Less Parametric Methods in Statistics[1]

K. Laurence Weldon[1]

Abstract

Despite forty years of revolution in the tools available for statistical analysis, the current academic tradition in statistics is remarkably similar to the pre-computer tradition. This tradition is rooted in parametric modeling, estimation, and testing, and optimal procedures based on these models. The paper argues for a shift in emphasis away from parametric modeling and estimation to graphical summary, from omnibus optimal techniques to those that are more context-specific, and from goals of objectivity to goals of revelation. It is suggested that the emphasis in statistical education should be rebalanced to reflect certain modern computer-based techniques.

1 Introduction

It is widely recognized that the advent of the computer has changed the scope of the statistics discipline. (Efron(1993), Moore et al (1995), Kettenring(1997), Tukey(1997), Cleveland, WS (2001), Moore(2001). However there are still aspects of the theory currently taught as essential basics, that were motivated in pre-computer times, but whose relevance in the modern context is diminishing. Some new computer-based methods that tend to be portrayed as second-choice alternatives need to be re-appraised and upgraded in status. For example, the power of *resampling* has been gaining momentum, especially since the landmark paper on the bootstrap by Efron(1979). Implications of the resampling idea for the teaching of introductory statistics has been stressed by Simon(1993).

In addition to the impact of resampling on statistical education and practice, the increasing importance of graphical methods (Cleveland(1993), Bowman(1997), Fisher(2001)), simulation (Efron(1993), algorithmic methods (McLachlan (1997)), exploratory data analysis (Tukey(1997), and Hoaglin et al (1991)), and Bayesian

[1] Larry Weldon, Associate Professor, Department of Statistics and Actuarial Science, Simon Fraser Univertsity, Vancouver, BC, Canada. V5A 1S6. email: weldon@sfu.ca

inference (Efron (1998) have all derived from the impact of computers. Are these simply add-ons to the traditional theory, to be treated as peripheral topics and left to advanced courses, or should they have a more fundamental impact? The answer matters because the future of statistics, both theory and practice, depends upon what is offered in the mainstream programs of statistical education. The conservatism of our teaching has already lost huge areas of data-based methods to other disciplines, such as data mining (Friedman (1999)).

The mathematical theory of statistics has its roots in the pre-computer age. The advent of the computer was seen to expand the realm of many of the theoretical models originally proposed, like Bayesian analysis, Monte Carlo methods, or even multiple regression. However, another effect of computers was to make entirely new approaches to statistical analysis feasible (Efron, B., and Tibshirani, R. 1997). While many authors have drawn attention to these new approaches, the developments have had a fairly minor impact on undergraduate textbooks. Since these textbooks largely determine the content of their associated courses, the teaching of statistics has been slow to adapt to the modern context of statistical practice. As Moore (1995, p 251) has suggested:

"I suggest that the limited impact of technology on teaching is rooted in cultural resistance to change in colleges and universities, …"

A similar view is expounded by Chambers (1993) in his outline of the difference between "greater statistics" and "lesser statistics". He makes it clear that contemporary statistics instruction as relating mainly to "lesser statistics".

Kettenring(1997) says it this way "The question I wish to raise is whether the 21st century statistics discipline should be equated so strongly to the traditional core topics and activities as they are now. Personally I prefer a more inclusive interpretation of statistics that reflects its strong interdisciplinary character."

Finally, to emphasize the slow nature of change to our discipline, note that Tukey(1962) warned of the narrow tradition the discipline was developing.

In the following sections, some topics that tend to be viewed as central to our discipline are re-examined in the light of the modern statistical context. We wish to add to the growing doubt concerning the conventional approach to instruction in statistical theory, a movement that has a long history even though its impact has been small. In addition, we suggest with examples the directions for change. Course designers and textbook authors may wish to consider these arguments.

2 The Changing Environment of Statistical Practice

2.1 The traditional focus on parametrics

Are parametric models and parameter values really the best way to summarize distributions, and make comparisons of distributions? The issue here is not whether to use the t-test or the Mann-Whitney test - both tests are concerned with a location parameter, and the Mann-Whitney differs only in the model for the data, not in the feature of interest, the location parameter. The issue I want to address is whether the feature of interest is really a simple parameter value, or is it a whole distribution. Has our focus on the estimation of parameters been the result of limited options for distribution summary in the absence of modern computers and software? ? Are investigators really interested in a distribution's parameters, or in the whole distribution? The following example will help to illustrate this choice. Although the example is worked through in some detail, it should be stressed that it is the general issue of the focus on parameters, not the particular example, that is the issue here.

Imagine the outcomes shown in Figure 1 from a clinical trial aimed at comparing the effectiveness of two treatments. Assume the quantity measured is something that is to be maximized for clinical effectiveness.

Figure 1 shows dotplots of simulated samples of size 25 from N(0,1) and N(1.3,3) distributions. The Treatment variables represent responses of patients, with a large value being a good response. The question is whether Treatment 2 is better than Treatment 1, based on these data.

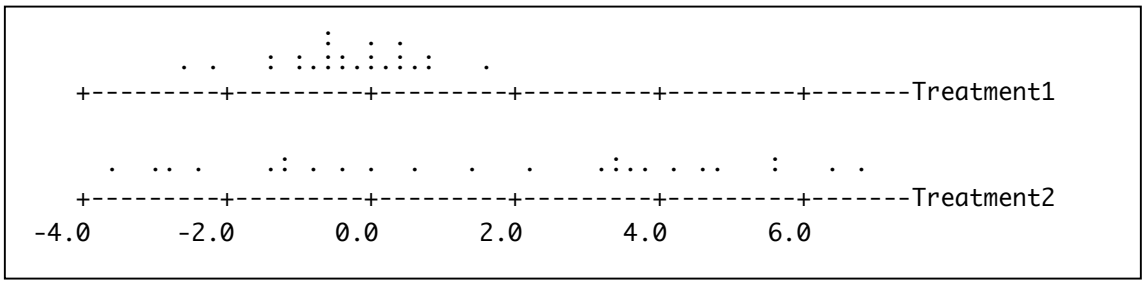

Figure 1 Data for a comparison of treatments

The parametric approach to this data would be to test for a difference in mean and/or standard deviation between the two treatments. Informally there does seem to be a higher mean and a higher standard deviation associated with Treatment 2. But how should this be verified with standard tests? An F test for the variance ratio will certainly show the second distribution to have the greater variability. If we use a t-test or a Mann-

Whitney test for a location shift, in spite of the clear violation of assumptions, we get equivocal results, rather than a clear indication of the superiority of treatment #2. In the particular example shown, the p-value is about .05.

The data is actually simulated from two normal distributions:

1. N(0,1)
2. N(1.3,3)

In this case it can be verified that the shift in mean will only be statistically significant at the .05 level in approximately one-half of the pairs of samples from these populations, and this is true whether one uses the t-test or the Mann-Whitney. So there is a very good chance that the researcher in this situation would be led to conclude that treatment 2 is no better, on average, than treatment 1, if these inappropriate location tests were used. The example we have chosen from the hypothetical population for illustrative purposes is one that has a median location differential: that is, there are equally frequent samples that show more and less differential than this one. In this section, we explore both conventional and unconventional analyses of these data.

A test of the variance ratio would almost always be significant, so a traditional conclusion could easily be that, not only is Treatment 2 is no better on average, but it also appears to be less consistent! Of course, the researcher would likely believe, based on the graph, that for most patients, treatment 2 provides a superior result. Perhaps the researcher would invent a test that uses the observation that about half of the treatment 2 responses are better than the highest treatment 1 response, and get the result published anyway. This last-resort strategy is not recommended for obvious reasons. The point is that *the traditional approach fails because of a technicality that has nothing to do with the science of the situation*. The unnecessary assumption is the assumption of homoscedasticity, or approximate homoscedasticity, for the t-test or for the Mann-Whitney test.

What would be an appropriate report of the above outcome? First of all, the graph itself does contain all the information in the outcome of the experiment, and the only question is how this information should be interpreted. This interpretation is not entirely a statistical question - there are some medical and ethical issues involved. However, if we leave aside these issues, the graphical display itself shows without question that the second treatment is associated with generally higher values. We might suggest that the second treatment looks better. A radical suggestion might be that formal tests are unnecessary.

However the issue of reproducibility can be studied formally without using a parametric model for the data, through a re-sampling strategy. If one asks, "How often will Treatment 2 produce a result that is superior to Treatment 1?", based on the data available, the resampling estimate is an estimated probability of 0.65. This is obtained by sampling with replacement from each distribution and comparing pairs of simulated values.

The fact that this is greater than 50% suggests the superiority of Treatment 2, but does not establish it. To discount sampling error, we need to know the precision of the 0.65 probability estimate. One way to tackle this is with the raw bootstrap. If one performs the same estimation procedure starting with bootstrapped re-samples of each treatment response, the result is that the estimated sampling distribution has a standard deviation of 0.08 around the mean of 0.65, and the distribution appears normal. This suggests that the nominal proportion, 0.5, is 1.9 standard deviations below the observed mean, 0.65, of this sampling distribution, and that the superiority of Treatment 2 is reasonably well established.

This information is based entirely on the particular sample of 50 data values, and not at all on the models used for the simulation of the data. However, if we use the normal models N(0,1) and N(1.3,3), one can compute that the true proportion (underlying our estimate of 0.65) was actually 0.66, and the standard error of our bootstrapped estimation procedure can also be found (by simulation) to be about 0.12. This is somewhat greater than the 0.08 based on our single sample in this case, indicating that the significance 0.66 might have been missed from a single sample (with a one-sided P-value of about .10). However the example was chosen to be marginally significant by parametric tests so it is not surprising that it would be marginally significant by non-parametric approaches.

It might be argued that the above estimate of 0.65 is itself a parametric estimate that follows the traditional parametric approach. However, we have shown that, if one really must have a parametric summary of the data, then it should be in terms of parameters of practical interest – outcome parameters rather than parameters of the data distribution. Moreover, the estimation of this parameter and its variability was done without reference to a parametric model for the data itself.

This data-based analysis would not be as reliable as a parametric-data-modeling approach *when the parametric model for the data is correct*. However it is an attractive alternative that is very adaptable to a variety of contexts and very easy to explain to a scientific researcher. Moreover, it avoids the technical problems having nothing to do with the science of the situation. Parametric modeling of the data can be an unnecessary, and even confusing, step.

Let us now further review the parametric analysis of this data with the "not-parametric" analysis. The term "nonparametric" will be avoided here since it suggests an approach such as the Mann-Whitney test that is actually parametric in the location parameter.

The not-parametric analysis focused on the real feature of interest to the clinician, the relative performance of a typical patient under the two treatments. The clinician is not really interested in how the average patient does, but rather in how many patients will do better using Treatment 2, and how much better or worse they will do. Of course, without intra-patient experiments, we do not have an accurate intra-patient estimate of the differential, but we do have an estimate of how many patients do better (65%).

The analysis just performed on this hypothetical data is "not-parametric". The feature of interest to the clinician is described as directly as possible. No parametric model of the data itself is used for this description. The degree of consistency in the superiority of Treatment 2 is estimated. The reproducibility of the superiority of Treatment 2 is assessed.

These results are achieved easily through use of computer intensive techniques: graphical display, simulation, and resampling. All the problems inherent in the usual "parametric" approach (which includes the so-called "nonparametric" tests) are avoided.

If the distributions were markedly non-normal, as with very skewed distributions, inappropriate conclusions could result from the parametric approach. Note that this technical problem disappears when the distribution is considered as a whole, and a not-parametric approach is used.

We recommended a graphical display as an important part of the data summary. For many purposes, graphical displays are adequate distribution summaries, and allow efficient communication of distributions. The ability to communicate sample distributions graphically makes a numerical summary, via parametric estimates, less important. Moreover, numerical information can still be communicated graphically, but in a less-restrictive way. For example, estimated quartiles can be portrayed graphically along with the data values themselves. Estimates of location and scale parameters are far less informative than the entire distribution, portrayed graphically. The reputed efficiency of parametric descriptions is questionable in an age of easy graphical summaries provided through software.

The overall suggestion from this example is this: in a world of easy access to powerful statistical software, parametric summary of the observed data is often inappropriate. Information from data-based studies is often best summarized graphically, and questions of reproducibility can be asked with respect to whole distributions, and not merely with respect to certain parametric summaries. Statistical practice seems to be evolving towards a less parametric approach.

2.2 The focus on estimates of regression coefficients.

Regression is another area where the focus on parameters may be overdone. Regression methods are often proposed for studying "relationships between variables". See for example Devore (2000) p 488, Moore and McCabe (1993) p 117, Wild and Seber (2000) p.503. Based on the model $Y = \beta X + e$, we want to study the relationship between X and Y. However, a more precise objective is revealing: regression models are used either to predict Y from X, or to study the model that would do this prediction. When X is random and we merely want to describe the relationship between X and Y, with no focus on prediction, then minimizing the sum of squares of $Y - \hat{Y}$ is not appropriate. If a fit were required in this situation, the least squares approach should find a fit which

minimizes the sum of squares of $\| Y - \hat{Y}_{\perp} \|$, the Euclidean distances of data points to the fitted line. Such a fit is not called regression, and it is certainly not suited to prediction or describing a predictive relationship. Regression theory assumes prediction of Y from X, or a predictive model for this, is required.

What we do with our predictive fit depends on whether we are interested in the ability to predict Y from X, or in the detection of consistent predictors X of Y. In the former, we focus on $Y - \hat{Y}$ to judge the quality of the model, whereas in the latter, the actual coefficients of X are of primary interest.

In the case of our actually using the fit for prediction, the estimate of the regression coefficients is really of minor interest. We simply want the predictor to work well, and how this is achieved is not so important. The problem of colinearity is not really a problem in this setting since interpretation of the predictive equation is not important as long as the predictions are accurate. In other words, estimates of β in this setting are not very useful by themselves -- they are only useful as employed in the predictive equation. There is no need in this case for confidence intervals for β_i .

On the other hand, if our interest is in the identification of useful predictors, to find out what variables actually do make a difference to our response variable Y, then we usually frame the question as a null hypothesis that $\beta_i = 0$. Again, the estimation of β_i is not of interest. Of course, the question of whether or not $\beta_i = 0$ is unanswerable in a strict sense, and it may be argued that an estimate of β_i is more useful than a test of the hypothesis $\beta_i = 0$. However, if the research question is really whether an explanatory variable has *any* role in prediction, the hypothesis test is really the more direct approach. Our methods for testing $\beta_i = 0$ or any other parametric statement have been criticized over the years from many directions, but at least the question we are asking seems to be the right one. A wrong question is "Is the population value of β_i close enough to zero that we can conclude it is equal to zero?" In other words, estimation is not our goal in this context.

In some situations, it may be of interest to know the actual size of the regression coefficient, but the suggestion is that this is fairly rare. Even in comparisons of regression coefficients, when the relative importance of predictors is desired, the question will often reduce to a test of $\beta^* = \beta_2 - \beta_1 = 0$. Even in this situation, the estimate of regression parameters is not really of primary interest.

These arguments suggest that interval estimation of β in the regression model $Y = \beta X + e$ is not really as informative as a focus on its estimation would suggest. The estimation of regression coefficients is not often the ultimate goal, but rather an intermediate computation. Again, the focus on parametric estimation may distract the statistician from the questions of practical importance.

Note that it is not the theory that is at fault here, but rather the customary use of the theory. So much emphasis has been put on fitting a parametric model for the data, that thoughtful consideration of the information that is the goal of the analysis, is minimized. This emphasis appears in practice and in instruction. In presenting statistical theory, we must not emphasize the model so much that students forget that it is reality that we are trying to describe, and that the model is merely a simplified approximation. Moreover, our optimization methods are usually model-based, and a healthy disrespect for "optimal" procedures should be encouraged, since the optimality is usually conditional on the model being correct.

In this section the failings of the data->parametric model->estimation of parameters paradigm have been highlighted. We have tried to suggest the relative merit of the data->graphics->not-parametric summary and the suggestion is that we give this more emphasis in our training of statistical practitioners and theoreticians.

2.3 The parametric view of prior knowledge

How should prior knowledge about data context influence data analysis? This clearly depends on the purpose of the analysis. The Bayesian approach is appropriate for updating knowledge (either subjectively or objectively), while the classical approach attempts to extract information from data that is context-free, or in other words, free of prior belief. However neither approach is very successful in achieving its respective goal. The Bayesian approach attempts to summarize prior knowledge with a prior distribution of a parameter, whereas prior information is usually more complicated than that. The classical approach incorporates prior knowledge in the specification of a class of probability models from which a fit is selected, but the impact of the prior information in this case is quite hard to judge. The robustness of the fit to the specification of the class of probability models may be viewed as a strength or a weakness, but either way the impact is poorly controlled.

The robustness movement attempted to solve the problem of model mis-specification by arranging that the model specification have little impact on inference. However, researchers would ideally like to have the data-based information combined with their prior information in an organized way, as well as having the data-based information separated out. Researchers will act on the basis of the combination of prior information and data, even if they only report the information in the data itself.

Even descriptive statistics procedures can be altered by prior information – for example the choice of a smoothing parameter in a non-parametric smooth.

It may be useful to explore methods for incorporating prior information of all kinds into the analysis of data, but with a measure of the sensitivity of the inference output to the prior information.

A start in the consideration of a theory incorporating model-building into statistical theory has been suggested by Cleveland and Liu (1999). They extend the Bayesian idea of a prior to *external information*: through scientific knowledge, such as linear or exponential growth, one judges the credibility of neighborhoods of models. The Bayesian likelihood is extended to *exploration information*: through data analysis, such as probability plots, one uses intuition and graphical techniques to judge the plausibility of data given a neighborhood of models. One might call this "holistic statistics"! Cleveland and Liu call it *model specification inference*.

Hierarchical Bayes models have been proposed to accomplish this end. However, this very mathematical approach seems too parametric to capture all prior information of general scientific value.

The goal of a purely objective method of extracting information from data should perhaps be set aside. As the discipline of statistics is used by an expanding range of researchers, we should aim at exposing the hazards of subjectivity, rather than trying to eliminate it from our analytical processes. The value of intuition in scientific progress is, with good reason, widely believed, and statistical methods should adapt to this reality. We have been reminded of the need for informal inputs to data analysis by Cleveland (2001).

The argument in this section has been that the incorporation of prior knowledge into data analysis is still a very informal process, and current theories, Bayesian or classical or other, still need a lot of development before the informality can be reduced. We need to put less emphasis on optimal procedures (conditional on the model being correct) and more on the hazards of jumping to conclusions.

2.4. Do statistical techniques provide the tools necessary for decision-making?

When decisions depend on data-based information that is subject to unexplained variability, we try to reduce the impact of the uncertainty as much as possible. Such decisions are typically about whether two or more samples are from the same population, or not. Our prior beliefs about the state of nature usually determine what strength of evidence is required to change these beliefs. A surprising result usually requires very strong evidence to change our opinion. In medical studies, usually more than one study with strong evidence is required to establish a surprising result. The observed p-value that would actually change a belief will certainly depend on the strength of the prior belief. Decision making must take into account prior belief.

Another factor affecting the strength of evidence for a "significant" result is the cost of making a wrong decision. Yet neither the strength of prior belief nor the cost of wrong decisions is formally incorporated into the p-value approach, whether "significance testing" or "hypothesis testing". This reminds us that the p-value approach is not a formal decision-making procedure (in spite what we tell our introductory students).

Statisticians have known for decades how to incorporate loss functions and prior belief into a decision-making procedure. But the difficulty in specifying these aspects of the decision-making framework in a scientifically acceptable way has retarded their widespread use. An interesting text for senior undergraduates "Making Hard Decisions" (Clemen and Reilly, 2001) has been recently published attempting to close this gap. Too many textbooks attempt to portray the p-value approach as a respectable decision-making procedure. These textbooks give the impression that the p-value approach is an objective method for decision making.

However this objectivity is clearly an illusion. The choice of critical p-value is arbitrary, and the definition of the null hypothesis and alternative hypothesis is based on prior belief. Moreover, since the cost of making wrong decisions (or the utility of making correct decisions) is usually not perfectly known, correct quantification of the decision-making process is not usually possible.

Nevertheless, decision-making should incorporate these vaguely known features as well as possible so that good decisions can be made. To back-off to the part that can be done objectively and mechanically seems to be ill-advised. And yet this is what we often urge our students and applied researchers to do. To be acceptable to editors of scientific journals, data-based research conclusions must be supported by traditional testing. Scientists who do not understand the limitations of traditional inference are strong advocates for its use. But it is teachers of statistics that generate this misguided belief.

Textbook authors need to be more careful in describing how statistical testing is to be used, and how it relates to decision-making. Moreover, we need to be more flexible in our acceptance of the respectability of the subjective aspects of decision-making. In fact, we should be emphasizing the importance of these subjective aspects of the process of decision-making under uncertainty. We also need to publicize to journal editors the limitations of "tests of significance" for evaluating whether or not a certain scientific outcome is worthy of publication.

This theme ties in to our overall theme of over-emphasis on parametric models. If we suppose that the focus of data-based research is a decision about the value of a parameter, in many cases we will have oversimplified the science of the situation too much.

2.5 The least-squares criterion

Is least squares a reasonable criterion for estimation and curve-fitting? Certainly an important aspect of our theory of parametric modeling is the least-squares criterion.

For numerical or graphical summary of a relationship between two variables, when one variable is considered dependent on another, we usually search for the function in our model class that minimizes the sum of squared deviations in our data set, or sometimes a weighted sum of squared deviations. There is an implication in this criterion

that our fit should try hard to accommodate outlying observations, because of the amplification effect of squaring the deviations. The practical implication of this is that a decision needs to be made in the curve-fitting process whether or not certain "outlying" points are to be ignored, at least temporarily. This complicates the fitting procedure. Methods robust to outliers like the bisquare procedure as described by Cleveland (1993) modify least squares in a semi-objective way to avoid this complication. The bisquare approach iteratively re-weights the data values according to the residuals. The bisquare is just one of many such procedures, but the idea is that procedures exist to improve on the purely subjective decision to keep or set aside certain outlying observations.

However, what at first may seem to be the methodological problem of dealing with outliers may actually be a problem with the least squares criterion itself. Is the squared distance of a data value from a proposed fit really the appropriate contribution of the value to the measure of lack-of-fit? Perhaps a procedure for seeking a conditional mode would be more generally acceptable. This would be based on nonparametric density-estimation procedures. The smoothness parameter in the density fit would allow us to effectively modify the least squares criterion. Fits determined in this way can be assessed using residual analysis, as usual. This modal method is not seriously proposed as the answer, but is mentioned as typical of the kind of option available when least squares is questioned. Modern statistical software makes these options feasible.

The least squares criterion fits nicely with normal models, maximum likelihood estimation, the central limit theorem, minimum variance unbiased estimation, and the familiar averaging process. But the mathematical tractability of this criterion may have blinded us to the ease of more flexible methods that have been made feasible by modern software. The practice of statistics should include more of these smoothing kernel approaches to estimation. Available software removes the algebraic and arithmetical obstacles, and also the problem of reporting the result of such smoothes, since a graphical report is ideal. Nonparametric smoothing should be recognized as a method appropriate for the general user, and not only for advanced data analysis. Least squares should be recognized as a mathematically appealing technique that is gradually being replaced as its limitations for applied data analysis are recognized.

3 Summary and Proposal

Academics in statistics share a common interest in exploring probability models, developing methods for fitting models to data and making inferences about model parameters. However this activity has led us into a larger discipline which includes graphical methods, data summary, research design, exploratory inference, and the construction of complex simulation studies. Thus the discipline as it is practiced has undergone a major intellectual expansion - but it is turf that is being claimed by other

disciplines (Friedman(2000)). We must adapt our teaching traditions to meet these new challenges. Our textbooks must adapt more completely to the modern world of computerized data analysis.

To be specific about the topics that need more emphasis, I suggest the following topics be added to the early courses and their textbooks:

1. The limited role of parametric models and the opportunities for resampling approaches to inference.
2. Nonparametric smoothing and the graphical representation of relationships among variables.
3. The influence of prior information of all kinds on statistical procedures.
4. Data-based decision making and the contrast with the p-value approach.
5. Alternative loss functions to least squares and the impact on practice, especially on the handling of outliers.

These extensions of our traditional material are suggested to remove the dependence of statistical practice on parametric models for observed data. Other current trends such as greater use of simulation and resampling techniques have not been discussed here but are covered in Simon(1993) and Weldon (2001).

Even the best popular textbooks used for undergraduate statistics courses do not have much to say about these topics. (Moore and McCabe (1993), Freedman, Pisani, and Purves (1995), Wild and Seber (2000), Devore (2000), De Veaux and Velleman (2004).) Work needs to be done to see if these very practical ideas and tools can be presented successfully at the undergraduate level. Some preliminary results from a course taught to first year students with no prerequisites, in which many of the recommended topics were included, shows that the material can be absorbed.
(Weldon, 2004). It is not enough to include these topics in graduate courses and professional seminars. As Cleveland(2001) says:

"A very limited view of statistics is that it is practiced by statisticians. … The wide view has far greater promise of a widespread influence of the intellectual content of the field of data science. "

References

[1] Bowman, A.W and Azzalini, A (1997) *Applied Smoothing Techniques for Data Analysis*. Clarendon Press. Oxford.

[2] Chambers, John M, (1993) Greater of Lesser Statistics: A Choice for Future Research. Statistics and Computing 3:182-184.

[3] Clemen, R.T and Reilly, T. (2001) *Making Hard Decisions*. Duxbury. Pacific Grove, Ca.

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

[5] Cleveland, W.S. and Liu, C. (1999) A Theory of Model Specification Inference. Talk transparencies: Joint Statistical Meetings, Baltimore, August, 1999.

[6] Cleveland, W.S. (2001) Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics. International Statistical Rev. 69:21-26.

[7] De Veaux R.D., Velleman, P.F. (2004)
IntroStats. Pearson. New York.

[8] Devore, J.L.(2000) *Probability and Statistics for Engineering and the Sciences*. Fifth Edition. Duxbury. Pacific Grove, California.

[9] Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics 1:1-26.

[10] Efron, B. (1993) Statistics in the 21st Century. Statistics and Computing, 3:380-382

[11] Efron, B. (1998) R. A. Fisher in the 21st Century. Statistical Science 13: 95-114

[12] Efron, B., and Tibshirani, R. (1997), Computer-intensive statistical methods', Encyclopedia of Statistical Sciences, 1 139-148.

[13] Fisher, NI (2001) Graphical Assessment of Dependence: Is a picture worth 100 tests? American Statistician 55: 233-239

[14] Friedman, J.E. (1999) The Role of Statistics in the Data Revolution. Bulletin of the International Statistical Institute: 52nd Session. Book 1:121-124.

[15] Freedman, D., Pisani, R. ,Purves, R. (1998) *Statistics*. Third Edition. Norton. New York.

[16] Hoaglin, D. C., Mosteller, F., Tukey, J W. (Ed) (1991) *Fundamentals of exploratory analysis of variance*. Wiley-Interscience; Somerset, NJ

[17] Kettenring, Jon (1997) Shaping Statistics for Success in the 21st Century. JASA 92:1229-1234.

[18] McLachlan, G. J, Krishnan, T. (1997) *The EM algorithm and extensions*. Wiley-Interscience,Somerset, NJ.

[19] Moore, DS (2001) Undergraduate Programs and the Future of Academic Statistics. American Statistician 55:1-6.

[20] Moore, D.S., Cobb, G.W., Garfield J., Meeker, W.Q. (1995) Statistics Education Fin de Siècle, The American Statistician, 49:250-260.

[21] Moore, D.S. and McCabe, G.P. (1993) *Introduction to the Practice of Statistics*. Second Edition. Freeman. NY.

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

[23] Tukey, John W. (1962), The future of data analysis, The Annals of Mathematical Statistics 33:1-67

[24] Tukey, J.W. (1997) More Honest Foundations for Data Analysis. Journal of Statistical Planning and Inference, 57:21-28.

[25] Weldon, K.L. (2001) Informal Probability in the First Service Course. Unpublished manuscript. (Weldon@sfu.ca)

[26] Weldon, K.L.(2004) Experience with a case-oriented introductory course in statistics. Unpublished manuscript. (Weldon@sfu.ca)

[27] Wild, C.J and Seber, G.A. (2000) Chance Encounters: *A first Course in Data Analysis and Inference*. Wiley. New York.