® FROM DATA TO GRAPHS TO WORDS - BUT WHERE ARE THE MODELS?

WELDON, K Larry
Simon Fraser University
Canada

The pioneers of statistics focused on parametric estimation and summary to communicate statistical findings. The tradition of basing inference on parametric fits is a central mode in statistics education, but in statistics applications, computer-based graphical summary is playing an increasingly important role. A parallel development has been the spread of statistics education to almost all disciplines, and thus the need to communicate statistical results to non-specialists has become more acute. These influences of more graphics and a wider distribution require adaptation in our statistics courses. This paper provides examples of, and arguments for, the use of simulation and graphical display, and the role of these techniques in enhancing the verbalization of analytical results. The immediate goal of the paper is to persuade those who design curricula for early statistics courses to provide a serious introduction to graphical data analysis, at the expense of some traditional parametric inference. The ultimate goal is to enable more students to communicate statistical findings effectively.

INTRODUCTION

The founders of statistical theory were not thinking of communicating to the masses when they invented the jargon needed to make precise their concepts. Consider the following:

"Normal" distributions are not always the usual ones.
"Significant" results are often of no real importance.
"Expected Value" is almost never expected.
"Sampling distribution" is not quite the same as a sample distribution.
"Standard Deviation" is not a deviation that is generally acceptable.
"Error" is not a mistake.
"Standard Error" is not an acceptable mistake.
"Regression" has little to do with moving backward.

This jargon is very confusing to the lay person, and even to many students of statistics. And, to further obfuscate the details, we summarize our methods and parameters using Greek alphabet symbols: $\mu,\sigma,\alpha,\beta,\gamma,\varepsilon,\rho$ and Σ are the common ones.

The formal definitions of these various terms do eliminate the suggested ambiguity, and so the statistics jargon does have a useful purpose. The problem is that jargon is not the language of the masses. We have devised a jargon language designed primarily for parametric inference. But are parameter estimates and inference statements really the only way to communicate statistical results unambiguously? The suggestion here is that graphics can be a sort of language that is accessible to a broad audience and that, with the help of ordinary English, can communicate statistical results clearly. This is especially useful when not only the summary, but the analysis of the data is done using graphical methods. This suggestion is not without risk – graphs can be as misleading as formal statistical summaries. However, it will be argued that the apparent results from a graphical display can be put into words more easily than the apparent results from a formal statistical summary. Moreover, the chance of gross errors with graphical analysis may actually be less than with a parametric analysis, when the analyst is less than expert in statistics. I will refer to literature to support these views but also find support from my own extensive experience in teaching and consulting.

In the following we illustrate by examples the power of graphics to extract the important information from data, and to facilitate the communication of the findings in English. These particular examples are chosen because they each have a fairly simple context that might easily arise in applied work, and yet are not easily analyzed and reported by the methods usually taught in early courses. The graphical route extracts the information from these contexts.

EXAMPLES: DATA → GRAPHICS → INFORMATION

EXAMPLE 1: TRENDS IN A TIME SERIES

Figure 1 shows five years of gasoline usage for a particular automobile. The data recording was done in the usual way by the driver at each fill-up, and has the usual sources of measurement error, so the trend is not very clear from the raw data.

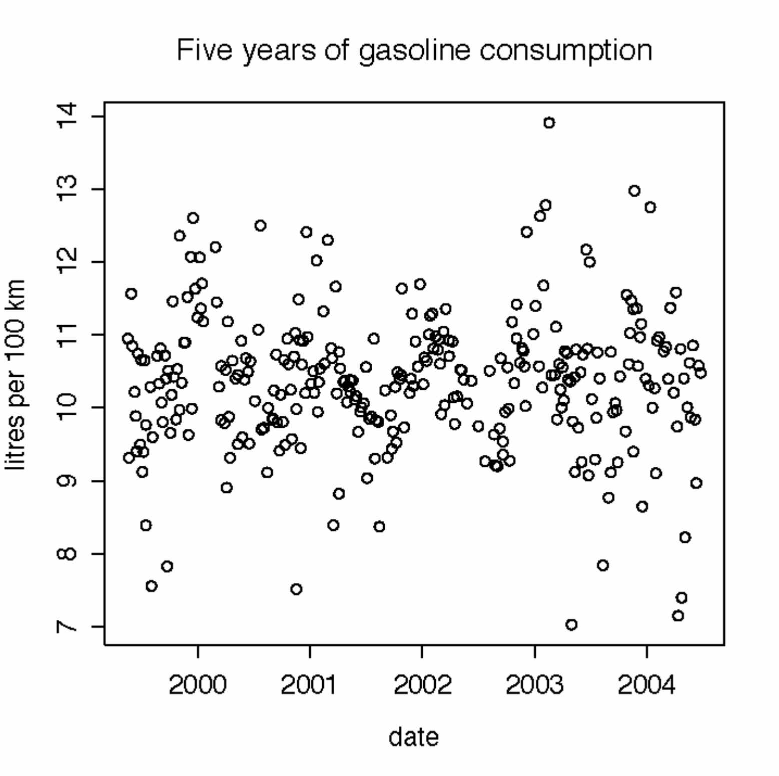

Figure 1. Raw data on five years of gasoline consumption.

However, using almost any nonparametric smoothing technique, the seasonal pattern is clearly evident – see Fig. 2.

This vehicle was used primarily for commuting 100 km each day, in the area of Vancouver, Canada. The almost-sinusoidal trend is clearly related to the pattern of temperature, even though the annual variation is moderate. One can report that the gasoline consumption rate follows a smooth increase from mid-summer to mid-winter and then smooth decrease until mid-summer. A numerical report of a sinusoidal parametric fit to the data would not be as successful a description. Even leaving aside the report of significance or measurement errors, the estimated sinusoidal parameters would have less meaning to most audiences. The graphical/verbal report seems to be all that is needed here to reveal the seasonal pattern.

Another excellent example of the power of graphics to extract information from time series data is the Melanoma incidence story described in Cleveland (1993). The identification of the sunspot cycle and even some long wave climatic changes from this data is remarkable, and obtained entirely from graphical methods. An important feature of the nonparametric smoothing method used there was that the matching of the sunspot cycle was a finding of the analysis rather than an assumption of it – it was not necessary to know the wave-length of the cycle in order to extract its signal, as is usually the case for seasonal data. Again parametric fitting was neither necessary nor even helpful.

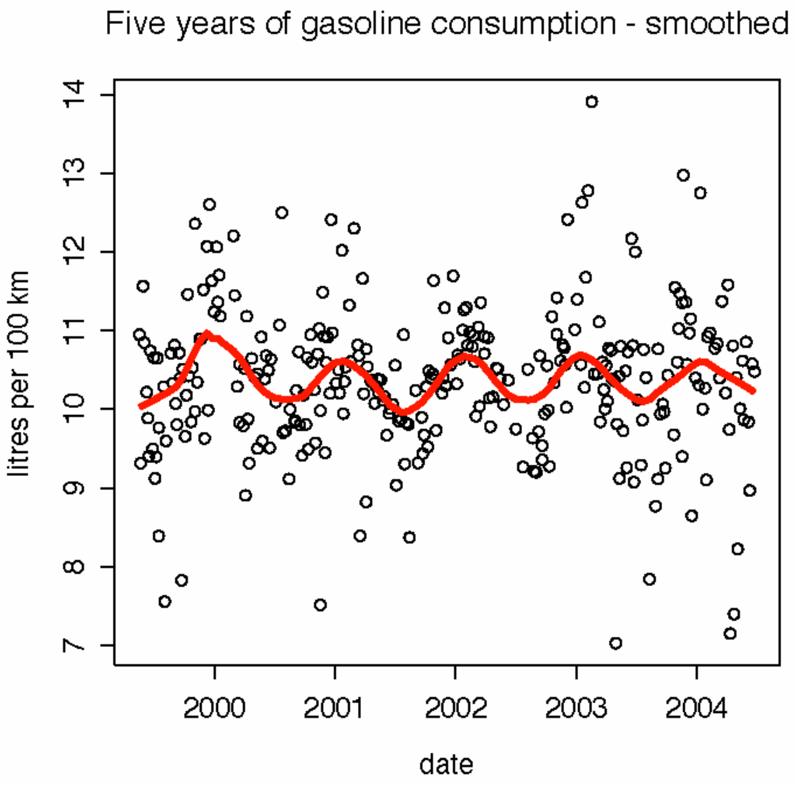

Figure 2. Lowess-smoothed data for gasoline consumption showing seasonal pattern.

EXAMPLE 2: RELATIONSHIP BETWEEN TWO VARIABLES

An interesting data set that can be used freely for non-commercial purposes is available from http://lib.stat.cmu.edu/datasets/bodyfat. It can be used to study the degree to which human body density can be predicted from several easily obtained measurements, in a population of Caucasian males. While it is an ideal set-up for a multiple regression model, since age, weight, height and ten other body-part circumference measurements are provided for predicting body density, the initial focus here will be on the relationship between height, weight and body density. It is common to use the body mass index (BMI) as a rough measure of body density and hence of excess weight. BMI is the ratio of body weight to the square of height, usually in metric units of kilograms and meters. It is known that excess weight can be characterized by body density, but that body density is not so easy to measure in a public health setting. The question of how well BMI predicts body density is of some importance.

Fig. 3 shows the relationship of body density and BMI. While body density has a clear relationship to BMI, which could be ascertained by simple regression, it is not so clear whether this is clinically useful or not. With $R^2 = .55$ and the standard error of prediction being about .013 it seems that it might be useful. We know the relationship is not perfect; but is it good enough?

An eyeball look at the graph shows that any prediction of Density from BMI could often be in error by as much as .01 or even .02, and that the high BMI readings could be associated with a Density from 1.015 to 1.055. It turns out this corresponds to a percent body fat between 19% and 38%, while usually the threshold for concern is about 30%. One would have to conclude that the relationship was not precise enough to be useful as a public health screening device.

The point of this analysis is that no parametric model was required to read the relevant information. It was not necessary to worry about linearity of the fit, or to report the standard error of prediction. Not only was the parametric analysis unnecessary, but also the communication of such an analysis would have been more complicated, and useful to a much smaller audience, than the graphical analysis.

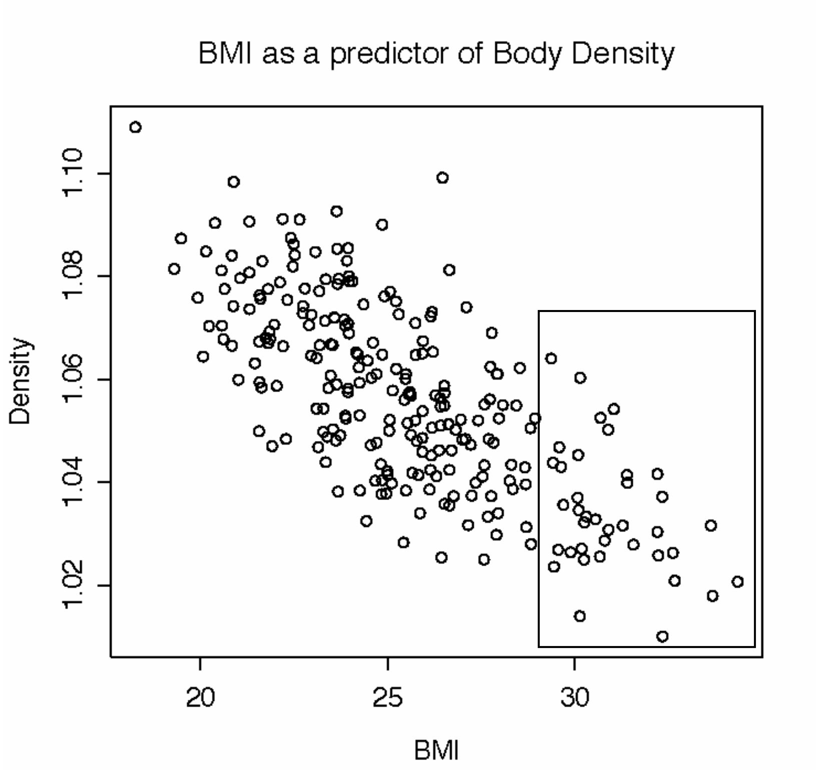

Figure 3. Is Body Mass Index (BMI) a useful public health index of body density?

What analysis would a new graduate in statistics perform in this situation? Would the result be presented graphically and verbally, or presented primarily with models and parameter estimates? Would the new graduate feel that the data had not really been analyzed until the regression had been done? We need to ensure that the bias toward parametric analyses is reduced in our teaching as graphical data analysis becomes more respectable and more effective.

EXAMPLE 3: MORE THAN TWO VARIABLES – INTERACTION

When we teach students about interaction between two variables in the association with a response, we may use a tabular approach via analysis of variance, but this alone does not usually convey the important scientific idea behind interaction. A more effective method is to use a graph like Fig. 4.

This data comes from Cleveland (1993), and relates to a study of wear rates on thirty different samples of tire rubber. The coplot method used for this graph is described in the Cleveland book and the software is now widely available. For simplicity, only two ranges of the conditioned variable are used in this example – usually more ranges are used even for a small data set like this.

Fig. 4 shows that there may be an interaction between tensile strength and hardness in predicting abrasion loss. The presence of interaction is debatable but the nature of it, if it is present, can be easily described. An interaction is indicated by the lack of parallelism of the smoothed fit in the two panels. To describe this suggestion in words to a scientist one might say: “Although lower levels of tensile strength tend to be associated with greater abrasion loss, there is a suggestion that at the high levels of hardness, this tendency may be absent. In other words, samples of rubber with low tensile strength may still have low abrasion loss if the hardness is great enough.” It is difficult to detect this kind of information through parametric modeling. Even if the interaction were detected by some persistent trial-and-error of parametric modeling, the graph would still be very helpful in supporting the fairly complex interaction described in words.

The point of this example is that parametric modeling really has no place in the analysis of this data. Graphs and words are all that is needed for the analysis and the communication of the result. How many students majoring in statistics graduate with an appreciation of this approach to analysis and summary?

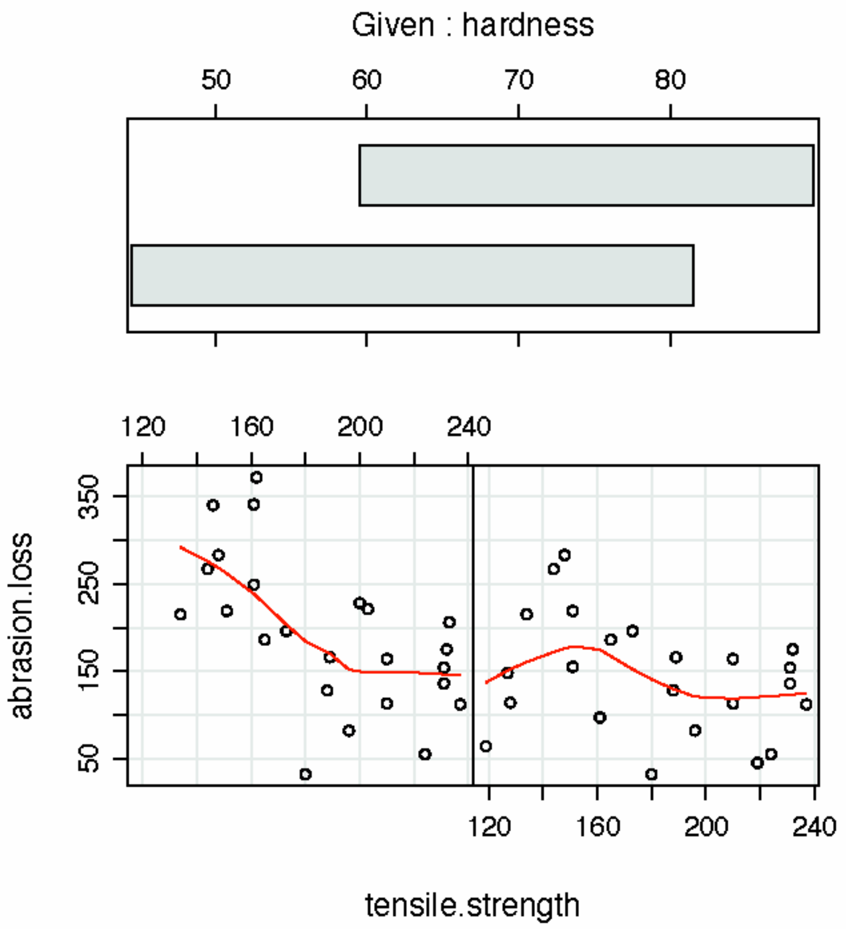

Figure 4. Coplot of Abrasion Loss as a function of Tensile Strength, conditioned on two overlapping ranges of hardness. There is a suggestion that there may be an interaction between hardness and tensile strength in the association with abrasion loss.

This approach is extendable to any number of variables, and data need not be quantitative, as Cleveland shows in his book and also in the Trellis software included in Splus (2005). When a multivariate relationship is complicated by higher order interactions, then of course the verbal explanation will also be more complicated. However, this graphical approach allows us to go a little further than the ordinary scatterplot approach in producing statistical results that can be communicated to a non-statistician.

EMERGING FROM THE PARAMETRIC TRADITION

The tradition of parametric modeling is deeply ingrained in statistical education. One often hears “First we must construct a model for the data”. This instruction would likely include the recommendation to plot the data first, but once the data is viewed and anomalies dealt with, further analysis would be based on a parametric model. Once one starts with a parametric model, it does seem compelling to focus on inference related to parameters of that model. This section provides an example which illustrates that modeling the data can distract the analyst from information that is directly accessible through a simpler graphical approach. While it may be argued that any information in data can be deemed to be parametric information, in practice “parametric modeling” has a more restrictive connotation, suggesting the fitting of simple parametric distributions to the data itself. The point of the example is that a more “data analytic” approach is often more effective than the traditional parametric approach, and at the same time will be easier to communicate to the end user. This argument has been made forcefully by Simon (1993) in his work on Resampling Statistics, but has yet to be widely adopted in statistics textbooks.

Most introductory courses start with a discussion of distributions and of measures of central tendency and variability, and soon the focus is on means and standard deviations. This is partly motivated by the need to prepare students for the parametric summary of distributions so that group comparisons can be reduced to a comparison of a few statistics, and partly to prepare students for the details of inference procedures for means and variances. In the focus on summary statistics, the interest in the actual form of the distribution itself is diminished. But with the aid of graphics, verbal communication of statistical results can proceed without these numerical summaries. One needs to question whether the best "standard" way to compare two distributions is to examine the difference in means. Consider the hypothetical data in Fig. 5

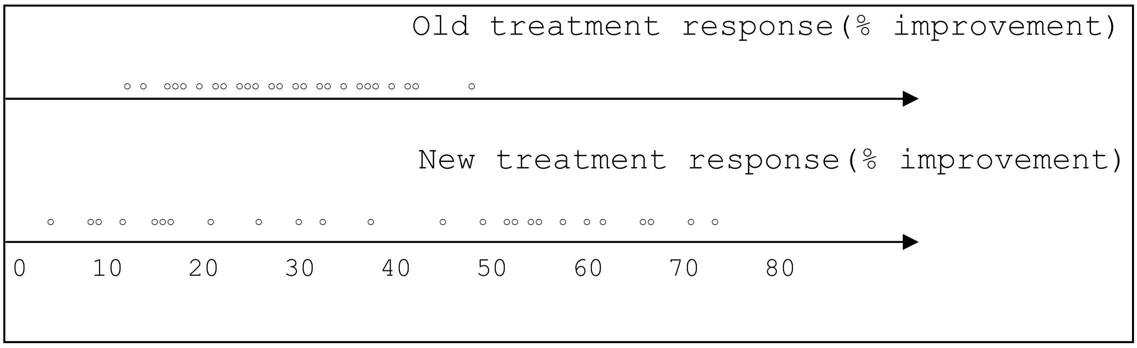

Figure 5. Simulated data of clinical decision problem.

Imagine that the above distributions are for two independent groups of 25 patients in a clinical trial to evaluate the efficacy of a new treatment, and that large values reflect better outcomes. The graph shows that the new treatment has promise for a large proportion of the group and would usually be of practical interest. However, the test for a shift in mean in a situation like this will often turn out "Not significant" (explained below): furthermore it is not clear that such a test is even relevant. The fact that the equi-variance assumption (of a t-test or Mann-Whitney test) is violated here is a technical problem but this is not the main problem. Rather the real problem is that a reduction of the comparison to a question about mean and standard deviation is inappropriate. Of interest in this case is a possible change in the whole distribution: location and spread, and possibly shape as well.

The example was simulated from two distributions, N (0,1) and N (1.3,3). These particular populations were chosen since the conventional tests would be significant for approximately fifty percent of the samples generated. This is neither a defense or a criticism of the use of an equi-variance test in this situation – only that the technical difficulty of doing an appropriate test is caused by an "irrelevant requirement" that a parametric test be done. The graph tells the story. A verbal description of the result is easy to do: "The treatment appears to be effective in a majority of patients, although possibly harmful to a few". Too many students learn that the first thing to do with data is "fit a model". Even if this includes a preliminary look at the graph, fitting a model would seem to play no additional role in extracting the information from this data.

Bryce et al (2001) reports an in-depth assessment of the needs of an undergraduate program in statistics. An emphasis on data analysis in the early courses is supported:

> "*There was unanimous agreement on the need for the undergraduate statistics curriculum to include a heavy emphasis on data analysis (perhaps more weight should be given to the "data" than the "analysis"). This emphasis reflects modern statistical practice, and also helps position students for careers in either industry or graduate school.*"

Moore (2001) makes the following observation in his discussion of trends in academic statistics:

> "*It is, of course, technology that is driving all the trends we are watching. Technology has changed statistics, so that our field has moved somewhat away from mathematics back toward its roots in data analysis and scientific inference*".

Difficulties with teaching traditional inference are well-known. One documentation of this is the thesis by Lipson (2000) in which the near-futility of teaching the sampling distribution of the mean in a first course is established, if a useful understanding of the concept is the pedagogic goal. The need for a broader approach to teaching statistics is evidenced by Harraway (2003), who surveyed employers about statistics needs of their operations. A proposal for broadening the scope of statistics education is suggested by Cleveland (2001), and his perspective is from an extensive immersion in consulting. A similar view is presented by Kettenring (1997) based on his extensive experience in industry and with the American Statistical Association. The common theme is a recommendation to change the way we teach statistics, at all levels, to include a broader set of tools and concepts. See also Cobb (1993). One way to do this is to put more effort into data analysis at the expense of traditional parametric inference, an idea supported by Singer and Willett (1990). Graphical methods are central to modern data analysis and increased emphasis on these graphical methods is likely to follow from these recommendations.

One trend in the practice of statistics that is bound to further enhance the ability of modern technology to assist the scientist is the availability of *interactive* software tools (Hammerman and Rubin, 2004; Efron and Tibshirani, 1997; Cleveland, 1993). This will require another round of revision of materials for teaching statistics. Perhaps the increase in emphasis in our early stat courses of graphical data analysis is an intermediate step preparing for this future change.

SUMMARY

The examples support the proposal that graphics can be a replacement for parametric analysis, and in some cases provide a better method of analysis. Because the graphical methods are often simpler to learn, to explain and to report, the methods are accessible to a wider audience than are parametric methods. Early courses in statistics should provide a serious introduction to graphical analysis - with much more emphasis than is usually given to parametric analysis. This is not only recommended for statistical literacy courses, but for all introductory courses. The impact of the introduction of statistical literacy courses could be magnified if the same revolution in content were applied to all introductory courses.

In summary, the argument by example has been this: if we teach students to take graphical data analysis seriously, they will be better prepared to communicate the results of statistical analyses. The opinion expressed here is that there is too much emphasis on parametric modeling and its associated inference, at least in the first one or two courses. Most users of statistics will not take more than one or two courses, and we need to provide more deliberately for this. Moreover, many of the students that apparently are mastering parametric inference techniques are actually learning only how to follow procedures. When these students stumble on the inevitable complications of applied research, their associates will think the discipline of statistics is lacking, and this does our discipline a disservice. By limiting our teaching to concepts that are understandable and broadly useful, we may better teach our students and also benefit the reputation of statistics. This may mean an increased emphasis for graphical methods of analysis and summary.

ACKNOWLEDGEMENTS

The author wishes to thank John Brewster, Douglas Stirling and John Harraway for their comments on an early draft of this paper. The argument expressed here is nevertheless my own. The graphs have been produced in R (2004).

REFERENCES

Bryce, G.R., Gould, R., Notz, W.I., & Peck, R.L. (2001). Curriculum guidelines for Bachelor of Science degrees in statistical science. *American Statistician, 55*, 7-13.

Cleveland, W.S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International Statistical Review, 69*, 21-26.

Cleveland, W.S. (1993). *Visualizing Data*. Hobart Press, Summit, NJ.

Cobb, G.W. (1993). Reconsidering statistics education: A National Science Foundation Conference. *Journal of Statistics Education, 1*(1), available at http://www.amstat.org/publications/jse/.

Efron, B., & Tibshirani, R. (1997). Computer-intensive statistical methods, *Encyclopedia of Statistical Sciences, 1*, 139-148.

Hammerman, J.K., & Rubin, A. (2004). Strategies for managing statistical complexity with new software tools. *Statistics Education Research Journal, 3*(2), 17-41.

http://www.stat.auckland.ac.nz/~iase/publications.php?show=serj#archives/

Harraway, J.A. (2003). The use of statistics in the workplace: A survey of research graduates in several disciplines. Bulletin of the International Statistical Institute, Volume LX, Book 2, 2-5. (Proceedings of the 54th Session of the ISI).

Lipson, K. (2000). Determining the relationship between the concept of sampling distribution and the development of understanding of inferential statistics. PhD thesis, Swinburne University. Melbourne.

Moore, D.S. (2001). Undergraduate programs and the future of academic statistics. *American Statistician, 55*, 1-6.

R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org.

Simon, J. (1993). *Resampling: The "New Statistics".* Wadsworth.

Singer, J.D., & Willett, J.B. (1990) Improving the teaching of statistics: putting the data back into data analysis. *American Statistician, 44*, 223-230

Splus (2005). www.insightful.com and Trellis Display at http://cm.bell-labs.com/cm/ms/departments/sia/project/trellis/s.html