STATISTICS EDUCATION IN A CONSERVATION ORGANISATION: TOWARDS EVIDENCE BASED MANAGEMENT

Ian Westbrooke
Department of Conservation, Christchurch, New Zealand
iwestbrooke@doc.govt.nz

The Department of Conservation manages 30 percent of New Zealand, protecting biodiversity and promoting recreation. Our 1800-plus staff include several hundred science graduates. DOC's first dedicated statistician started in 2000, with improving design and analysis skills as first priority. Many staff require basic skills in data entry and exploration. Some need specialist statistical knowledge, especially modelling skills. Most data is observational in nature, with interest centring on estimation of effects (and their error). However most graduates' training focused on experiments and hypothesis tests. We developed a course to broaden knowledge of the linear model, and extend it to glms; then added one on longitudinal data analysis using mixed models. The next priority is training in practical sampling design. For our purposes, a stronger statistical modelling approach is needed in university training, with less emphasis on hypothesis tests.

INTRODUCTION

This paper uses a case study of statistics education in my own workplace to look at statistics training in industry. I'll draw out some conclusions that may apply more generally, especially for training of graduates of other disciplines in applied or management (rather than research or academic) organisations, and have implications for what is needed in university courses.

Key areas considered are:

- Data management and exploration,
- Dealing with observational more than experimental data, and the need for effect size estimation rather than hypothesis tests, and the need for statistical modelling skills,
- The differences in approach needed for statistical training in a workplace context.

DOC'S NEEDS

New Zealand's Department of Conservation (DOC) is the central government organisation charged with promoting and implementing conservation of the natural and historic heritage of New Zealand. Thus DOC is responsible for managing about a third of the country's land area along with 19 marine reserves; protecting and managing much of the country's indigenous biodiversity including many unique species; promoting recreation; and facilitating tourism. Our 1800 plus staff include several hundred science graduates undertaking science and technical work at national, regional and local levels.

To carry out DOC's conservation management effectively requires evidence based on data. Typical questions facing managers include the following examples. What are the trends in abundance and health for native species and ecosystems, and how can management make a difference? How are visitors using parks and conservation lands and facilities, and what issues need to be managed? Moving beyond broad qualitative statements to answer these questions in an evidence-based approach demands quantitative assessments, based on data. As in all environmental studies, there is plenty of variability involved in conservation, so statistics become essential.

I started as DOC's first dedicated full-time statistician in 2000, and soon found that by far the most effective role for a sole statistician was to spend a high proportion of my time in statistics education and advocacy. Staff need skills in design, collection, analysis and reporting. A very broad layer of staff also need basic skills in effective data entry, management and exploration, including effective graphing. A key layer of staff need training in statistical modelling skills, starting from the linear model, through its extensions, including mixed models for repeated measures. In addition, smaller numbers need knowledge of specialist areas such as estimating animal abundance or survival analysis. We have developed and promoted courses in these areas, using a mixture of in-house and external expertise.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

Initial emphasis has been on providing tools to allow for increased and improved data analysis. Like many management organizations, DOC collects much more data than is ever analysed. The importance of work in this area was reinforced in survey of the several thousand ongoing biodiversity monitoring projects (Figure 1) undertaken by the organisation. Data analysis scored substantially lower than any other category, in an overall assessment of quality.

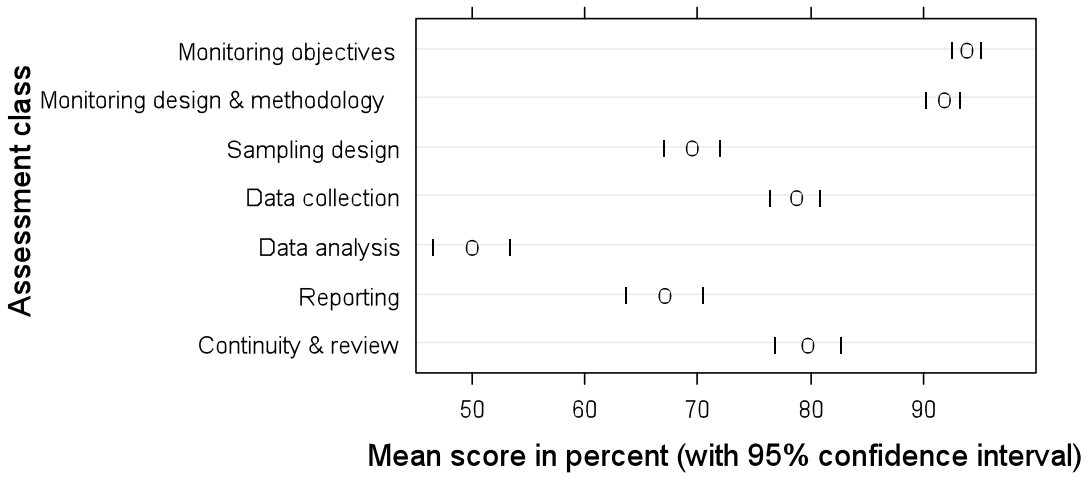

Figure 1. Quality scores on a stratified random sample of 448 (of more than 2000) ongoing biodiversity monitoring projects at DOC
(Projects were scored with yes/no responses on 31 criteria, grouped into seven categories as indicated. Note the weakness of data analysis, along with reporting and sampling design.)

EFFECTIVE DATA ENTRY, MANAGEMENT AND EXPLORATION

Data entry and preparation is an important but neglected area of statistical practice, and is essential for the key tasks of data exploration and analysis. In fact, there is a key phase where data preparation and exploration need to interact to ensure that the data is in a state fit for analysis.

Initially many DOC staff apologised about their data, a typical statement being "my data is only in Excel". In fact, we have found that Excel to be a very good general tool for data handling, providing appropriate guidelines for data format are followed. A useful reference is http://www.reading.ac.uk/SSC/publications/guides/topsde.html. The key thing to emphasise to students is the importance of standard data formatting. Each observation has its own row; each variable goes in a column with a meaningful name; and only raw data goes in data sheets, with no blank rows. Analysis and summaries go elsewhere. Fortunately, this layout works both when data is transferred to statistical packages, and for using Excel's excellent cross-tabulation tool, the Pivot Table. When staff see the advantages of the new layout, particularly through creating summary tables quickly and easily with Pivot Tables, the new approach is happily adopted.

Legitimising the use of Excel for data entry, storage and initial exploration has helped facilitate getting data off bits of papers and into computers for analysis. Many hundred staff have taken part in a course we call *Using Excel to enter, manage and explore data*, originally developed by statisticians at New Zealand's largest agricultural research institute, AgResearch. Along with tools for assisting entry and validation, it covers exploration of data, using tables and production of graphs.

GRAPHS

To facilitate data exploration, and improve the quality of presentation, we developed a course on graphs, together with a manual (Kelly, Jasperse & Westbrooke, 2005). This draws heavily on Tufte (2001) and Cleveland (1994). We developed exercises including one (Figure 2) to allow students to learn for themselves about Cleveland's recommended order of visual perception (Cleveland & McGill 1985). Other exercises included examples showing how easy it is to make

default Excel graphs much better. After the manual was published, we added exercises demonstrating the inadequacies of pie graphs based on an excellent book by Robbins (2005).

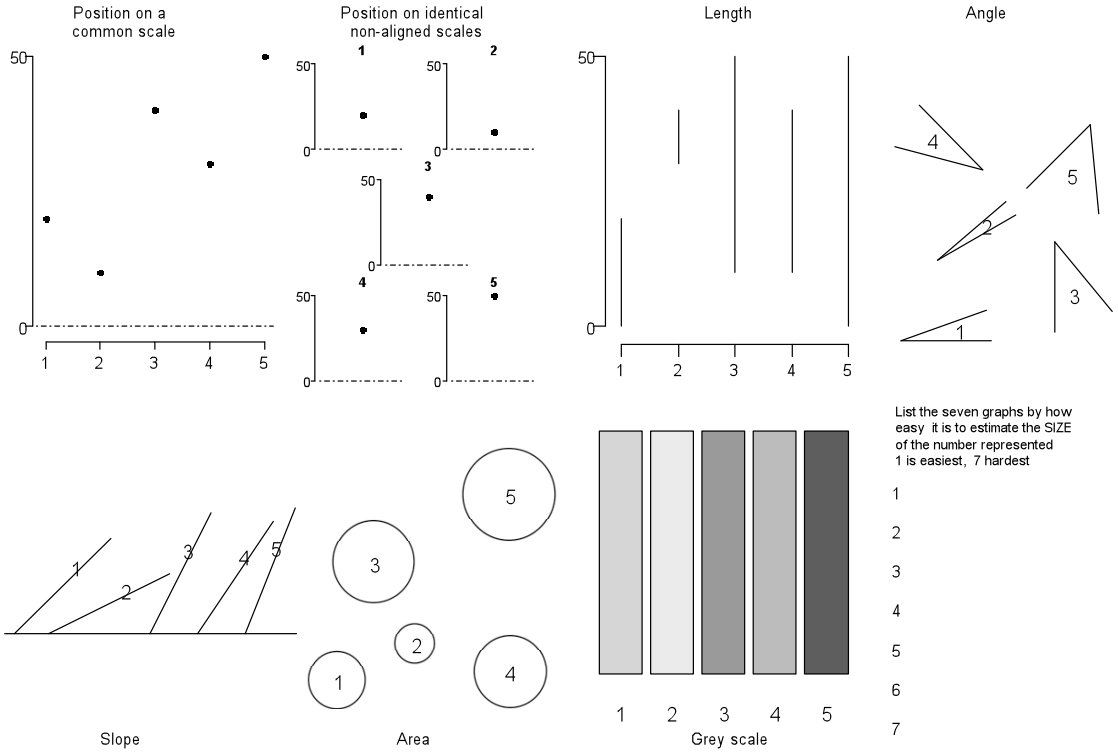

Figure 2. An exercise from our graphs course
(Students are asked to order the seven graphs by how easy it is estimate the size of the number represented. This exercise allows students to learn for themselves about the options for presenting quantitative data in graphs and leads into considering the accuracy of visual perception of different approaches.)

Originally, each of the Excel and graphs courses took a full day, but slightly shortened versions work well combined into two half-day workshops. However, these courses, especially the data entry aspects, do not necessarily require the skills of a trained statistician, and we are moving to establish a course with an external provider to present this material for both DOC staff and others interested.

Practical data handling and exploration are essential pre-requisites for successful application of statistics in the workplace, but often insufficiently covered in training.

SPECIALIST STATISTICAL AREAS

Most subject areas require some specialist statistical techniques, and conservation is no exception–for example in estimating parameters of animal populations, such as survival, reproduction and density. With New Zealand's unique and threatened species, mark/recapture and distance sampling approaches are important areas for development. For these, we have brought in experts from outside to provide training. Survival analysis, originally developed for human populations, has been needed to provide a firmer basis for estimation and for evaluating conservation management options. We developed materials in-house, with a written manual and series of workshops (Robertson & Westbrooke, 2005). Along with this, we found it useful to teach a basic course in deterministic matrix population modelling, which was an adventure into mathematics rather than statistics training for biologists. We have found that a combination of developing in-house expertise and drawing in outside specialists works well.

STATISTICAL MODELLING – THE KEY AREA

The most important statistical training area has been providing science and technical staff a grounding in modelling skills. Almost all DOC's data is observational in nature, and we are almost always more interested in estimating the size (and confidence in) effects or differences than in testing null hypotheses. However virtually all graduates have received statistical training at university focused on designed experiments and hypothesis tests. ANOVA is the main tool most had been exposed to. "How do I apply an ANOVA to this data?" was a question encountered more than once early on.

To remedy this, we have developed a three-day internal course to promote basic knowledge of the linear model, and extend it to generalised linear models. The prerequisite is a university statistics course or equivalent. This means most students have encountered ANOVA, plus probably linear regression. We start into modelling by explaining why we need statistical modelling for DOC data, reviewing simple linear regression and ANOVA, with each case starting with a single explanatory covariate. We then progress to multiple regression, then to mixtures of continuous and categorical covariates, drawing then together in *the linear model*, a unifying concept that is new to most participants. Next we extend this to generalised linear models (glms), with sessions on logistic regression and Poisson regression. Providing tools for handling binary and count data is critical, as these are very common in the data DOC staff are encountering. We include a session on using statistical software effectively, and additional material on graphing. Plus if time permits within the three days, we briefly introduce generalised additive models and/or tree-based models.

The course is taught workshop style, using real data wherever possible from DOC's work. Context and relevance of data is very important when teaching in the workplace. Each student works at a computer, accessing data, creating graphs and applying models as the trainer demonstrates.

Statistical software

We have used R (R Development Core Team, 2009) as the software in modelling courses. R is amazingly powerful and flexible, plus it is free. However, with this comes a steep learning curve. There are two main barriers to widespread adoption of R by staff who need it. Firstly typing code is new for those used to a point and click approach to data management and analysis. R Commander (Fox 2005) provides a good way of overcoming much of this barrier. It provides an introductory menu-based interface to R, while deliberately providing a bridge to creating and developing R code. We have recently converted our basic modelling course to be almost entirely in R Commander, and have found that this works very well. The second barrier to adoption of R is that accessibility for help is highly variable. Most help files within R are of very limited use to the uninitiated, and this tends to put people off. This is an area that needing further development if R is to be used more by people who are not professional statisticians.

Mixed models

More recently we have introduced a 3-day course extending to analysis of longitudinal-type data using mixed models. Identifying trends in various biodiversity measures is a common challenge for conservation management. Studies involving repeated measures on the same sample units are often useful for estimating such trends efficiently. However, extensions or alternatives to the linear (and glm) framework are required to deal with the correlations that may result from this design. In the course we introduce mixed models. The first day works through a data set with a continuous response variable, first misapplying a linear model, then applying a linear mixed model; while the second day works through a similar example for count data, applying a generalised linear mixed model. The third day involves applications–extending analysis on the first two days, and applying the approaches to trainees' own data.

SAMPLING DESIGN

The next priority is training in practical sampling design for applied conservation ecology, an area needing strengthening as shown in the recent survey of monitoring (Figure 1). There is a course currently under development. A major challenge is ensuring students absorb the basics of

randomisation and replication, and why they matter. Another important aspect is clarifying the differences between experiments and observational studies, particularly in strength of inference, and providing guidance on when and how to implement different types of studies.

DIFFERENCES IN WORKPLACE-BASED TRAINING

While the content of these courses has similarities to those presented in educational institutions, the workplace context has led to some distinct features. First, we emphasise practical applications and examples using real data, with a basic outline of the theoretical background. Formulae and mathematical notation is kept to a minimum, with no derivations or proofs. Second, we teach intensive block courses (typically one or three day) rather than multiple sessions over a longer period such as a quarter or semester. We have found it is much easier for staff faced with many competing priorities to commit to attending for a short block of time. Plus our staff are necessarily dispersed across New Zealand, as conservation management is often in remote areas. Third we work with small classes, a maximum of about 12, and with a high trainer to student ratio. One trainer can cope with up to five or six students, so we usually aim for two trainers. Fourth, we do not carry out formal assessment of students, as it would take up precious classroom time, and there is less need for formal qualifications in the workplace context. Instead, we ask the students to assess the course and its applicability to their work.

CONCLUSION

Key lessons that emerge from this case study are:

- Training that allows distinguishing between observational and experimental data and provides modelling skills applicable to different types of data is critical. Emphasis needs to be on estimation of effect sizes rather than hypothesis tests.
- Effective data management and exploration (especially graphing), skills, are needed, to provide the basis for data analysis.
- The workplace context means that courses work best as intensive block courses, rather than as series, with a practical rather than theoretical emphasis. The courses work best with hands-on computing integrated in, using real examples and datasets that attendees readily understand. Evaluation of the courses in meeting workplace objectives is more important than evaluation of individuals.

The key to a statistician making a difference in a much larger workplace is a strong training and advocacy role. One (and more recently two) statisticians can make a difference, in our case to help protect New Zealand's unique biodiversity and protected areas. We receive great support from the wider statistical community through consulting, receiving and providing specialist training, and through the availability of resources like R and more specialist software. For workplace application, our experience points to the need for a stronger statistical modelling approach, and less emphasis on hypothesis tests in academic statistics training for biological, ecological and social science students.

ACKNOWLEDGEMENTS

I'd like to thank the many statisticians and others who have helped with development of training at DOC, especially Jennifer Brown, Neil Cox and Maheswaran Rohan who have played major roles in developing some of the courses.

REFERENCES

Cleveland, W. S. (1994). *The Element of Graphing Data.* Summit NJ. Hobart.

Cleveland, W. S., & McGill, R. (1985). Graphical Perception and Graphical Methods for Analyzing Scientific Data. *Science, 229,* 828-833.

Fox, J. (2005). The R Commander: A basic-statistics graphical user interface to R. *Journal of Statistical Software, 19*(9), 1-42.

Kelly, D., Jasperse, J., & Westbrooke, I. (2005). Designing science graphs for data analysis and presentation: the bad, the good and the better. *Department of Conservation Technical Series 32*. Wellington: Department of Conservation. Online: *www.doc.govt.nz/upload/documents/science-and-technical/docts32.pdf*

R Development Core Team. (2009). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Online: *www.R-project.org*.

Robbins, N. B. (2005). *Creating More Effective Graphs*. Hoboken NJ.: Wiley.

Robertson, H. A., & Westbrooke, I. M. (2005). A practical guide to the management and analysis of survivorship data from radio-tracking studies. *Department of Conservation Technical Series 31*. Wellington: Department of Conservation. Online: *www.doc.govt.nz/upload/documents/science-and-technical/docts31.pdf*

Tufte, E. (2001). *The visual display of quantitative information*. Cheshire, Connecticut: Graphics Press.