Some under-used, but simple and useful, data analysis techniques

K. L. (Larry) Weldon
Dept. Statistics and Actuarial Science
Simon Fraser University
Canada

Introduction

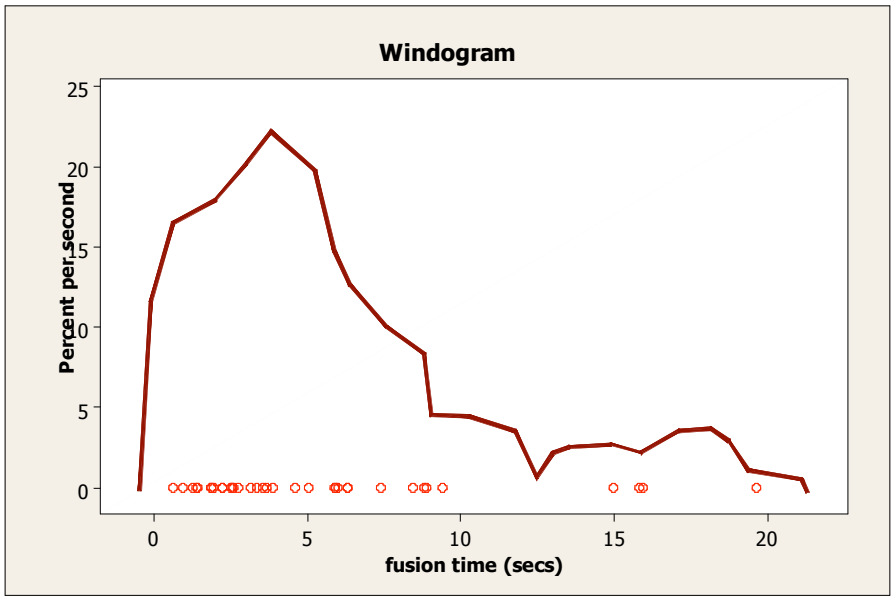
Resampling and graphical methods have experienced remarkable growth in both the theory and practice of statistics, largely as a result of advances in computer speeds and software ease-of-use. However, textbooks that support introductory courses in statistics tend not to emphasize these methods; moreover, the courses themselves often stay with the textbook-based syllabus. Advanced courses in time series, multivariate analysis, and often do bring these computer-based topics to advanced students of statistics. But data is often analysed by researchers with only introductory statistics training. To serve these people, there are some modern techniques of data analysis that could easily be included in early courses in statistics. In this paper I will describe some simple approaches to kernel estimation of densities and nonparametric smoothing, to multivariate data display, and to the use of the bootstrap as a general-purpose inference procedure. It is argued that these primitive versions of the modern techniques are as useful as the traditional topics usually taught in an introductory course, and easier for students to understand. One consequence of an increased emphasis of these techniques might be more widespread use of statistical science.

Statistical theoreticians strive for optimal procedures. A reminder of what is optimal in practice is embodied in the definition of "Eisenhart Efficiency" (quoted in Pregibon and DuMouchel(2001)): statistical efficiency x probability that it is used, if appropriate x probability that it is used correctly, if used. Procedures that are sometimes considered sub-optimal may be optimal in a larger context. Sometimes simplicity is underrated in academia, and narrow efficiency may be overrated.

Kernel Estimation

First we consider a variant of the histogram that we can call the "windowgram" to distinguish it from the traditional histogram, although it is a very close relative. Consider the data provided by the Cleveland (1993, Chapter 2, pp 42-67 and pp 82-85, the VV group) on the "fusion time" for subjects examining a stereogram. First define an equally-spaced grid across the range of the data. Then for each grid value, scan the entire data set noting the distance of each data point from the grid point. For some arbitrary distance d (such as a tenth of the range), count up the number of data values within d of the grid point, and plot that value at the grid point. Note this is a weighted sum where the weights are 1 or 0. Do this for each grid point and then plot the result. Except for the vertical scale, the resulting graph would look like:

Figure 1: Frequency Distribution based on primitive kernel density estimation



This crude kernel density estimator can be scaled to percent per year as shown, if desired, so that area = 100%. The obvious relationship between a weighted sum and a weighted average can be explained when the scaling is important. Also, a good default value for the half-width of the kernel is $Range/\sqrt{n}$. However neither the scaling nor the default value of d need be understood to make use of this alternative to the traditional histogram. Trial and error can be used to choose a reasonable value of d – even the naïve user will realize that some local deviations are likely due to random variation.

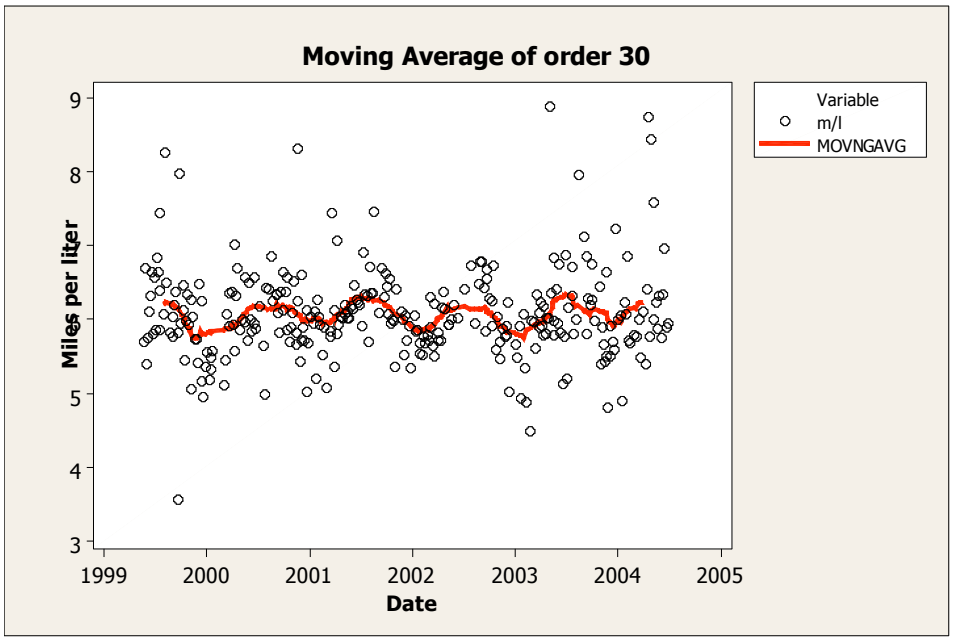
What has been shown is that a density estimator that is really proportional to a sum of local frequencies of data values. This not only makes the density idea more transparent, it provides an easy-to-explain way to portray the distribution graphically.

A secondary use of this approach is that the local-frequency-count idea can be re-expressed as a rectangular smoothing window. The modification of the rectangular window to something more gradual is a natural step. For example a trapezoidal-shaped window, or a Gaussian window, or even a tricubic window (Cleveland(1993)) could be motivated as this extension. The point of these extensions would simply be to obtain more smoothing with less bias – the particular kernel used need not be presented algebraically for a novice audience.

Nonparametric Smoothing

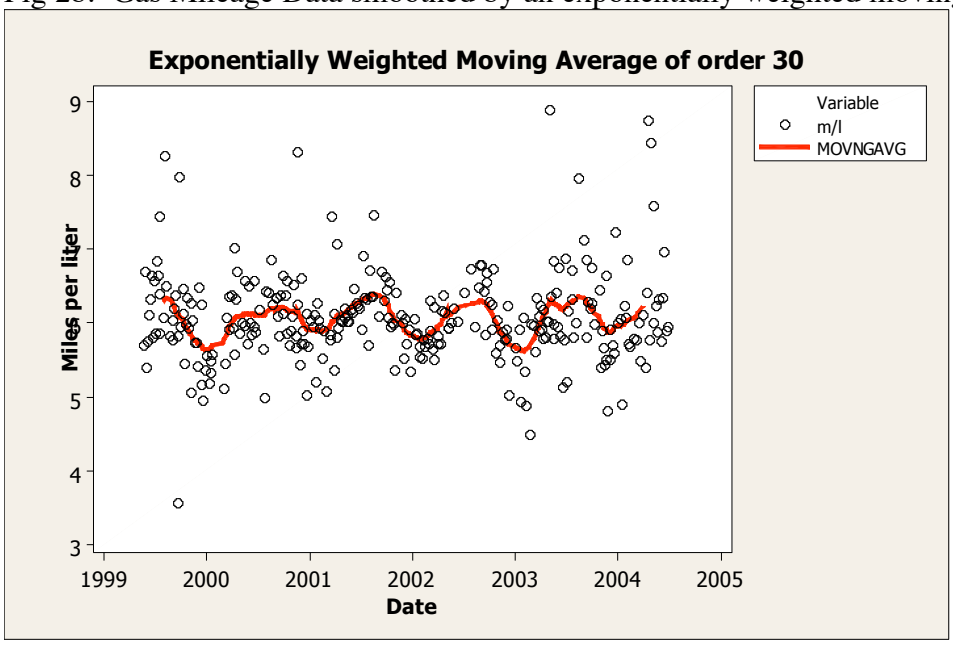
Once the idea of kernel estimation of a density is presented, it is a small step to suggest a similar approach to smoothing a time series. In fact the primitive smoother of using a moving average is directly analogous to the primitive local-count density estimate just outlined. For a time series with equally spaced time intervals between observations, an m-order moving average will be an average of the m closest series values. Instead of counting the close values, one averages them. The graph below shows the moving average of some gas mileage data collected over a five-year period.

Fig 2a: Gas Mileage Data smoothed by an ordinary moving average



The kernel extension of this is to use a weighted average rather than an ordinary average, with more weight on closer values. The step of moving from a rectangular kernel to a more gradual one is just as with the histogram variant. The weighting function in the following graph was $\exp(-cd^2)$ where d is the distance to the grid point, and c is a constant determined either by a default based on the x-range of the data, or by trial and error. The result of the modification is shown below with the same data as the previous graph. Again, the general shape of the weight function is all that the novice need appreciate, rather than the exact algebraic form, so this extension is also feasible to a lay audience.

Fig 2b: Gas Mileage Data smoothed by an exponentially weighted moving average

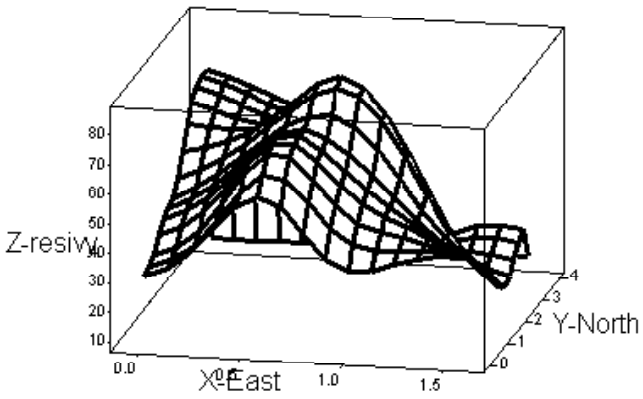


The exponentially weighted moving average of the same order will always be a little less smooth, and also a little less biased, than the unweighted moving average of the same order.

This nonparametric smoothing method is not limited to time series. For any data in which the objective is to fit a curve y=f(x) to predict y from x, the same weighted average smoother is appropriate. The avoidance of guessing the parametric form of the fit is a great benefit. Residual analysis is still possible based on this nonparametric fit.

To illustrate, suppose we have a trivariate data set (y,x_1,x_2). The same approach is easily extended. We use a two-dimensional grid in the x_1-x_2 plane, and again compute exponentially weighted averages of nearby values. The result can be displayed as a contour plot or as a wireframe plot.

Fig. 3 Wireframe Plot

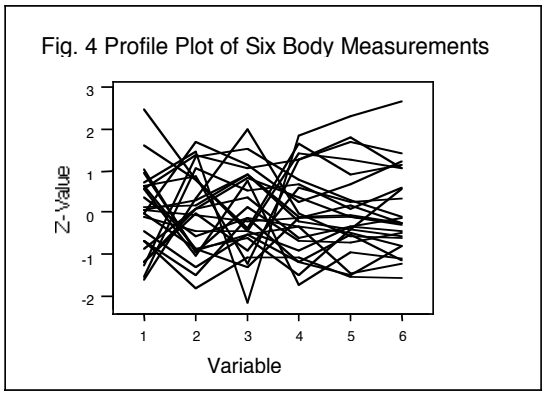


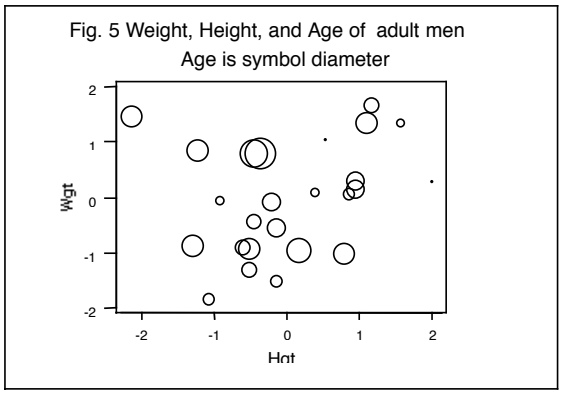
The data on which this surface estimate was based is taken from a data set of 1000 soil resistivity measurements over a rectangular agricultural area, taken from Cleveland(1993).

The techniques just described are really just applications of weighted moving averages – this technique is not too complex in its simplest forms and can be motivated by techniques no more complicated than the histogram.

Multivariate Data Display

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.



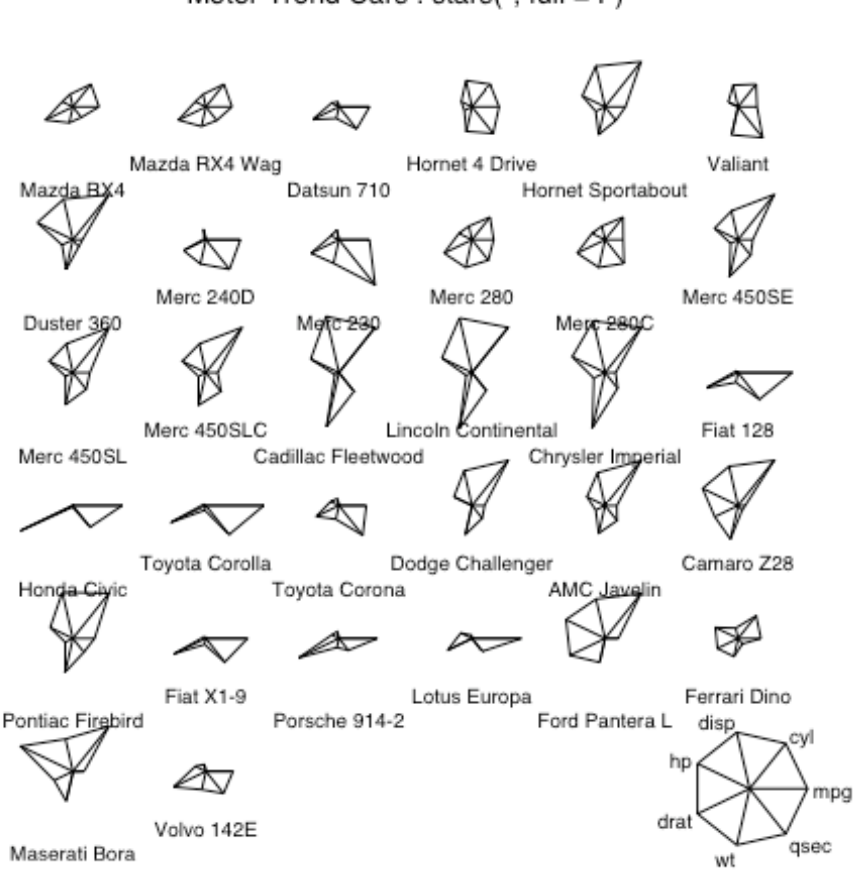


The Profile plot is not a useful summary plot but is very useful for data analysis: detecting coherent subsets, outliers and correlations among several variables.

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. Additional variables can be added by adding features to the plotted figures.

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 stars themselves, but these are optional enhancements. This technique is useful for 10-15 variables – more than that causes confusion in identifying the variables from the direction spokes.

Fig. 6 Example of Star Plots from R examples. (R(2004))

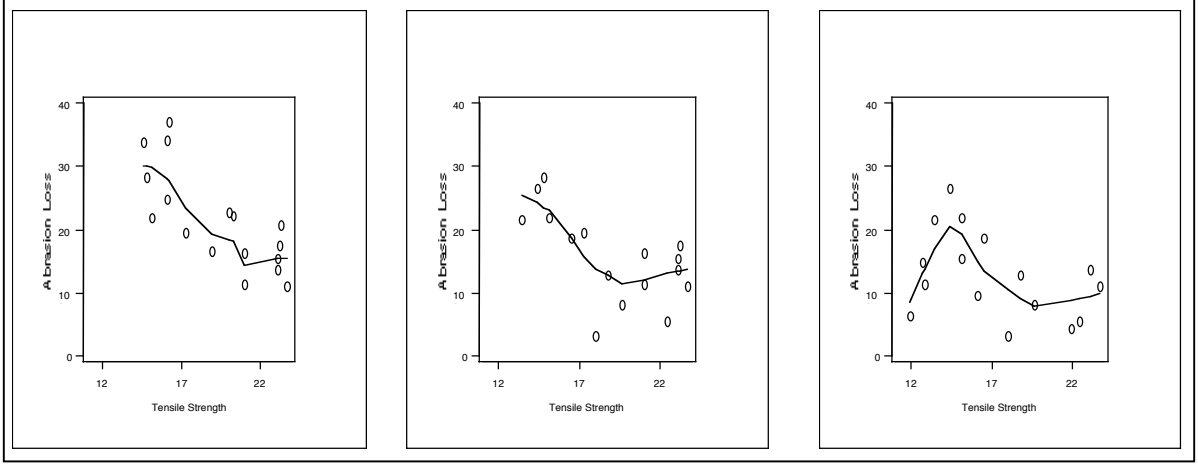


Use of these plotting methods in an early service course may be one way to alert students to the very common multivariate nature of real-life data. Also, it gives an appropriate emphasis to graphical methods, which is surely one of the most popular methods of data analysis.

A technique that works well for 3-6 variables is the coplot, or conditioning plot. 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 a smoothing technique such as the moving average technique just described. The loess procedure is also available in software and is better than the moving average but is a bit more difficult to explain. However, as long as it is understood as a smoothing technique, it could be used instead of the moving average.

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. 7) 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. 7 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. Note that this kind of plot is ideal for explaining the concept of interaction.

The ability to graph data in more than two dimensions is very useful for realistic, and usually multivariate, data sets. Researchers involved in data-based studies may only have a single statistics course in their background. The suggestion here is that these techniques be exposed to students in a first course.

The Bootstrap

Recent texts by prominent statisticians have given new respectability to the resampling method – it is now starting to be used as a basis for introductory courses (Good(2001), Lunneborg(2000)). Proponents argue that resampling approaches can be understood with less background than traditional inference, and is usable in more complex settings.

A well-known example of a resampling procedure is the permutation test. It has been given brief mention in introductory statistics texts, presumably because it depends in practice on computer software that was not readily available to the broad spectrum of users until more recent years. However, this is no longer a constraint, and the extensive use of the permutation test for introductory courses has been demonstrated by Good(2001). In this paper I will limit my comments to another resampling procedure, the bootstrap. It addresses directly one of the basic goals of the discipline – to understand the impact of unexplained variation.

The basic idea of the bootstrap is that the data itself can be used as a proxy for the target population, for the purpose of evaluating precision of a statistic. One description of the method is to resample from the data, with replacement, to get some "bootstrap" samples of the same size as the original sample, and observe the resulting variability in the calculated statistic for each bootstrap sample. This simple procedure can be explained to students with very little background. Another approach is comforting to theoreticians but harder to explain at the introductory level: simply use the inverse transformation from the empirical cumulative distribution function to simulate new samples, and again compute the variability of the computed statistics. One can argue that the ecdf is a good estimate of the population cdf, and it requires no assumption about the parametric form of the population cdf. This explanation of the bootstrap requires knowledge of the probability integral transformation and would usually not be appropriate for introductory courses. But this explanation is not necessary for plausibility of the technique or for directions for the procedure.

The fact that the bootstrap can be improved using advanced techniques, for example, smoothing the ecdf, does not detract from the very general utility of the ecdf itself, since the process of resampling the ecdf is so simple to explain.

It is important that students have been exposed to the idea of simulation for a painless introduction to the bootstrap. All that is needed is some simple computer-based versions of coin tossing or selection of numbered tickets. 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. To demonstrate the ease with which bootstrapping can be introduced, consider the following.

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. Part of the confusion is that students do not know what the sampling distribution of the mean is. Seeing the bootstrap resamples, and their means, might help. Consider the artificial population of integers {0,1,2,…,500). The reason for choosing this population rather than a real one is that the resulting estimates can be checked and students more convinced that the method works.

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, and ignoring for now the fact that the population is actually known?

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, 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 simulated resamples is that the sample mean has a standard deviation of 32.8. To report the estimate of the population mean and its precision, the estimate could be stated as 225.4 ± 32.8 Now we shall see how the more traditional estimate results. The actual sample SD is 167.3 and we sue this as our estimate for σ_x and compute "theoretical" standard deviation of the sample mean = $\hat{\sigma}_x / \sqrt{n} = 167.3 / \sqrt{25}$ = 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 ± 33.4 as our estimate of the population mean. 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. 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 $= WGT / HT^2$, in which WGT is in kg and HT is

in meters. We want to estimate the value which would be greater than 90 percent of the men in this age group, as possible use for clinical evaluation of excessive weight.

The data for the BMI looks like this:

22.1 23.8 26.8 28.2 24.8 24.6 29.9 23.2 32.0 29.3 27.1 26.7 20.1 20.3 22.7 33.5 23.9 20.0 25.4 21.2 29.4 26.5 20.8 21.6 27.0

And so the 90^{th} percentile might be estimated as half-way between 29.4 and 29.9 or 29.65. A concern however is that the sample is too small to provide a reliable estimate of the 90^{th} 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 90^{th} percentile estimate we have used to arrive at 29.65 has a standard deviation of about 1.5. This suggests our estimated 90^{th} percentile could be easily be the 60^{th} percentile up to the 96^{th} percentile, or even more in error, and so would be inadequate for clinical use. The sampling theory of the 90^{th} 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.

A criticism of the raw bootstrap is that it gives estimates that are inferior to other methods in situations where more is known about the population, such as symmetry of the population distribution. However, the challenge of teaching methods to address these special situations in an introductory course is daunting, and may be unrealistic. It may be better to describe a method that is advertised as sub-optimal, and easy to understand, than several methods that are optimal in various circumstances, but not understood in any of them!

How important is the bootstrap for the future of statistics? Kotz and Johnson (1992) select Efron's 1979 paper as one of the most important breakthroughs in the development of the subject. The commentary by Beran(1992) explains this breakthrough in the light of more recent research. As with any method in statistics, the details of the bootstrap are complex. But the basic idea is understandable without much preparation, and the hazards of using this idea are possibly less than the hazards of using p-values and confidence intervals with minimal preparation. The bootstrap idea should be included in introductory courses.

Conclusion

A feature of academic life is that new developments are initially "research" and considered to be advanced topics. It seems that a long time is required for new

techniques to filter down to the introductory courses. Part of the reason for this is that important new developments reported in research journals are further developed in these same journals. When this happens, the complexities of the development mark it as an advanced topic, and if it is taught at all, it is taught to graduate students. But the original idea may be useful to novices even if not refined. I have tried to identify some examples of topics in this category, and hope that instructors will experiment with inclusion of them in introductory courses.

References

Lunneborg, C. E. (2000) *Data Analysis by Resampling: concepts and applications* : Duxbury, Pacific Grove, CA.

Good, P. I. (2001) *Resampling Methods : a practical guide to data analysis*. Second Edition. Birkhauser, Boston.

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

Kotz, S and Johnson, N.L. (1992) *Breakthroughs in Statistics.* Springer-Verlag. New York.

Beran, R.J. (1992) . Introduction to "Efron(1979) Bootstrap Methods: Another Look at the Jackknife. In Kotz and Johnson (1992), Vol II, pp 565-568. (cited above).

Pregibon, D. and DuMouchel, W. (2001) What is Data Mining? Handout in Tutorial on Data Mining, 53rd Session of ISI, Slide 29. Soeul, Korea.

R Development Core Team (2004) R: A Language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. (Note: url is http://www.R-project.org)