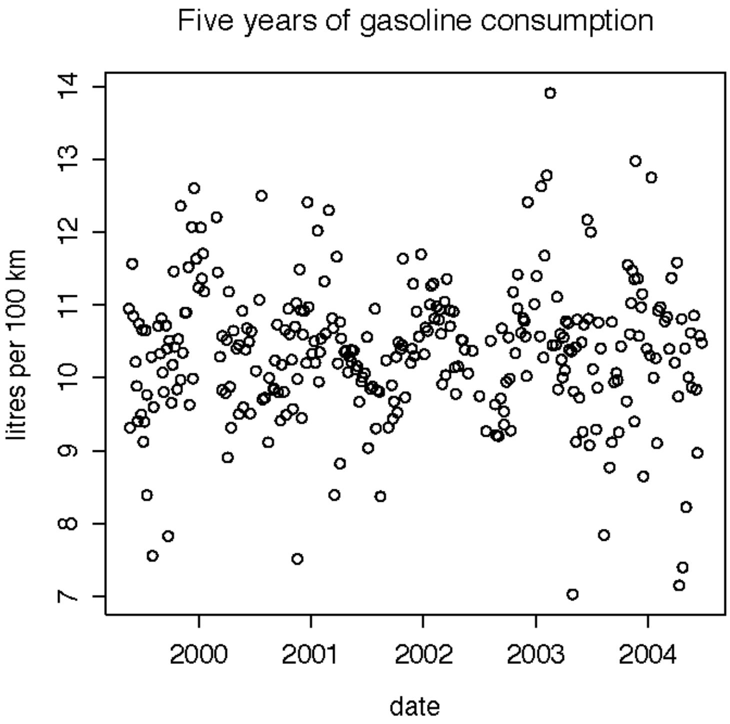
Some Controversies of Statistics Education and Practice

In a young discipline like statistics, debate over the basic tenets is inevitable. It is not usual in statistics to include these debates as part of our course content – the subject is already confusing enough without questioning the foundations. However, there is a need for faculty to discuss the various points of view and the new developments in the subject so that an appropriate balance is included in our panel of required courses. This presentation includes some suggestions for moving away from the framework presented in the textbooks we currently use. The ensuing discussion may help to indicate whether or not changes in our courses are advisable.

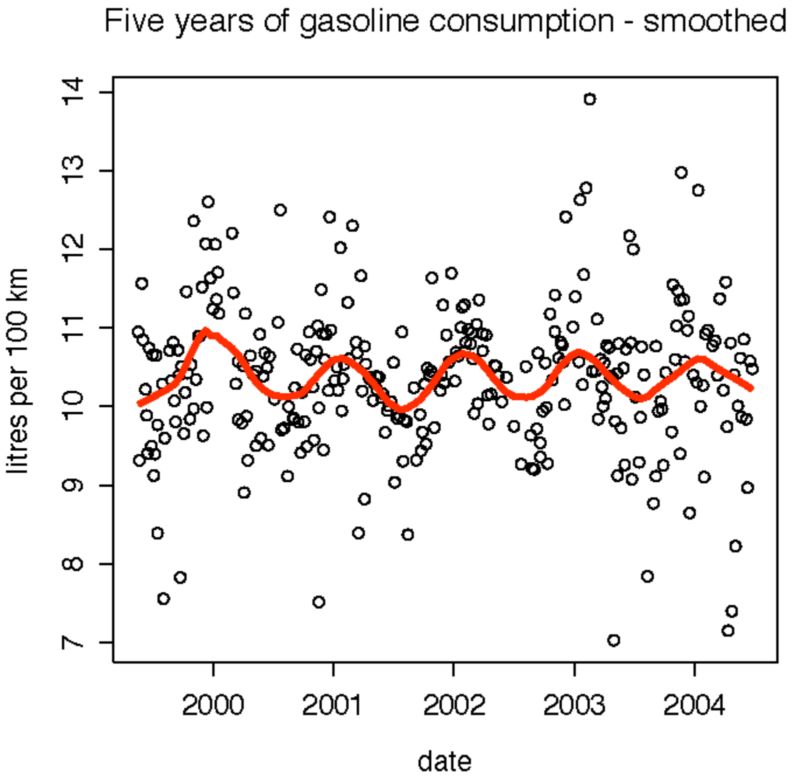
The controversies will be described under three headings:
1. the role of parametric inference
2. the practice of statistics
3. problems of pedagogy
I suggest a decreased role for parametric inference, outline how we might provide an improved service to statistical practitioners, and list some ways to improve the effectiveness of our courses. My objective is to invite criticism and encourage debate!

1. The role of parametric inference – do we emphasize it too much?

Let me start with an example:
Ex. 1 : Monitoring Fuel Consumption. Here is a time series for 1999-2004.



It is typical of data collected to monitor a process: for example, such data sets are collected by traffic engineers, by paper producers, and even by little retail stores, often represented a routine measurement. My purpose was to try to detect any engine problems with my car by noting the gasoline consumption at each fill-up. I was also interested in anything I could learn from the data about my fuel consumption. For example if consumption per km increased suddenly, I might seek a remedy. However, like many such series, there is a lot of noise in the system and I need to have some way of amplifying the signal. Since I have no idea what parametric form the series might follow, other than perhaps y=c, I will smooth the data:



This is clearly an improvement on the y=ybar line. In colder climates, the seasonal effect is more pronounced and more valuable for the monitoring task.

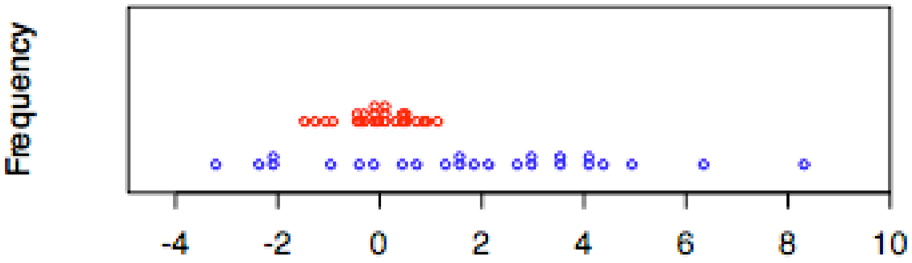
Some statisticians would be uncomfortable with this analysis: what is the model? My answer is that there is no need for a parametric model and in fact it is unlikely to produce the information obvious from the smooth. The graphical analysis completely satisfies my needs from this data.

What would our students do? Would they fit a model, study the residual plot, and test the fit of the model? Would they learn anything about the data?

This loess smooth provides a basis for properly detecting any future problems with the car.

Ex 2: Asessing a new drug:

Another example where a parametric approach is more complex than a not-parametric approach.



Imagine the red data is the response of the standard drug, the blue data is the response of the proposed new drug, and a large value is a good response. The study has been badly designed and the 25 in each sample are not paired – suppose they are independent samples.

What do you conclude about the relative effectiveness of the new drug? It is obvious that the new drug is more helpful that the old drug for a majority of people, but is worse for a small subgroup. It is not so easy to propose a hypothesis test that would test whether this result is reproducible, even though it seems clear that it is reproducible. The issue is not one of normality nor of homoscedasticity, but rather that the parametric modeling in this case is irrelevant to the information sought. Isn't the graphical analysis enough?

What would our students do here? t-test regardless? Often (40-60% of the time) it will be non-significant, as one can show using simulation. Would they first test for the variance difference? The conclusion is relevant but only a small part of the story here.

Our students might well think of fitting two normal distributions to the two samples. But would they then use simulation to examine the stability of the outcome? This might be a reasonable approach and would make use of our probability modeling instruction. This underlines the importance of probability modeling for statistical inference.

I suggest the following: Students need to be taught the effectiveness of simulation and graphics for data analysis, and also the importance of probability modeling for simulation.

Regression Modeling:

One aspect of parametric modeling is the estimation of functions f() in y=f(x) + e. Linear models figure very prominently in this activity. There are at least two distinct uses of such models – one is descriptive, where we want to compare the relative importance of various possible factors in predicting y, or possibly we simply want to predict y from x. The other use is analytical, where we wish to check a theory about the functional form of links between x and y. For the descriptive mode, the functional form is not so important since we usually are content with locally linear fits. Parsimony is a guiding principle in this, although we also aim at minimizing the prediction error. We want a simple way to describe the relationship f(). But for the analytical mode, we want to get the right function – not just a local relationship described by tangent approximations, but a relationship that corresponds to a causal theory. We would sacrifice predictive precision in favor of good estimates of our causal parameters.

In statistics education we concentrate on the descriptive modeling. But in this case estimation of the parameter values is not so critical: these parameters are not measuring anything physical. In other words, in the context we most often present regression models, inference concerning the model parameters is not the main interest. Rather it is how well the model enables the value of y to be predicted from x (over the range of x in the data). My suggestion is that a nonparametric smooth could more easily provide this information – residuals from a nonparametric smooth have the same role as residuals from a parametric smooth. We should give more emphasis to the methods and uses of nonparametric smoothing.

Unbiasedness:

To seek a method that is right on average. Like the proverbial statistician who feels comfortable on a blistering hot day by standing in a pool of ice water. It is surprising that this criterion has become so ingrained in the stat theory toolkit. Unbiasedness sounds like it is a good thing, and for many purposes it is. But as a criterion for an estimator, it does not seem appropriate. In estimation of a parameter, we usually hope the estimate is close to the population parameter value in a particular instance, and whether it would be exactly right on average over many such estimation events is of little relevance.

How many of our students understand the limitations of the unbiasedness estimation criterion?

Analysis of Variance:

A very nice feature of analysis of variance is the ability to separate the attribution of total variation to two or more sources of variation. However, there is a problem with using variance as a measure of variability: it is in squared units. So if 50% of the variance in Y is attributable to X, and 50% to error, how much have we improved our precision of estimating Y by measuring X? In this case $r^2 = .5$. The precision of estimate of Y, in a simple case, is $\sqrt{1-r^2} * \sigma_Y$ and the reduction from σ_Y is by the factor $\sqrt{1-r^2}$ = .71 or in other words only a 29% reduction. This is a lot less than the usual report of a 50% reduction in variance. In squared units we have this nice additivity of variance (under certain conditions) but this simplicity is partly an illusion since the units are not the units of interest to the investigator. Analysis of variance is a computational tool that, vital in the days of hand calculations, seems a diversion from the relevant outcome of the analysis. When students hand in assignments that included tables of sums of squares, it is apparent that they think these sums of squares have some useful interpretation shedding light on the data. Do students realize that it is an anachronism of pre-computer days, and totally irrelevant to the interpretation of the analysis?

A common response to these criticisms is "how would you do it better"? My answer is that I would put less emphasis on parametric fitting of data, on estimating regression coefficients, on unbiased estimation, and I would never use variance as a measure of variation. I would replace the space created by these omissions by giving more emphasis to nonparametric smoothing, graphics, resampling, and simulation. In fact that is exactly what I do in STAT 400 (Data Analysis). Is it enough to leave these topics to STAT 400? Well, STAT 400 is not a compulsory course, and even if it were, it seems a shame to leave these simple issues to the final year of undergraduate studies.

Least Squares

and outliers ....

The Practice of Statistics

Another reason to reconsider our course content is that the changing nature of statistical practice, and of the cadre of workers who need to use or understand our methods.

Often it is said we are educating students rather than training them for particular jobs. However, we have tailored our courses to meet certain educational needs partly to maintain control of teaching in our discipline and partly to cater to the perceived needs of various disciplines. The specialization has been partly geared to particular application areas: life sciences, social sciences, natural sciences and engineering. Another partition of the students in our course might be as follows:

INTRO: There are some who need to know what statistics is, and what the big ideas in it are, but have no desire to actually analyze data
SERVICE: There are some who need to be familiar with statistical concepts and tools for use in their field of study, but have no desire in inventing or adapting solutions to new situations
AMATEUR: There are those who need to understand the concepts and tools well enough to perceive new opportunities for inventing or adapting solutions to new situations they meet in their own work, but may not have an interest is developing tools for others
EXPERT: There are those who want a thorough introduction to the basic concepts and tools so that they can involve themselves in research for transmission to others

Do these subgroups need different courses, or merely a different number of courses? We currently have

INTRO: STAT 100
SERVICE: STAT 101, 201, 203, 302, 403
AMATEUR & EXPERT: All the rest.

My question here is: Do our majors/honors need the exposure to the content of the INTRO and SERVICE courses? Will they be able to communicate effectively with co-workers if they do not have the exposure to the questions of applied statistics? STAT 300 should help, but perhaps our major/honors streams should include a bit more applied stats at the lower division levels.

Imagine a list of students who have ever taken a stats course as part of their recent undergraduate degree. Consider partitioning the list according to the highest level reached in statistics: 000, 100, 200, 300, 400

The number at each level might be 1000, 1000, 600, 325, 75 based on a graduating class of 3000 students. In other words only about 2.5% of students become minimally competent in statistics before graduating and only 75/2000 or about 4% of those taking any stats course continue to the highest levels. Our market share is quite small, however assessed. Is it true that only a small percentage of students need to be competent in statistics? Is the competence we inculcate in students so little needed by graduates? Is our target too narrow?

I think most of us would agree that for students to be expert enough to teach, they need at least a MSc and then probably some work experience as well. If that is the case, we should not be aiming to prepare "experts" in statistics at the BSc level – leave that for graduate school. Or, to see it another way, perhaps we should be producing students who understand the basic ideas and can research the details when they need to. Might a change in emphasis allow us to attract more students, each with a broader range of skills for life?

Let me be more specific about the kinds of topics that could receive greater or less emphasis than is currently done:

Decision Making vs Significance Testing

We do not teach students how to make decisions based on data. Significance testing does not do that since it ignores priors and utility functions. Do our students appreciate the components necessary to make an optimal decision? Business statistics courses do try to include decision trees, but where does this appear in our courses? It is true that the scientific context decisions are not so important since data-based findings, when they are really new, are usually not widely accepted until reproduced by several investigators in several locations. However, most students will not be employed in a scientific context. A 2004 follow up survey of SFU graduates showed 13% in science and medicine, but 63 % in business and social science. The basics of data-based decision-making is an important topic. We teach it in STAT 460 but that is an optional course in our major program.

One feature of hypothesis testing that reveals its failure as a decision making technique is the arbitrariness of the Type I error specification. Different alphas can give different "decisions". The suggestion of using the observed p-value, instead of specifying a critical value, does not solve this problem since the observed p-value is usually compared with an arbitrary alpha. This whole procedure seems more formal than the establishment of an index of credibility would warrant. The crispness of the result of a hypothesis test seems to have little to do with the genuine equivocation required by a marginal result. In practice, users are more relaxed about interpretation of p-values than the dogma would suggest. We need to build this realism into our presentation of the subject. Would our students be compelled to report p=.06 as Not Significant? Do they think p=.001 very significant (in spite of possible study shortcomings)?

Designed Data Collection vs Data Mining

Our courses focus on inference in the case of sample data, either inference to a hypothetical population (as in a measurement situation) or to a concrete population (as in surveys or environmental sampling). But it must be a majority of statistical analyses that are based on data that has been collected for some routine purpose, rather than a specific research goal: for example, census data, weather data, stock market data, financial transaction data, etc. These data tend to be time series, but not always: in the case of credit card transaction information, the time trend of the data is not as important as the relationship among the variables (e.g. Do residents of West Vancouver shop for cars in West Vancouver?). We do not say much about time series in our courses. We don't talk much about the study ot relationships among variables in data sets that are not random samples from any population of interest. The art and science of analyzing data that has been collected for some other purpose, often called data mining, is a topic deserving more attention. Do our students know when a convenience "sample" might contain valuable information, and do they know how to extract the information?

Graphical Communication

There are a references in our courses to graphical methods. Many courses start with a reminder about histograms, box-plots and scatterplots, and we also present probability plots, residual plots, and effect plots. But these are mostly for analysis purposes: checking assumptions, suggesting models, etc. We seldom propose to use plots to communicate results of data-based studies. The tradition in statistics is to treat graphical communication as something only for a lay audience or for pre-school exercises! It seems hard to convince researchers that a good graphical display can make clear some complex research outcomes. A graphical display may be the simplest way to report a result. With nonparametric smoothing, it is sometimes the only way. Do our students understand that a graphical display can sometimes be the best way to convey study results? Will they feel that the summary is incomplete without a parametric test of significance or parameter estimate?

Optimization vs Gradual Improvement

The SPC movement introduced the idea of gradual improvement as a more practical alternative to optimization. Complex systems cane be studied in two ways: construct a simple model of a real-life process and optimize it, or, modify the real-life process in a way that should improve it. While the latter approach takes longer, it may lead to better results. The classical work of Box with his EVOP was one early proposal of this approach. Deming's QC and related strategies were another. These post-war efforts have not had much impact on our statistics curriculum. Is this a technology that we need to give more emphasis to in our courses?

In this section the suggestion has been that we need to provide more emphasis to data-based decision-making, data mining, graphical communication and gradual improvement techniques. This increase would be offset by a reduction in the attention given to significance testing, designed data collection, parametric numerical summary, and optimization methods. Moreover, an argument has been made to prepare students less for graduate work and more for statistical practice.

Pedagogy

A philosophical shift from the math-based teaching of statistics to the data-based teaching of statistics requires a change in teaching methods. Instead of being concerned that students will not bridge the gap from math to applications, we would instead have a concern that students would not bridge the gap from data to concepts. We outline a few suggestions for how to adapt to this change in philosophy.

Case Studies vs Logical Progression

Compare a series of lectures in wildlife population estimation, sports team quality, accident-free driving survival, randomness in the stock market, junk mail filtering, ... with a course in one-variable tabular and graphical distribution summary, two-variable relationships, types of variables, sampling distribution of the mean, confidence intervals, regression analysis,..... Which course sounds more attractive? The case study course would be selected by most students on the basis of interest. The pedagogic question is whether the students would derive the helpful logical structure of statistics theory from the case study approach. I changed my mind about this a few years ago and now believe that the motivation feature is more important than the advantage of a logical progression, and that students do indeed see the logical structure once they understand the case studies. Some evidence for this last claim was derived from a "minute paper" exercise done a few times in my STAT 100 course in 2002. For anyone primarily trained in mathematics, this finding is hard to accept, and indeed I disagreed with it for many years. But I have become a convert. I would urge faculty to consider this approach if they have not already tried it.

Tests and Examinations

Statistical software is used in virtually all statistical analysis. But our tests and exams usually do not allow students access to software. This is unfortunate. Of course, there is more to learning statistics than learning how to analyze data: data collection, concepts of analysis, interpretation of results, communication of results, are all additional skills that need to be learned and tested. Do our tests and exams assess these skills? Do we ask students how to design a survey, whether, for example, they understand signal amplification, whether they appreciate that all models are wrong, whether they can put results in clear English for a technical or a lay audience? The fact that it is hard to test these things is well-known. If faculty do not devote adequate time to constructing effective tests and exams, they will be tempted to use ready-made "calculate" questions that can be done by hand. But what is the incentive to spend this extra time? Might the department have a review process that reviews tests and exams? This would provide data helpful in assessing faculty teaching, would provide an incentive for faculty to spend the time to make the exam effective, and would provide students with an incentive to obtain a more complete education in statistics.

Use of Common Sense:

Many of the concepts we teach are very sophisticated – students can learn how to use various methods without fully understanding the whys and whens. In order to pass the exams, the student accepts the dogma provided without question. We must encourage students to say "That seems unreasonable to me for the following reason. Please explain why it is nevertheless accepted practice". In obtaining the answer to such questions, the student

is better able to answer the same question when they are asked it by clients or co-workers

is aware of the particular context in which the method applies

is motivated to consider how to overcome the limitations of the technique

Of course, it is commonly said that the best way to obtain a deep understanding of a statistical strategy is to teach it to others. So while engaging in this kind of dialogue, the faculty member learns as well. The seminar format is not very common in statistics instruction, even at the graduate level. This may account for the impression sometimes conveyed to students that the "proper" statistical procedures have all been invented and an application merely requires consultation to the proper published authority on the subject. The suspension of common sense sometimes that is fostered by this impression often leads to failed analyses, since the conditions of a widely accepted approach are so rarely present.

One antidote to the failure of students to use their general intelligence for solving statistics problem is to require students to verbalize the strategies they learn. This is not to require students to perfect their English – even verbalization in their native language will help. Most people are comfortable with expressing their thoughts in their native language, and so expressing statistical strategies this way helps to make the strategies part of their general mode of thinking. Students who learn strategies by remembering their symbolic (or mathematical) form do not have this native tongue advantage.

Summary

I have argued that current courses in statistics give too much emphasis to parametric inference, and not enough to nonparametric approaches, graphics and simulation. I have suggested that the practice of statistics already requires a broader understanding of the discipline than they get from our courses. Finally, I have suggested some pedagogical issues that I feel would help to encourage students to better prepare themselves for effective use of statistical strategies.

If my suggestions seem ill-advised, then I encourage faculty and students to be openly critical of them. The ensuing discussion will hopefully take us all to a better place.

Some References:

Weldon, K.L. (2005) Less Parametric Methods in Statistics. Metodoloski zveszki 2:95-108.

Weldon, K.L. (2005) From Data to Graphs to Words - but Where are the Models? Proceedings of the ISI/IASE Satellite Conference in Sydney, Australia, April 2005.

Weldon, K.L. (2005) Modern introductory statistics using simulation and data analysis. Proceedings 55th Session of ISI, April 2005, Sydney, Australia.