ADVANCED TOPICS FOR A FIRST SERVICE COURSE IN STATISTICS

K. L. Weldon
Department of Statistics and Actuarial Science
Simon Fraser University
Vancouver, Canada. V5A 1S6

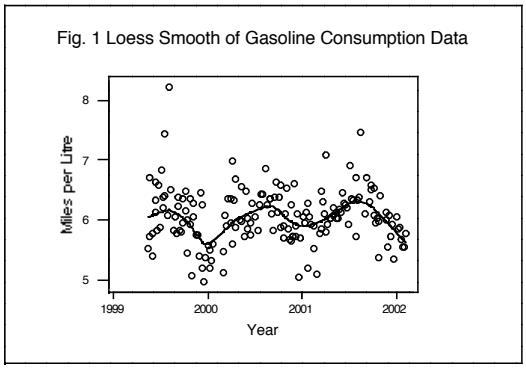
Statistical software has made traditional statistical calculations accessible to almost anyone, but it has also stimulated new methods that are usually reserved for advanced courses. In this paper I argue for inclusion in the first course of non-parametric smoothing, density estimation, coplots, simulation, the bootstrap, time series forecasting, and plots of multivariate data. It is argued that the logic underlying these techniques is simpler and more useful than the logic underlying the inference usually included in first service courses in statistics.

Recent popular texts aimed at the first service course in statistics reflect the development high level statistical software. Hand calculations are de-emphasized, large data sets are included, and graphical methods of analysis and display are given more prominence. (See for example, Moore and McCabe (1993), Freedman, Pisani and Purves(1998), and Wild and Seber(2000). However the frontiers of statistics have also been impacted by the advent of computing power, and some of these frontiers have produced very useful and conceptually simple methods. Because these methods have been developed by statistical researchers, they have tended to be reserved for senior undergraduate or graduate courses, along with more complex material. We discuss the merit and the feasibility of bringing the simpler methods into the first service course.

1. NONPARAMETRIC SMOOTHING

The adjective "nonparametric" would be unnecessary except for the historical fact that statistical theory was mostly parametric. The parametric model was needed to simplify the communication of statistical results. Graphical methods were too time consuming to prepare for all but the smallest data sets. A simple regression analysis could be summarized by reporting an estimated intercept, slope, and residual standard deviation. However, the ease of graphical displays with statistical software has made parametric models less crucial. Consider the following example: gasoline mileage was recorded at every fill-up over a 33 month period, and the interest is to study the pattern over this period, to help in the assessment of future readings. The data was collected in Vancouver, Canada for a 1986 Mercedes 190E.

The smooth curve (Fig. 1) which clearly shows the seasonal effect is generated by the lowess procedure in MINITAB, which is described in detail in Cleveland (1993). No clever use of sinusoidal functions or time series modeling is required. A smoothing parameter must be chosen but trial and error is an adequate approach - in this case the use of .1, .2, and .3 was enough to show that .2 was a reasonable choice to reveal the form of the anticipated seasonal effect.



The details of the lowess procedure can be explained fully in a first course if desired, provided simple linear regression is included in the course - only the least squares fit of a line needs to be

covered, not the inference associated with the regression. Then lowess can be described as the result of simple linear regression on each of a grid of values using the points
close to the grid values. The refinement of weighting the regression can be added once the basic idea is conveyed. The smoothing constant is the proportion of data values used for the fit at each grid point. The "smooth curve" is really a series of line segments joining the fits at the grid points.

The result of the lowess smoothing procedure can be conveyed graphically or in a table if numeric table if required. The lack of an explicit function might be a disadvantage in some applications - for example extrapolation. However, the time and expertise required to fit a parametric function to this data would have limited returns. The lowess would already send the enterprising data analyst to sources of information on weather and traffic over the period, for a more detailed analysis.

The utility of a procedure like lowess should not be underestimated. Exploratory statistical methods are generally aimed detecting signals that are hidden by noisy data. Lowess does this is way that is applicable to a very wide range of situations - those in which a response is modeled as a function of an independent variable. It is logically as simple as simple linear regression, and very useful for anyone who has learned the basics of a statistical software package. It should be included in a first service course.

2. DENSITY ESTIMATION

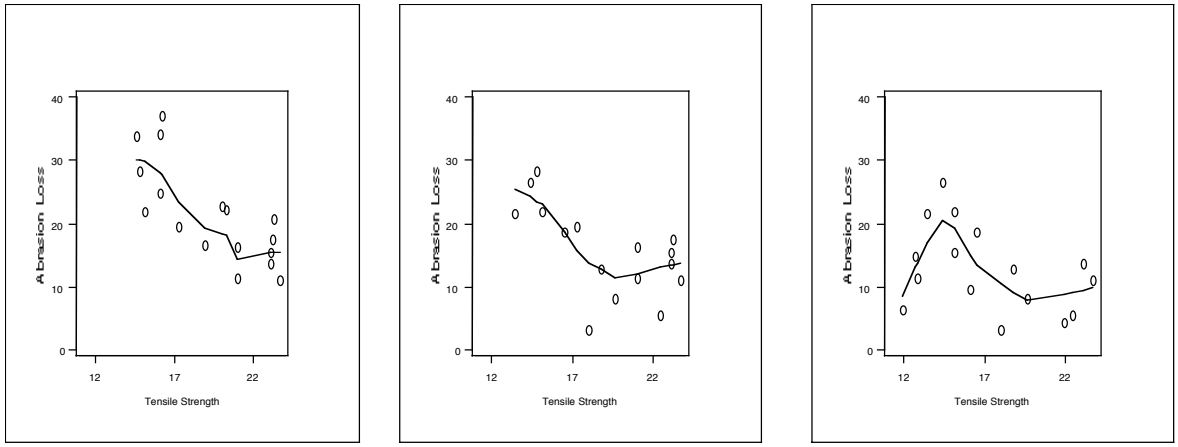
This is another technology steeped in higher mathematics that actually has a simple and useful aspect to it. We spend too much time talking about how to construct histograms: number of bins, location of bins, and treatment of unequal bin sizes. All these troublesome details disappear when one uses the approach of estimating frequency on a grid of values. As a first step, have students count the number of values within +- d of each of say, 20 grid points. Clearly the choice of d is similar to the choice of the smoothing constant except that its size is not limited to the range (0,1). These counts are pseudo-frequencies that can easily be turned into relative frequency of even density if desired, but of course this last step will not change the shape of the graphical representation. A default value of d might be 1/15 of the range of values, or some convenient rounding of this value. Fig. 2 is the result of using this method for a set of 36 age-at-death of the U.S. presidents.

Contrast this with the default histogram of the same data. The histogram is neater, and easier to compute by hand, but the window approach used here is more revealing and no more effort when using a computer program. Of course refinements are possible - for example, the rectangular window could be replaced by a triangular window, but the advantage does not seem worthwhile at the elementary level. This "windowgram" does not require any decisions to be made by the user. Moreover, the logic of the windowgram is not much more difficult than the histogram - one just counts up the number of data values close to each grid point, and plots this on the vertical axis at the grid point. There is only the slight complication of explaining the vertical scale so that the percents add to 100. The definition of "close" can be chosen by the student to observe its effect, or can be left to the software.

3. COPLOT

Cleveland (1993) describes coplots and shows them as very useful is analyzing more than two variables. The first service course usually limits the discussion to two variables partly because of the complication of displaying phenomena in three or more variables, but also because of the conceptual complexity of fitting surfaces. The Coplot (short for Conditioning Plot) is a powerful technique for examining relationships among several variables, and for three variables is conceptually very simple. The basic idea is to present a sequence of scatter plots of two variables in which the third variable advances through its range. This is particularly effective when combined with the lowess technique just described. Consider the following example which is based on the example in Cleveland's book. The three variables in this example are a dependent variable Abrasion Loss and two independent variables, Tensile Strength and Hardness. These measurements relate to the material loss in 30 specimens of rubber caused by a certain amount of rubbing of the specimens. The sequence of graphs (Fig. 3) reflects increasing hardness: from left to right, the graphs reflect the effect on the Abrasion Loss - Tensile Strength relationship as the Hardness increases. More specifically, each graph shows the data associated with the 17 smallest values of hardness, 16 middling values, and the 16 highest values of hardness. The three data sets overlap approximately 50 percent - this overlap percentage determines the number of graphs.

Fig. 3 Coplots of Abrasion Loss vs Tensile Strength Given Hardness



The graphs reveal both a main effect of hardness on abrasion loss, and an interaction effect with tensile strength on abrasion loss. The coplots for abrasion loss on hardness for increasing tensile strength would show the interaction from another point of view. The information in these plots is not accessible from regression methods or from other plotting methods such as spinning the three-dimensional data scatter. An appreciation of scatter plots is all that is required to understand the coplot technique - the computer can work out the graphing details. The addition of loess to the coplot is helpful but not an essential element of the coplot.

4. SIMULATION

The role of simulation in helping to solve hard problems in both theoretical and applied statistics is well-known amongst statisticians, but many students never reach the courses in which it is used. The idea of having a computer mimic randomness and inferring useful information about random phenomena can be conveyed in a simple context. However, before computer simulations are used, a physical simulation needs to be demonstrated to ensure that students distinguish simulation from calculation.

Many instructors use simulation to demonstrate the relative stability of means. The details will be omitted here except to say that the same demonstrations are useful in a first service course. Moreover, the idea of simulation itself is very important. It is a very general way to explore the consequences of randomness, and this is something that can be done without a background in mathematics, and without much experience with computing.

5. THE BOOTSTRAP

This technique was named and promoted by Efron (1979), and has become a very popular way to estimate variability of a statistic that would otherwise be intractable. It appears in research papers and advanced courses, but the idea is quite simple and useful and could easily be added to an early statistics service course. It is important that students have been exposed to the idea of simulation for a painless introduction to the bootstrap. The first example will illustrate how the technique might be used to introduce and explain the relative stability of averages, and the second example will show how it might be used to solve an intractable estimation problem in a real-data context.

EXAMPLE 1 THE BOOTSTRAP APPROACH TO THE SEM

While students eventually internalize the idea embodied in the "standard error" formula $\sigma_{\bar{x}} = \sigma_x / \sqrt{n}$, initially there is usually quite a bit of confusion. The simulation example mentioned in the previous section should help, but students may rightly ask "What if the population is completely unknown, and I am interested in something other than the mean, then how do I judge the accuracy of the sample information?" The bootstrap provides a way to answer this.

Consider again the artificial population of integers {0,1,2,…,500). A random sample of size 25 from this population allows one to compute the sample mean. For example,

335 214 57 35 243 32 497 111 270 32 495 294 471 484 169
163 9 389 267 147 204 463 29 205 21

is such a sample and its mean is 225.4. If the population were unknown we would estimate the population mean to be 225.4. But can we assess the accuracy of this estimate based on the sample information alone?

One approach to suggest is to treat the sample as if it were the population, and take samples of size 25 with replacement from the original sample of size 25. For each of these samples, compute the new sample mean, (rounded to an integer in this case), and from all the samples taken calculate the standard deviation of the sample means. This standard deviation gives an idea of the precision of the original estimate, 225.4. The result of 100 re-samples is that the sample mean has a standard deviation of 32.8. This gives an answer to the question about the precision of the estimate 225.4. The estimate could be stated as 225.4 ± 33 Now compare this to the theoretical standard deviation of the sample mean = $\sigma_x / \sqrt{n} = 167.3 / \sqrt{25}$ = 33.4, and so with the standard theory behind us and the estimate of the population standard deviation = 167.3 we would report 225.4 ± 33. However we computed this same precision without using the standard theory, but only the first principles approach of re-sampling. The re-sampling method is not always exactly the same, and of course both methods are only estimates of the true accuracy based on the true standard deviation of this population, which is 28.9 or ± 29. However re-sampling gives reasonable answers in practical situations, and is easy to use and explain.

This example shows that the bootstrap can give information about variability of an estimate that is similar to that provided by standard sampling theory. The next example shows the bootstrap at work when no standard sampling theory is available.

EXAMPLE 2 BOOTSTRAPPING AN INTRACTABLE ESTIMATION PROBLEM.

The bootstrap method of computing the accuracy of a statistic applies to any statistic, no matter how complex. Consider the following example: From the following data set of height and weight of 25 men under 40 years of age, we have computed the so-called body mass index (BMI) which is $= WGT / \sqrt{HGT}$. We want to estimate the value which would be greater than 90 percent of the mean in this age group, as possible use for clinical evaluation of excessive weight.

The data for the BMI looks like this:

16.9 17.2 17.5 17.8 18.4 18.5 18.5 18.8 19.0 19.8 20.0 20.2 20.7 21.4
21.8 22.1 22.5 22.9 23.1 24.3 24.4 24.8 25.4 28.0 28.5

And so the 90th percentile might be estimated as half-way between 25.4 and 28.0 or 26.7. A concern however is that the sample is too small to provide a reliable estimate of the 90th percentile, and so we estimate the precision of this estimate with the bootstrap. The result of a thousand re-samples of the BMI data shows that the 90th percentile estimate we have used to arrive at 26.7 has a standard deviation of about 3.1. This suggests our estimated 90th percentile could be easily be the 76th percentile up to the 99th percentile, or even more in error, and so would be inadequate for clinical use. The sampling theory of the 90th percentile is not simple, and neither is the appropriate model for the ratio we have used as the BMI, and yet we have been able to determine, with the easily-explained bootstrap, the variability of our estimate for this particular sampling situation.

The point here is that we were able to assess the variability of the estimated parameter without knowing a parametric model for the data or any theory about the variability of percentiles. All that was needed was the bootstrap technique. The bootstrap is another example of a simple technique that we could and should include in our early courses, particularly service courses.

6. TIME SERIES FORECASTING

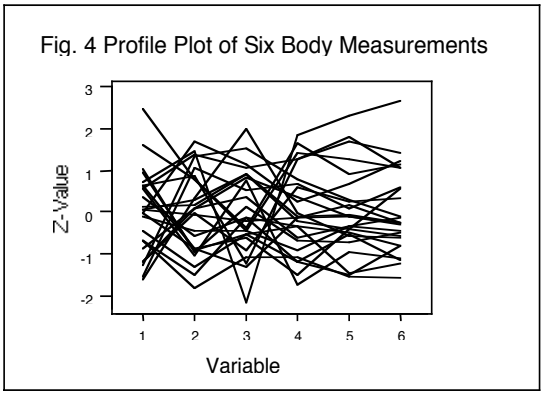
Time series methods are taught either as advanced mathematically-based statistics courses, involving ARIMA modeling or spectral theory, or as ad-hoc techniques specific to an financial or industrial forecasting. Undergraduate majors in business may be required to have some exposure to time series fitting and forecasting methods. However, the need to understand time series data analysis is important in almost any field which uses data at all. Chemists, oceanographers, environmentalists, and even medical doctors need to know about the opportunities and hazards of using time series data.

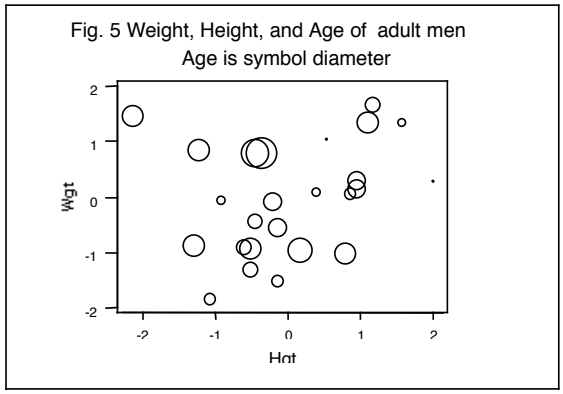
A few textbooks aimed at the general first service course do include a chapter on time series (Weldon(1986), Griffiths et al (1998)) outlining elementary methods of smoothing, trend and seasonal extraction, residual examination, and forecasting, but many popular books do not (Moore and McCabe (1993), Freedman, Pisani and Purves(1998), and Wild and Seber(2000)). Suffice it to say here that the gas consumption data described earlier in this paper is an example of a time series for which the analysis using elementary techniques like lowess produces useful information. In fact, there is an interesting question with this data of where the seasonal minimum occurs, which in this particular case is crucial for producing a useful forecast of the next few time points. It turned out to be right after the last collected data value: the subsequent values of that time series were 5.7, 5.8, 5.7, 5.9, 6.3.

The treatment of time series as a topic that is supplementary to the mainstream gainsays its ubiquitous appearance in almost all data-based disciplines. While advanced time series is beyond the grasp of those who do not have a solid mathematical background, there are many useful ideas that can be conveyed at the elementary level.

7. PLOTS OF MULTIVARIATE DATA

Once students get used to scatter diagrams, they realize their power in helping to analyze bivariate quantitative data. At the same time, they will realize that many practical data sets will involve more than two quantitative variables. While multivariate analysis has formidable mathematical problems associated with it, some very simple graphical strategies are possible to convey in a first course. An obvious one is the profile plot. Six body measurements on 25 men are displayed in the profile plot Fig 4. - the six variables are Age, Weight, Height, Neck, Chest and Abdomen.





The Profile plot is not a useful summary plot but is very useful for data analysis: detecting coherent subsets, outliers and correlations among several variables.

An even simpler method available for a small number of variables is to use an "augmented scatter plot". The simplest example of this is to have a third variable X_3 determine the size of the circular ring placed at (X_1, X_2). The Age, Weight, Height data looks like Fig 5.

Another primitive plotting method is the star plot, in which the profile lines are placed on spokes so that the profile plot looks a bit like a star. These starts must be placed in a sequence. There are various ways to make these plots more "viewer-friendly" such as choosing the order of the variables, and for the stars, choosing the order of the starts themselves, but these are optional enhancements.

Use of these plotting methods in an early service course may be one way to alert students to the very common multivariate nature of real-life data. Also, it gives an appropriate emphasis to graphical methods, which is surely one of the most popular methods of data analysis.

CONCLUSION

Several methods that are very useful for practical data analysis have been proposed for addition to the first service course in statistics. It has been argued that these methods all have a simple conceptual basis, with very little prerequisite knowledge required. Given the extreme difficulty experienced by students to understand the standard methods of inference (Lipson (2000)), my suggestion is to replace some of the time spent on inference with time spent on these items of data analysis.

REFERENCES

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

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

Griffiths,D, Stirling, W.D, and Weldon, K.L. (1998) *Understanding Data: Principles and Practice of Statistics*. Wiley. NY.

Lipson, K. (2000) The Role of the Sampling Distribution in Developing Understanding of Statistical Inference. PhD Dissertation. Department of Statistics. Swinburne University. Melbourne.

Moore, D.S. and McCabe, G.P. (1999) *Introduction to the Practice of Statistics.* Third Edition. Freeman. NY.

Weldon, K.L. (1986) *Statistics: A Conceptual Approach*. Prentice-Hall. Englewood Cliffs, NJ.

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