SIMULATION MODELS, GRAPHICAL OUTPUTS, AND STATISTICAL DISCOVERY

K. Laurence Weldon, Department of Mathematics and Statistics,
Simon Fraser University, Canada

Both graphics and simulation are tools of increasing use in statistical education and practice. However the parametric traditions of statistics still resist the legitimacy of these technologies. Yet the historical roots of the discipline feature data analysis and probability modeling as principal tools, and graphics and simulation can be seen as the transformation of these roots caused by the computer revolution. With this perspective, much of the theory of statistical inference can be seen as a temporary diversion. This view has implications for both the style and content of statistics courses. The proposal here is that we should be using much more graphics and simulation in our courses, and more focus on data analysis and probability for content. Graphics provide the links between the two areas. It is argued that graphics and simulation allow a broader-based understanding of statistics, something that is both attractive and useful for both students and practitioners.

INTRODUCTION

The roots of statistics go back to gambling and genetics (probability modeling), and official statistics associated with politics, taxation and vital statistics (descriptive statistics). Applied research in the fields of agriculture and silviculture required designed experiments which led to the development of the theory of inference. The subject has broadened in many directions in recent decades, to include exploratory data analysis, study design, decision theory, Bayesian inference, and resampling techniques. The current diversity of approaches invites an educator to reconsider the scope of the discipline since the historic sequence of topics is not necessarily appropriate for pedagogy. In fact, a re-examination of statistics curricula is helpful in investigating the foundations of the subject.

Ironically, there is a sense in which a “back-to-basics” movement in teaching statistics would be an innovation of some merit. More particularly, we argue for a return to a focus on probability modeling and descriptive statistics, the roots of the discipline, for introductory courses. The result would be an innovation because of the impact of computer software on statistical practice and theory. The increase in importance of algorithms, graphics, multivariate computations, simulation and resampling methods are fairly clear consequences. More subtle is the decreasing importance of some traditions: parametric models for inference, optimality of inference methods, the least squares

criterion, the common location-scale summaries, and the histogram. Another subtle influence is the increase of trial-and-error methods (or iterative methods) of modeling.

Descriptive statistics is the engine behind most modern data analyses: the number one rule for the applied statistician is “look at the data”. Pre-computer constraints limited the scope of data-analytic work with the result that the mathematics of very specialized inference contexts was the focus of the statistician’s attention for several decades. The restrictive traditions of this era have been painfully slow to relax and evolve. In an effort to fill modern needs, some educators in statistics have rejected mathematical statistics and probability modeling in favour of computerized data smoothing and other “informal” methods. While statisticians are aware of the hazards of informal inference, it appears these hazards are no more serious than with the formal pre-computer methods undertaken while the data could not be easily visualized. This view has been convincingly made by Cleveland (1993). His argument for a new paradigm of statistical analysis based on graphics is compelling: the advantages of a graphical approach for inference over classical methods is particularly evident with multivariate data, and in practice, multivariate data is the norm. Descriptive statistical methods have become much more than a first step.

Probability modeling is another root of statistical theory that has taken on a new importance in the computer age. Computer simulation has increased the practical utility of probability modeling. Unfortunately, teachers concentrating on data analysis and graphical statistical methods may have reduced the probability content of their courses. However, in the broad view of statistics, observation of a consequence of randomness may be considered data of a kind, and the combination of applied probability modeling and simulation is an effective way of studying such phenomena. The traditional probability models can be combined to mimic complex systems, and simulation can reveal the properties of these complex systems even when data is absent or incomplete. To understand the potential for this approach, students need to be exposed to it in elementary contexts. For some details of an introductory course in statistics based on probability modeling, see Weldon (1998).

A modern training in statistics should have a heavy dose of the use of *graphical methods*. Graphical methods are important for both data analysis and the study of applied probability models. Until statistical software made graphical methods feasible, both these approaches to the discipline of statistics were held back. In the case of applied probability

models, computers not only allowed the simulations themselves, but also the portrayal of findings over the parameter space. In this paper my theme is that graphical methods play an essential role in both statistical education and practice, since they are the key tool in both data analysis and applied probability.

THE ROLE OF GRAPHICAL METHODS IN LEARNING DATA ANALYSIS

In order for statistical education to be useful, the student must learn the intellectual processes needed for statistical practice. This process is not simply an exposure to facts, but exposure to the questions and answers associated with the analysis of data. For example, questions like

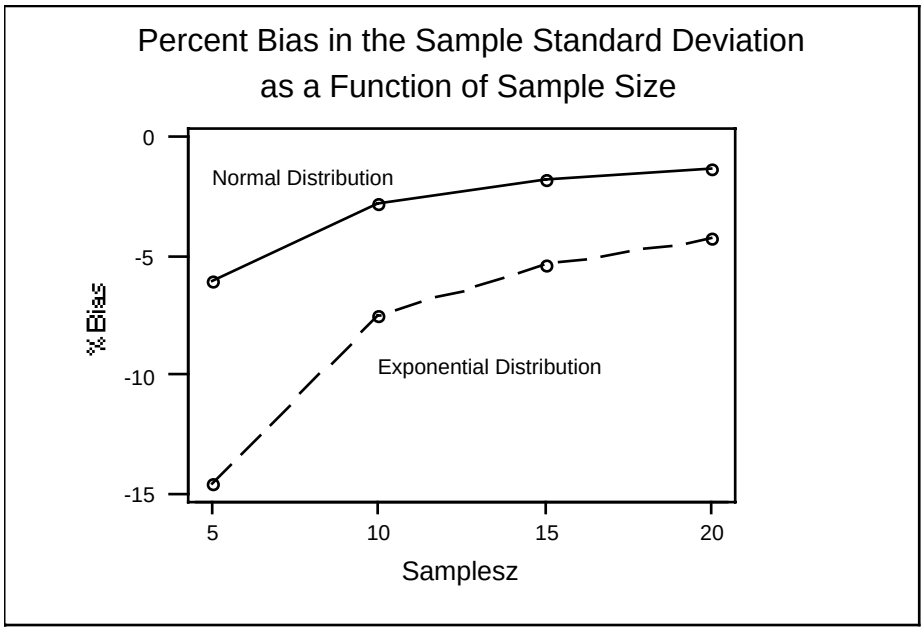
- Is this an exploratory study or a confirmatory study?
- Is there any reason that this data might not be typical of the population of interest?
- Are there any unusual features to this data, and if so, how will they affect my findings?
- How shall my findings be described for the primary audience of this analysis?
- need not be memorized by the student who has been led through some guided experiences.

In discussions of data analyses, the language of choice is ‘graphics’. The mere mention of correlation, regression, location and scale, smoothing, sampling or normal distribution will have the instructor and the student drawing graphs at each other. So a course in data analysis will require a few icons that students are very familiar with: the density curve, the dot plot, scatter diagram, the contour ellipse for bivariate data, the regression line and conditional densities, and the population-sample schematic. These icons can be combined and duplicated in creative ways that suit the particular needs of a given data analysis.

Clearly, graphics is the most important tool for the process of data analysis. This claim is increasingly accepted by statistics instructors. In the next section I explore the use of graphics to enhance the study of probability models, even when no data is available - this technique is less popular than the use of graphics for data analysis, but may have been underrated so far.

THE ROLE OF GRAPHICAL METHODS IN SIMULATION

Simulation in statistics is usually used to estimate probabilities or expectations in situations where the analysis is mathematically intractable. However, for all but the most advanced students, most results are unknown and may be considered “intractable” from their point of view. For example, the bias in the usual sample standard deviation may be simply assessed for a particular distribution such as the standard normal. The result for n = 5,10,15,20 is -6.0 %, -2.8%, -1.8%, -1.3%. Similarly for the exponential distribution the corresponding result is -14.6%, -7.5%, -5.4% and -4.3%. This simulation experiment can be summarized graphically as:

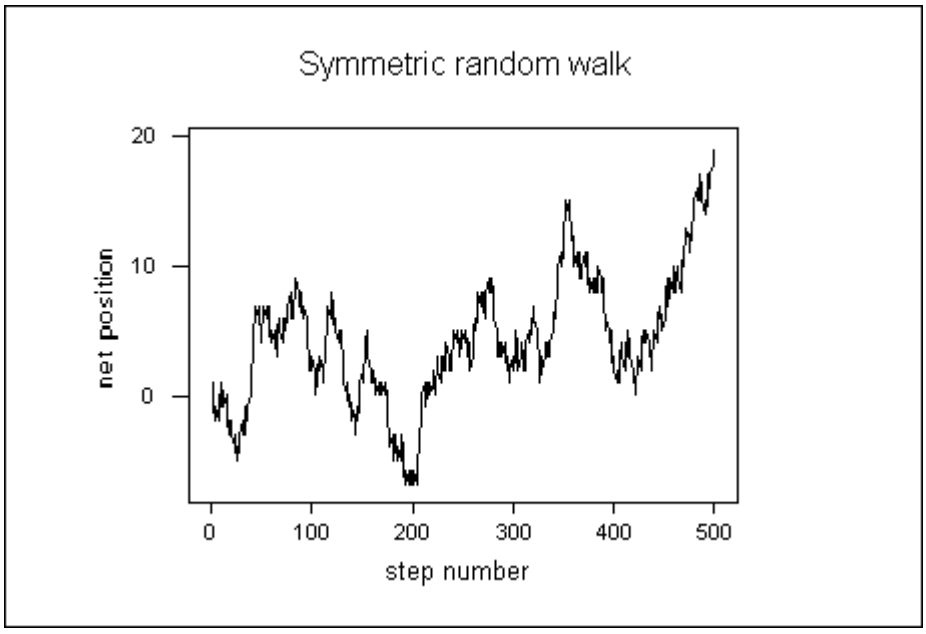


Most students would find the table of data less informative, even though the same information is there. It is doubtful that a formula could do a better job, even if the student knew a formula for this bias. The graph gives a summary of the bias in the sample SD for a wide range of situations, in an efficient and memorable way. The fact that it was produced with minimal knowledge of mathematical or statistical theory is also noteworthy.

Another use of graphics with simulation is to familiarize students with the consequences of random sampling. For example, many students think that a sample of size 30 from a normal distribution will look like a normal distribution. But a few dotplots will show that this is not so; moreover, the grouped-data histogram will not improve things very much for these moderate-sized samples.

Many basic statistical tools make an assumption of normality. How is one to judge whether a given sample of data is from a normal population? Of course the real question is, does it matter much that I will assume normality in this instance? In any case, if the student learns the futility of trying to recognize normality based on a small sample, this will be a valuable lesson. Graphical presentation of the simulated output is the key here.

This example illustrates the use of simulation and graphics to teach about the consequences of randomness in the simplest setting, that of random sampling from a primitive population. However the need for graphics is even more vital in summarizing the outcome of more complex systems. Consider, for example, the symmetric random walk. Actually, for this model it is rather easy to produce mathematical results, but the graphical representation makes clear visually things that are very sophisticated psychologically. Most students make the common error of confusing an expected value for the final displacement, 0, with the usual displacement which will be anywhere within $2n^{1/2}$ of 0. Consider for example the following typical symmetric random walk portrayal for 500 steps.



Not only is the increasing variation from 0 evident from the picture, but the apparently systematic trends in either direction are clearly illusions. With this simple example in mind, students will learn to be skeptical of claims made on the mere basis of past observations for trends in the stock market, world climate, earthquakes, etc. It is

difficult to make this point convincingly without both simulation and the graphical output of the simulation.

PARAMETRIC TRADITIONS AND GRAPHICAL COMMUNICATION

Statistical tradition is closely tied to parametric inference. The logical simplicity of a focus on a few parameters for data summary has been compelling. But some of this compulsion comes from an inability to produce graphical summaries efficiently. A simple linear regression is often used to describe the predictive relationship between two variables, such as the prediction of weight from height used to assess “ideal” weight. But there is no belief in the fiction of a linear relationship here - it is merely a simple way to approximate the relationship with a two-parameter summary, one that is easy to communicate. However, a good graphical summary can be based on an empirical nonparametric smooth of the data, with any degree of smoothness desired, and without the constraint of a simple parametric representation. Parametric summaries will still be needed with explanatory models, but empirical models are more common in statistical practice. With graphical relationships easy to generate and communicate electronically, there is less need for parametric summaries.

CREATIVITY IN DATA ANALYSIS AND PROBABILITY MODELING

Whether one is using graphics to analyze data or to study probability models, the creative stimulus of the visualizations cannot be denied. If traditional statistical inference has concentrated on reining in our creative urges so we are not overly enthusiastic about possibly transient effects, modern statistics is letting go of the reins in order to broaden the role of statistics to a more exploratory role. Instead of statistics playing the role of the inference police, it is becoming more like a research collaborator. While we may not want to go all the way, a journey in this direction is probably a good thing.

CONCLUSION

The discipline of statistics has been profoundly changed by the electronic revolution. Both the computation and communication aspects of this revolution have required that the mentors of our discipline reassess what they teach. Significant aspects of this change in attitude are the increase in importance of the graphical presentation of

results, the simulation of probability models, and the decreased importance of parametric inference.

REFERENCES

Cleveland, W. S. (1993). *Visualizing Data*. Hobart Press, P.O. Box 1473, Summit N.J. USA 07902-8473.

Weldon, K. L. (1995). The role of probability modeling in statistical inference. *Proceedings of the South East Asian Mathematics Society Conference: Mathematical Analysis and Statistics*: 37-49. Yogyakarta. Indonesia. July 1995.

Weldon, K. L. (1998). Probability for Life: A Data-Free First Course in Probability and Statistics. Submitted to the Journal of Statistical Education. (available from the author weldon@sfu.ca)