Lessons from Advanced Statistics Workshops for Researchers in Other Disciplines

John A. Harraway
University of Otago, Dunedin, New Zealand jharraway@maths.otago.ac.nz

Abstract

Data arise in most areas of research and analysis of such data invariably involves advanced statistical procedures extending far beyond what is taught in an introductory course on statistical methods. The researchers, who are frequently postgraduate students in subjects ranging from agriculture, biology, ecology, chemistry to the social sciences, health and psychology, have limited mathematical background which restricts methods that can be used in the presentation of topics. The goal is to produce researchers who understand concepts and are able to carry out data analyses confidently in future with limited guidance from a statistician. Several approaches to the problem of providing appropriate training in advanced statistical methods are discussed and student evaluations of the approaches are reported. The discussion features topic content, assessment, the method of teaching, the importance of context for different groups of researchers, the use of appropriate software and the timing of the teaching in a research programme which involves experimental or field work and data collection.

Key Words: Statistics workshops, post graduate training, context, software.

1. Introduction

Researchers including post graduate students and faculty in a wide range of disciplines at a university use advanced statistical methods in the design of their studies and the analysis of data. The researchers are familiar with the content of an introductory course in statistical methods covering data presentation, simple probability, sampling, estimation, hypothesis testing, contingency tables, simple regression, one and two factor analysis of variance and non parametric procedures. But the required methodology extends beyond these introductory topics and good understanding of advanced methods is compounded by limited knowledge of probability, inference and mathematics.

Two studies have informed on topics to be included in training researchers from other disciplines in advanced statistics beyond the introductory course on methods. The first investigated statistics use in 2927 research papers published in 1999 in 16 high impact journals from botany, ecology, food science, nutrition and marine science (Harraway, Manly, Sutherland & McRae, 2001). The second study surveyed 913 recent graduates, currently in employment, with PhD and Masters Degrees in the biological sciences, psychology, business and statistics. These studies identified gaps between topics and techniques learned at university and those used in the workplace (Harraway & Barker, 2005). Both studies highlighted deficiencies in regression and advanced modelling such as mixed models, multivariate statistics, experimental design including power analysis and computer intensive statistics. Context was suggested as an important issue and understanding the merits of various statistical packages.

2. Advanced Workshops on Statistical Methods

The first request for an advanced statistics workshop as a consequence of our surveys came from the Ecology Research Group at the University of Otago. They requested a workshop on multivariate statistics. In discussion it was decided to offer a four day course covering multivariate analysis of variance, principal components, factor analysis, discriminant function analysis, canonical correlation analysis, cluster analysis, multidimensional scaling and correspondence analysis. It was agreed to work with SPSS, to use case studies in contexts associated with ecology, to encourage those attending to bring their own data and to place an emphasis on hands on computer activity. Each morning and afternoon should have a one and a half hour taught "seminar" followed by a one and a half hour session on the computers. There were 30 students enrolled paying NZ$250 each. The income went to our Department to cover costs as the workshop was provided outside normal teaching loads. A portion of the money was set aside to help those teaching the course pay for later conference attendance. The workshop was presented to fit with student field work. A course book was printed and three tutors were employed in the laboratory sessions.

The following year a second workshop with the same structure of four days was provided on regression and generalized linear models. In discussion with the organisers SPSS was again used. Topics covered included a review of simple and multiple linear regression, underlying assumptions, model selection, logistic regression, multinomial regression, log linear models and an introduction to linear mixed models. The next two years the multivariate statistics workshop was repeated. The workshops were enjoyable to teach. There was no examination. Students brought their own data which provided a challenge in the hands-on computer sessions although data sets covering the topic in the previous taught "seminar" were provided, chosen to reflect an area in ecology.

The Director of the Ecology Research Group reported that "--the recent multivariate workshop sponsored by the Ecology Research Group received glowing reports from those who attended. The word 'enjoyed' was freely used. The workshop was highly successful, not just because of the way it was taught, but also in terms of the take-up rate by students." The student survey was positive and the student coordinator said that "for postgraduate students a timely compact workshop is highly desirable because it can be incorporated with other commitments like field-work, conferences and lab-experiments." Context was identified as an important feature in the survey as this helped students understand the type and structure of data that needs to be sampled and analysed..

3. Alternative Approaches

To make advanced statistics workshops more widely available to researchers in other disciplines and to overcome funding problems after ecology funds were exhausted it was decided in 2012 to provide a traditional course with lectures and an examination at the postgraduate level for a wider group of researchers at the university. The course *Statistical Modelling for Research* covering probability distributions, data collection, normal linear models, generalised linear models, model-selection and model-checking using the program R was introduced and widely promoted. Students could earn formal course credit and make use of alternative funding opportunities. But the attractive sounding course failed to attract a reasonable audience and has not been continued for 2013. Although the 12 students enrolled were highly motivated there were substantial differences in their statistical backgrounds which made teaching a challenge.

A second approach encouraged research students to enrol in three second year courses, all three having only an introductory statistics prerequisite. The three courses are *Study Design, Regression and Modelling 1 and Multivariate Methods* with both SPSS and R being used. An average of 60 students has enrolled in these courses in 2013, but only a few of these students are in our target group of current postgraduate researchers from other disciplines. The majority are students majoring in statistics or constructing a minor in statistics while majoring at the undergraduate level in another discipline such as zoology or psychology. The minor involving five taught statistics courses is an excellent way of preparing for research in other disciplines but does not attract current researchers. Hence for 2014 it is intended to introduce a single second year advanced statistical methods paper covering regression, study design and multivariate methods. This paper will be restricted against the three other second year applied statistics papers. The Department of Food Science has advised they have 40 students who will enrol in this new undergraduate course. However care will have to be exercised with course content as there will not be room to teach all the desired topics.

A recent forum, Statistical Training and Advice for Postgraduates, has recommended a return to a regular workshop series run by statisticians according to topic and availability of relevant experts. Each workshop will be two or three days long, at a time of year that suits as many staff and students as possible. To help formulate appropriate workshops, programmes at three other universities have been reviewed including:

University of Reading Short Courses (http://www.reading.ac.uk/ssc/n/resources/Docs/Publicity/SSC%20short%20course%20brochure%202012.pdf),

Lancaster University Statistics Workshops (http://www.lancs.ac.uk/fas/cpd/statistics/)

Consulting Centre courses at the University of Melbourne (http://www.scc.ms.unimelb.edu.au/courses.html).

At this stage in addition to the earlier workshop on *Multivariate Statistics* and one proposed on *Time Series* we plan on the following workshops being available on a regular basis:

An Introduction to R:
Learn the basics of R, with the material going beyond that in a two-hour workshop run by our Information Technology Service

Statistical Modelling 1:
Learn the basics of statistical modelling in R beyond what is provided in a standard introductory course. The material will include linear models, generalised linear models, model selection and model checking.

Statistical Modelling 2:
A follow-on from Statistical Modelling 1, covering more advanced topics such as hierarchical models, generalised additive models and model averaging.

Bayesian Statistics:
An introduction to an increasingly popular approach to statistical modelling.

The cost of attending a workshop will vary between NZ$100 and NZ$150 per day to cover administrative expenses and time preparing the workshop. The courses will be viewed as service to the university rather than being built into to teaching load for a staff member. Surplus income generated will be held by the Department of Mathematics and Statistics and used to assist the research of those presenting the workshops by helping fund conference attendance, by supporting post graduate statistics students, by building collaboration with colleagues in other disciplines and by joint publication using workshop data. It is recognised that PhD students will usually spend three years minimum at the University and if workshops are spread over a two or three year period it may be possible to combine four workshops into a special topic course in statistical methods specifically for researchers. This may help students fund attendance with costs being built into scholarships and spread over three years.

4. The importance of context

Student feedback from questionnaires emphasise context. This is also reported by Finch & Gordon (2010) and Cobb (2007) who stated that “context is an essential part of statistical thinking, and some of the worst teaching of statistics occurs when the teacher or textbook tries to treat context as irrelevant”. Appropriate examples can be selected depending on backgrounds of students in the workshop. I have been fortunate to have dealt with consultations from researchers in different disciplines and this has provided an interesting and modern set of case studies which I have used in my teaching. The first group of studies following involve regression modelling and sampling.

Parackal, Ferguson & Harraway. (2007). This research investigated low iron levels in New Zealand infants, a major health problem in New Zealand. Cluster sampling resulting in 320 respondents, regression models and confounder control related to the presence of an infection were used in the research.

Parackal, Parackal & Harraway. (2010). The sampling procedure involved 1800 telephone interviews and analyses applied logistic and multinomial regressions to investigate profiles of New Zealand women who would consume alcohol when pregnant. This research was funded by the New Zealand Alcohol Advisory Council.

These two studies and several others published more recently as part of this research project are relevant to a workshop on regression modelling for human nutrition students. Earlier studies relevant to a regression workshop for ecology students involved interactions between dolphins and humans (Bejder, Dawson and Harraway, 1999) and dolphin habitat selection (Brager, Harraway and Manly, 2003). This year, 2013, I am supervising a PhD student in Marine Science who is researching the behaviour of dolphins in three bays in the Red Sea on the coast of Egypt. Three types of tourist activity are being analysed, one activity in each bay.

Another project has involved a postal survey of people in a city to establish the attitudes of the citizens to the use of public money to pay for an expensive sports stadium completed in 2011. There were 1800 respondents to a postal survey of 5000 residents which asked if they supported the stadium, opposed it or were neutral. The predictor variables were sex, age, education level, household income and whether a respondent was a ratepayer. Logistic regression models provided answers.

A second set of consultations provides data for a multivariate statistics workshop.

Harraway, Niven, Peake & Weege (2012). This study investigated ginseng origin. A trace element analysis produced 28 elements important for identifying the country of origin of ginseng samples. Principal Components reduced the dimensions and discriminant analysis followed by a re-sampling procedure clearly identified country of origin which has important political implications. Similar trace element studies have been carried out on honey, oysters and other New Zealand sea foods.

Harraway, Broughton-Ansin, Deaker, Jowett & Shephard. (2012). This study defined four new factors, based on Dunlap’s 15 NEP scales, to investigate student attitudes to sustainability. A confirmatory factor analysis confirmed our four new factors. Subsequent research investigating how these four factors change during a student’s time at university has also been published. Multinomial logistic regression and mixed models for sequential data analysis over a three year period have been used. Aspects of the research have involved cluster analysis and multidimensional scaling.

In another study, recently reported from a local agriculture research station, multidimensional scaling (MDS) was used to group samples based on soil type. There are 10 sites and for each site soil chemistry information is available on the amounts of different elements at four depths in the soil profile. For each level of the soil profile a MDS map shows the similarities between the sites. Looking at the maps from the four profiles there appeared to be a trend with two of the sites clear outliers and a group of sites tending to cluster together. Grouping the sites made further analysis easier because rather than having parameters for 10 sites only five were needed.

Another example produced by a student from Botany attending the multivariate statistics workshop investigated the growth of vegetation at sites in Thailand as a result of different types of prior land use and vegetation clearance.

5. Software

Another issue in teaching the workshops is the type of software to be used. There is a preference for open source R but many students depending on their area of study find it easier to use menu packages for some of their analyses. The software to be used should be established through prior discussion with those attending the workshops. Feedback would suggest that, in addition to R, SPSS should be used for work in the social sciences, STATA for work in human nutrition and other health related areas, SAS for finance and GenStat for agriculture. There could be a debate about whether the statistician has a responsibility to promote R.

6. Conclusions

Focussed workshops appear to be the best way of conveying statistical information to researchers in other disciplines once they are in a research programme. These students usually have not had room, or even the motivation, during their undergraduate study to include anything in statistics beyond a first year methods course. If they do, a programme which includes a minor in statistics will be beneficial later. A minor in statistics is a set of five statistics papers with at least three at second year level and one at third year level. The content of these five courses will cover all that is taught in several focussed workshop. An alternative possibility for consideration is to set up regulations which allow four workshops to be counted as a postgraduate course in advanced statistical methods for research. This may help funding if approval can be gained.

Context helps motivation and the workshops introduce new data sets which can be used in other courses and which can result in joint publications for the statistician carrying out the instruction. Case studies should, however, be recent and hopefully from the experiences of the teacher. This provides an argument for locating the teaching within a unit directly involved with consulting in a Department of Statistics. The workshops which are enjoyable to teach with such a group of dedicated people are not built into a teaching load of a statistician but can be used to assist the research of the statistician.

Workshops should be tailored to specific groups of researchers whose needs may differ. For example a workshop for social scientists will be different to a workshop for ecology researchers or a workshop for nutrition and food science researchers. This is the emphasis on context again.

References

Bejder, L., Dawson, S.M. and Harraway, J.A., (1999). Responses by Hector's dolphins to boats and swimmers in Porpoise Bay, New Zealand. *Marine Mammal Science* 15(3): 738-750.

Brager, S., Harraway, J.A**.** and Manly, B.F.J**.** (2003). Habitat selection in a coastal dolphin species (*Cephalorhynchus hectori). Marine Biology* 143: 233-244.

Cobb, G. (2007) One possible frame for thinking about experimental learning. *International Statistical review*, 75(3), 336-347.

Finch, S. and Gordon I (2010) Lessons we have learned from post-graduate students. In C. Reading ed. *Data and context in statistics education: Towards an evidence-based society*. Proceedings ICOTS8, Slovenia. International Statistical Institute. Voorburg, The Netherlands.

Harraway, J.A., Manly, B.F.J., Sutherland, H. and McRae, A. (2001) Meeting the statistical needs of researchers in the biological and health sciences. *Training Researchers in the Use of Statistics*. C. Batanero ed. Granada. International Association for Statistical Education and International Statistical Institute p 177 –195.

Harraway, J.A. and Barker, R.J. (2005) Statistics in the workplace: a survey of use by recent graduates with higher degrees. *Statistics Education Research Journal*, 4(2): 43-58

Harraway, J., Broughton-Ansin, F., Deaker, L., Jowett, T. & Shephard, K. (2012) Exploring the use of the revised New Ecological Paradigm Scale (NEP) to monitor the development of students' ecological worldviews. *The Journal of Environmental Education.* 43(3), 177-191.

Harraway, J.A., Niven, B.E., Peake, B. M. and Weege, B. (2012) Statistical Considerations in Developing a Trace Metal Signature to Distinguish Ginseng grown in Wisconsin (USA) from Ginseng grown in Canada, China and New Zealand. WorldCongress of SQ Foods-2012. Shenzhen, China.

Parackal, S., Ferguson, E. and Harraway, J.A. (2007).Alcohol and tobacco consumption among 6-24 months post-partum New Zealand women. *Maternal and Child Nutrition*. 3 (1), 40-51.

Parackal, S., Parackal, M. and Harraway J.A. (2010) Warning labels on alcohol containers as a source of information on alcohol consumption in pregnancy among New Zealand women. *International Journal of Drug Policy*. 21, 302-305.