties of other summary statistics, such as sample medians, standard deviations, proportions, slopes of least squares lines, etc. It is important that students appreciate the factors that influence the distributions of such quantities. For example, Figure 3(b) shows the empirical distributions of correlation coefficients for samples of size 20 and 80 from bivariate normal populations with ρ = 0.5 and ρ = 0.9. The differences between the box-plots in Figure 3(b) should be understood by anyone using correlation coefficients, but are usually not explicitly taught.

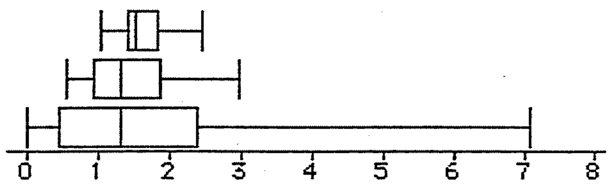

FIGURE 3(a)
Box-plots showing the empirical distributions of sample means from 50 samples of sizes 1, 4 and 16 from an exponential distribution

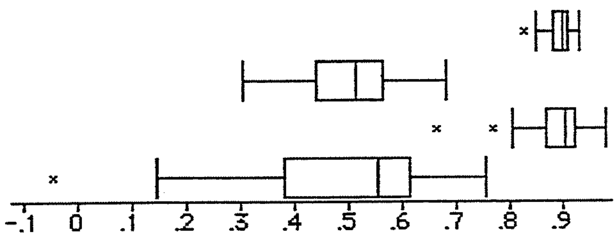

FIGURE 3(b)
Box-plots showing the empirical distributions of sample correlation coefficients from 50 samples from a bivariate normal distribution, with sample sizes 20 and 80, and correlation coefficients 0.5 and 0.9

The sampling distribution of least squares lines similarly helps explain their properties. For example, Figure 1 clearly demonstrates the greater variability of predictions away from the explanatory variable's mean.

The properties of confidence intervals are usually expressed in terms of repeated sampling; the intervals enclose the true parameter value in 90% of samples if the confidence level is set at 90%. This can be demonstrated from the empirical distribution of 90% confidence intervals. For example, in discussing confidence intervals for the mean of a normal population, students can verify that approximately 90% include the true mean. To emphasise their interpretation, similar exercises should be repeated for the various applications of confidence intervals to estimate different parameters in the course.

Of all the topics in an introductory statistics course, hypothesis testing is usually the hardest to understand and many students never grasp the interpretation of significance levels (p-values). Hypothesis testing is based on the following properties of p-values:

(i) When the null hypothesis is true, the p-value is uniformly distributed between 0 and 1.

(ii) When the null hypothesis is false, the p-value is more often close to 0 than close to 1. The farther the null hypothesis is from being true, the closer the p-values will be on average to 0.

Figures 4(a) to 4(c) demonstrate these properties with p-values for testing whether the mean of a normal population is 10, 11 and 12 when the true mean is 10. From these, students can appreciate that low values would be more likely when the null hypothesis is false and, even though they are still possible when the null hypothesis is true (with a probability we can calculate from the uniform distribution of the p-values), they provide evidence against the null hypothesis.

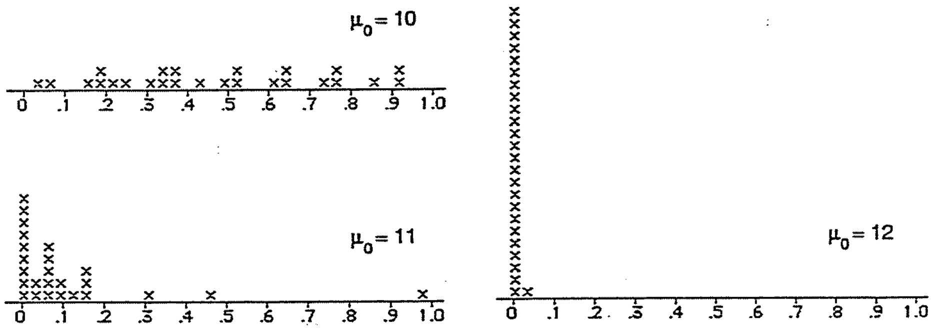

FIGURE 4

Empirical distributions of p-values for testing whether the population mean is 10, 11 and 12, based on 25 samples from a normal population with mean 10

Students tend to miss these central features of hypothesis testing when hypothesis testing is based on distributions of test statistics and rules for accepting and rejecting of hypotheses. The interpretation of p-values given above is the same for all types of hypothesis test, whether of proportions, least squares slopes, or the equality of means in two or more populations, and can be demonstrated with simulations in a similar way for all hypothesis tests in a course.

Simulations are also useful to demonstrate various other results in introductory courses. A few are listed below:

(i) independence of the sample mean and variance from normal populations;

(ii) the effect of sample size on the distribution of means from Cauchy distributions;

(iii) comparisons of the distributions of the sample mean and median from various distributions;

(iv) comparisons of the powers of parametric and nonparametric tests (via the distributions of p-values when the null hypothesis is false);

(v) comparisons of the properties of alternative estimators, such as the method of moments and maximum likelihood estimators for θ in samples from a uniform $(0,\theta)$ population.

4. Conclusions

Simulations can be used to illustrate most CDA topics in introductory statistics courses and, with currently available software, they can be widely used as a teaching tool. Ideally, students will perform the simulations themselves, though demonstrations in a lecture are also possible.

However, simulations do not need to be used only to support conventional teaching of statistics. Especially for students in service courses with weak mathematical backgrounds, simulations can play a more central role; they can replace mathematical explanations of the major statistical concepts. For example, hypothesis tests can be taught by means of the distributions of p-values as described above, without teaching formulae for test statistics or looking up tables. Computers would be used to apply the tests to any real data set so we only need to teach how to interpret the resulting p-values.

Computers are likely to fill an increasingly important role in statistical teaching, not only for data analysis, but also for the presentation and/or illustration of theoretical concepts.

References

Stirling, W D (1987a) The design of a microcomputer program for teaching statistics. *The New Zealand Statistician* **22**, 46-54.

Stirling, W D (1987b) *StatLab : Microcomputer-based Practical Classes and Demonstrations for Teaching Statistical Concepts.* The New Zealand Statistical Association, Wellington, New Zealand.