TRADITION SHOULD NOT SUPPLANT UNDERSTANDING AND INSIGHT

Richard Wilson[1] and John Maindonald[2]
[1]School of Mathematics and Physics, The University of Queensland, Australia
[2]Centre for Mathematics & Its Applications, Australian National University, Australia
rjw@maths.uq.edu.au

Technological changes and theoretical advances in the past several decades have created new demands and opportunities for cooperation between the statistical mainstream and application area specialists. Against this, traditions of statistical analysis have become embedded in some places that too often hinder understanding and insight. This paper will discuss: a) approaches that were never a good idea; b) common approaches which are outdated due to advances in modeling and computer technology; c) use of over-simplistic modeling assumptions; d) data pre-processing which removes key information; and e) discipline focus on one specific statistical paradigm, rather than choosing the paradigm that relates best to the research question. How can change best be effected through education and collaboration? What role may applied and mathematical statisticians have in such change?

OVERVIEW

It is a commonplace that statistical analyses, as represented in the applied literature, are disturbingly often cavalier, misleading or do not give an adequate account of the data and of the processes that generated it. When checks on published results from pre-clinical cancer trials find that upwards of 75% of results cannot be reproduced (Begley & Ellis, 2012), something is clearly wrong. Problems include faulty design and/or execution, overly simplistic analyses and failure to critique model assumptions. Incorrect, limited or partial conclusions are the result.

We will begin with comments on historical influences, both within the statistical discipline and in other disciplines. Entrenched traditions of statistical analysis, especially less-than-ideal practices that are considered “gold standards”, may not readily change in the light of modern critiques and understandings. For many in other disciplines, there may be limited exposure to modern statistical analyses. The data analysis paradigm that relates best to the research question should be chosen, rather than a paradigm that may be the current fashion in the discipline.

Examples will illustrate the uses and implications of modern tools for graphical exploration of data, for model diagnostics and for presentation of results. In particular, emphasis will be placed on the damaging effects of categorization and other preprocessing that lose information from data. Mention will be made on approaches that were never a good idea, including abuses associated with p-values. We note tools and approaches, such as resampling methods, that can help identify and deal with over-simplistic modeling assumptions.

HISTORICAL CONSIDERATIONS

Three historical developments will be considered briefly. First, what is the appropriate general theoretical framework, Bayesian with its longer history, or frequentist with its use of *p*-values in the style of R.A. Fisher (see, for example, Fisher, 1925)? A widely held view is that both methodologies have a place in the data analyst’s armory. Bayesian theory provides a broad framework within which to think about statistical problems. The *p*-value methodology may be regarded as providing a rough approximation to Bayesian inference that may be used in at least some cases where an “uninformative” prior is appropriate.

Second, some application areas have developed their own tradition of statistical analyses, separate from or on the fringes of the statistical mainstream. Approaches have been followed for which the main rationale has been mechanical convenience. A new methodology may emerge from time to time as a new fashion.

The third point is that development during the pre-computer era remains formative for much methodology that continues in use today, even though modern computational capability has made some methods less optimal or even redundant. The implications of modern technology have still to have the influence that is desirable on statistical training and practice.

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

EXAMPLES

The following examples will be used to illustrate issues that are of concern.

Example 1 - Reinhart and Rogoff (2010a, 2010b)

A recent controversy centered around papers (Reinhart and Rogoff 2010a and 2010b – referred to as R&R below) that used data on the economic performance of 20 developed economies to support claims that "once debt reached 90% of GDP, economic performance deteriorated sharply". A Herndon et al (2013) critique received huge attention, both publicly and in the economics community, focusing mainly on a serious Excel spreadsheet, but with other criticisms as well. Our focus is on flaws in the analysis methodology. These occur in the tabular and graphical summarization of the data, this being the only form of analysis used.

Example 2 - Student Study of Effect of Glucose Supplement and Activity on Blood Glucose

Students in a biomedical science designed a two week crossover experiment in which each subject was randomly allocated to one of two groups. Each Group A subject performed one type of physical activity while each Group B subject performed an alternative physical activity. Both groups were further split into two subgroups, where one subgroup received a loaded (with glucose) supplement and the other an unloaded supplement in week 1, with the subgroups receiving the alternative supplement in week 2. Each subject was to fast before the experiment and had blood glucose levels measured at five different stages: after fasting and before receiving the supplement; at 15 minutes after receiving the supplement; after 5 minutes of activity; at 10 minutes after finishing the activity; and at 20 minutes after finishing the activity. What detectable effect did the different activities and different supplements have on blood glucose levels?

Example 3 - Student Study of Rates of Production of Lycopene

Different *E. coli* strains were used to produce lycopene (which has industrial uses). Of initial interest was the rate of growth of the different strains of the microbes. Cultures were set up with "food" and the optical density of each culture was measured over time. Initial readings were unreliable and so ignored. After the food reduced, growth rates slowed, so later readings were also ignored. Approximately exponential growth might be expected (each cell should replicate itself in a short period of time). Accordingly, standard practice is to take logarithms of the measurements and then visually select the section of graph which "looks linear". A linear regression is fitted to this and the estimated slope taken as the observation to be used for comparison. In this experiment, three strains were used with three replicates of each, giving nine "observations" (slope estimates).

GRAPHICAL EXPLORATION OF DATA

Modern computer graphics have brought huge advances, beyond what could be done with pencil and paper. It is reasonable to expect that anyone involved in data analysis will make use of the tools that are now available for graphical exploration of data.

Example 1: The R&R Use of Tabular and Graphical Summary

R&R break the data into four categories, with cutpoints at debt to GDP ratios of 30%, 60% and 90%. No compelling reason, except perhaps previous use, is given for these three cutpoints. (R&R acknowledge that the sensitivity of the analysis to the choice of cutpoints should be checked.) Barchart data summaries, showing means and medians, then appear throughout the paper. R&R do not present any measures of statistical reproducibility.

For exploratory purposes, it is reasonable to plot GDP growth against debt to GDP ratio, with identification of country and year for outlying points. Figure 1 presents individual data values, with solid red horizontal lines showing mean GDP growth for each of the categories. The spreadsheet error reduced the mean for the over 90% category (see dashed line). Outlying points are labelled according to country and year, and following Herndon et al (2013), a smooth curve has been fitted (using the generalized through the points additive modeling abilities of R's mgcv package (Wood 2006)). Simplistic assumptions (errors about the underlying relationship are independent) are involved in the generation of this curve, so the pointwise 95% confidence bounds (in gray) give a broad indication only, but they do give a much fairer representation of the data.

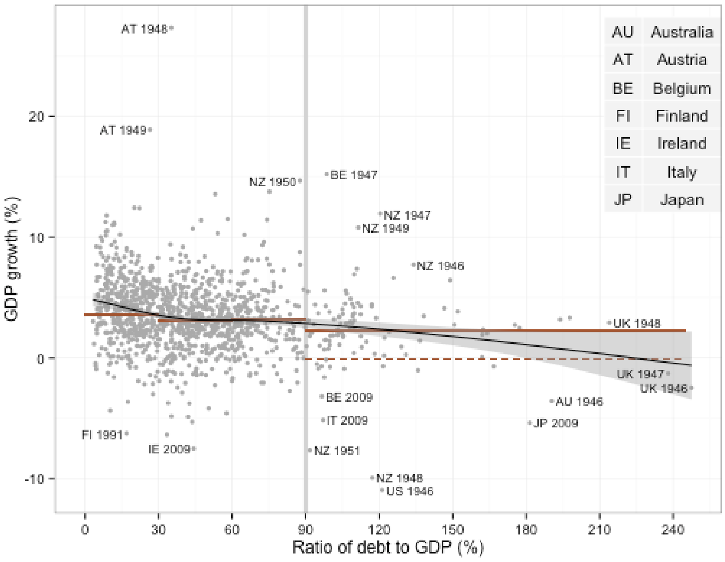

Figure 1: GDP growth versus Ratio of debt to GDP: individual data values.

Figure 1 gives useful insight on problems with the analysis. It makes it clear that points from the years following World War II have a large effect on the means for the over 90% category. The NZ GDP growth for the years 1946 - 1951 follow the pattern h/h/l/h/h/l, where h is high and l is low, with a debt to GDP ratio that ranged between 88% and 134%. Wild swings in the economic growth measure were clearly unconnected with changes in a high debt to GDP. The separation of the 1950 figure in the 60% - 90% category post-war figures is artificial.

More general criticisms are:

- The breaking down of data by discrete categories loses information. With modern tools, breaking data down in this way is unnecessary.
- There was no investigation whether the pattern was consistent over time.
- Why focus on the relationship between these variables in the same year? Such effects as can be identified for the medium term future are more important.

Exploratory data analysis tools, showing individual data points would probably have identified the calculation error, as it skewed the results by omitting several data points. Simple graphical investigations would make it clear that points for a 90% debt to GDP ratio were unduly influenced by data from when economies were still recovering from the effects of World War II.

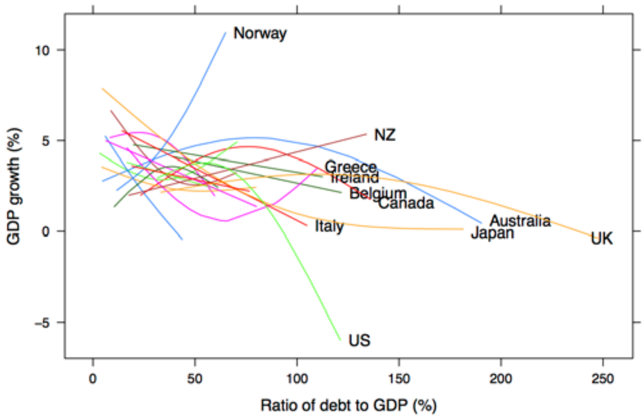

Figure 2: Smooth curves have been fitted to the individual country profiles.

It is possible to indicate what curves fitted to the individual country profiles would look like. This is important towards asking whether any relationship that seems detectable from the data is consistent over countries. Figure 2 plots smooth curves (dependent on spline basis) that have been fitted to the individual country profiles. There is no consistency across countries.

For an issue such as this, it may be reasonable to publish results from exploratory analyses, with the limitations of doing so acknowledged. Some testing of the sensitivity of the results to the influence of individual country profiles would be appropriate, extending perhaps to leave-one-country-out-at-a time cross-validation.

Example 2: Presentation of Biomedical Science Data

Typically, means and standard deviations for different groups are presented as bar charts in reporting results of biomedical science investigations. Consequently, students were initially encouraged to present the results of their experiment in the form shown in Figure 3(a) (showing results for week 2). This presentation hides the clear unusual values for some subjects and the repeated measures nature of the experiment. These aspects are clearly seen in Figure 3(b).

As can be seen, the bar chart presentation hides the student who did not fast, the student who ate something during the rest period and students whose results did not follow the general pattern of the others, as well as subject effects. In addition, the barchart presentation encouraged students to simply do multiple pairwise comparisons (treating groups as independent samples and ignoring any other assumptions), rather than exploring more fully the differences comprehensively. In analyses of this type, it is important to decide which contrasts are of primary interest and therefore may require testing and estimating.

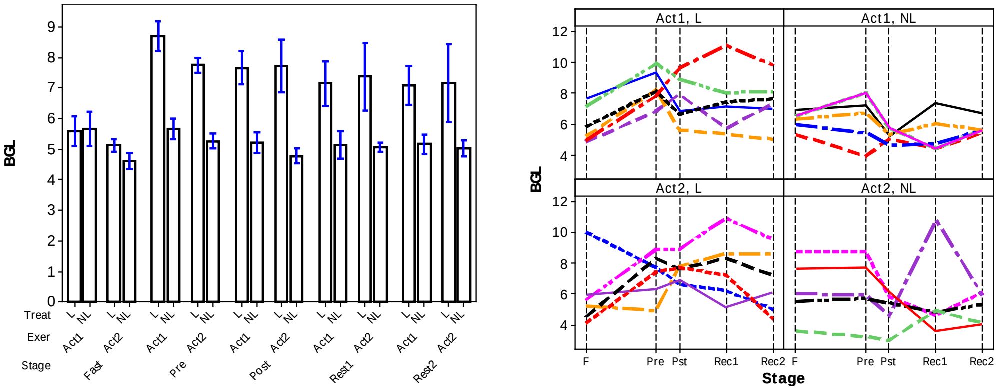

Figure 3: (a) Bart chart of means with intervals given by one standard deviation from the means. (b) Scatterplot of data with connecting lines for each subject and split into panels for each group.

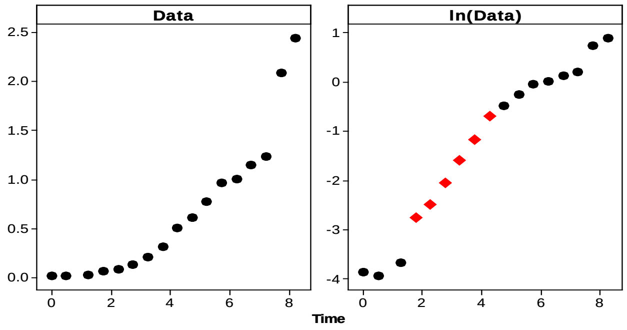

Figure 4: (a) Raw data (b) Logged data

Example 3: Preprocessing and Visual Guessing with Bioengineering Data

In Example 3, a visual guessing game is used to preprocess the data so that that estimated slopes from fitting linear regressions can be used as data. This is done with little regard for influential values. In Figure 4, the data for a single run are plotted against time in (a) and the logarithms of these values are plotted against time in (b). The points indicated in red would then be used in a linear regression to estimate the rate of replication in the raw data. In this example, removing one or other of the end values will significantly change the estimate. Using the original data (for all runs) would use far more information, remove the unreliable guessing game and give a fuller picture of the changes of interest.

MULTIPLE P-VALUES

All methodologies are open to abuse. This seems especially true of the mechanical use of *p*-values, particularly when many hypothesis tests are used in a single context. Multiple testing can often be avoided by a judicious choice of hypotheses, so focusing on issues of major importance instead of ad hoc testing (as in Example 2). The choice of methodology should be tuned to the question of interest. Furthermore, confidence intervals should standardly be given, even if the main emphasis is on a test of hypothesis. Classical hypothesis testing assumes a single hypothesis. Confidence interval calculations assume a single known model.

The following do not, without adaptation, fit the hypothesis testing scenario:

(a) Testing of multiple hypotheses, in contexts where it is unacceptable for the probability of accepting one or more hypotheses to be much greater than the common 0.05 significance level.

(b) Preliminary use of hypothesis tests in the process of selecting the model, including choice of an assumed error distribution and/or deciding on whether variable transformation is appropriate. Variable selection procedures, when followed by hypothesis testing and/or calculation of confidence intervals for the selected model, have this character. See Figure 5 below.

Both (a) and (b) have the character of fishing expeditions. In (a), several anglers set out to find a fish, usually with limited communication between the different anglers. The expedition will be judged a success if just one of the anglers succeeds in finding a fish. In (b), account has to be taken of the extent to which the odds have been stacked, in advance, in ways that affect the chances of finding fish. Unless, that is, resources allow large amounts of time and effort to be wasted on fish that, upon further investigation, will prove unmarketable.

Multiple Testing in Parallel

Controversy over the use of individual tests, versus multiple range tests, commonly arises because the aims are formulated too vaguely. If just two or three tests are conducted, is a modest increase in the probability of rejection of one or more of the null hypotheses acceptable? The issue is serious for high throughout genomic data where tens of thousands of gene transcripts are checked for expression. If 20,000 transcripts are checked where none are expressed, and if the probability of a positive is 0.05 in each individual case, then the expected number of false positives is 1000. The false discovery rate approach is one of several developed to handle this situation. The researcher has to manage the trade-off between missing cases of true expression, and fruitless work on transcripts that are not genuinely expressed — a classical decision analysis problem.

Variable Selection in Regression

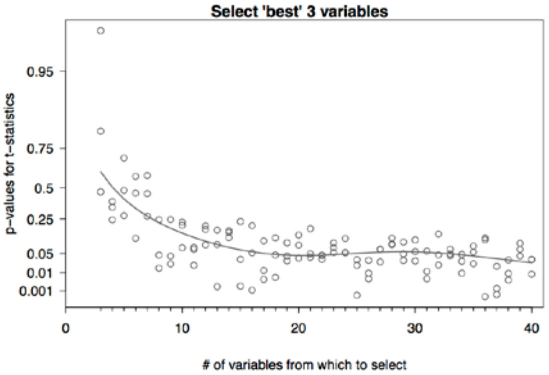

Figure 5: *P*-values for the coefficients for the three selected variables.

This is perhaps a good idea that has (mostly) gone wrong. As an example, if standard variable selection techniques are used to select the "best" three explanatory variables out of ten, *p*-values for the coefficients for the three selected variables, as given by standard regression software, are likely to be spuriously small. This can be readily seen by using such a selection process for a randomly generated dependent variable, with no relation to the explanatory variables. Figure 5 plots *p*-values from such regression analyses, against number of variables that are available for selection, with numbers available ranging from 3 to 40. The bias, which increases with the number of variables available for selection, is evident. However, the bias may be exaggerated here due to the independent explanatory variables as compared to the situation where this is not the case.

Variable selection approaches that avoid the biases shown include the use of separate data subsets for model selection and for model assessment. With a small data set, this is not, however, an efficient use of the data. Other possibilities include cross-validation and the bootstrap.

OTHER ISSUES: INDEPENDENCE ASSUMPTIONS AND RESAMPLING METHODS

If data have been collected according to a complex sampling design, then data analysis ought to respect any dependence (for example, from cluster sampling and/or from stratification) that is inherent in the design. Here the concern is rather with data that are to an extent observational in character. Spatial and/or time dependence are common in such data. Use of an autocorrelation plot to look for simple forms of time dependence in the residuals should be routine for data where there is a time sequence. This is clearly the case with all three examples above.

Former excuses for guessing, in contexts where the strict requirements of the classical theory are not satisfied, no longer apply. Providing that the assumptions can be stated in precise mathematical form, simulation can often be used to check the extent to which the theory can be trusted or it may give clues on how the theory should be modified. The provision that it must be possible to state the assumptions in precise mathematical form is important. This severely limits the use of simulation to check out the effects of lack of independence.

If it is suspected that assumptions may be wrong, it is reasonable to try a methodology that requires relatively weaker theoretical assumptions. Here, note the use of a training or test data approach, of cross-validation, of bootstrap methods and of permutation tests. None of these are an answer in every situation. They do however, between them, greatly extend the range of problems where answers can be given with fair confidence. The situation is clearest for prediction.

IMPROVING THE SITUATION AND CONCLUSION

Perhaps the most useful step would be to insist that all data and accompanying analysis be made available on a web page. Funding bodies should consider funding well qualified and experienced statisticians to undertake reviews of published data and analyses. This is especially important where the results have major implications for public policy, health or safety. In this area at least, the processes of science themselves require closer scientific scrutiny.

There is a huge gap between what statistical professionals and statistically informed subject area specialists consider a reasonable standard of analysis and much of what appears in the literature. The damage to science and to the scientific process is immense, resulting in both errors and a lack of confidence in statistical results.

REFERENCES

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature 483*, 531–533. doi:10.1038/483531a

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Oliver and Boyd.

Herndon, T., Ash, M., & Pollin, R. (2013). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Working Paper 322*, *Political Economy Research Institute.* http://www.peri.umass.edu/fileadmin/pdf/working_papers/working_papers_301-350/WP335.pdf

Minitab (2013). MINITAB 16.2.4. Minitab, State College, PA. (For Figures 3-4)

Reinhart, C. M., & Rogoff, K. S. (2010). Growth in a time of debt. *American Economic Review: Papers & Proceedings, 100,* 573–578. See also http://www.nber.org/papers/w15639

R Core Team (2014). R: A language and environment for statistical computing, *R Foundation for Statistical Computing*, Vienna, Austria, http://www.R-project.org (For Figures 1,2 and 5)

Reinhart, C. M., & Rogoff, K. S. (2010). Growth in a time of debt. *American Economic Review: Papers & Proceedings, 100,* 573–578. See also http://www.nber.org/papers/w15639

Reinhart, C. M. and Rogoff, K. S. (2010b). Growth in a time of debt. Working Paper 15639, National Bureau of Economic Research. http://www.nber.org/papers/w15639

Wood, S. N. (2006). *Generalized additive models. An introduction with R.* Chapman & Hall/CRC.