STATISTICS FOR POSTGRADUATES AND RESEARCHERS IN OTHER DISCIPLINES: CASE STUDIES AND LESSONS LEARNED

John A Harraway
Department of Mathematics and Statistics, University of Otago, New Zealand
jharraway@maths.otago.ac.nz

Postgraduates and researchers in many disciplines use advanced statistics procedures. Statistics backgrounds often extend to at most an introductory course on statistical methods. Effective ways of providing training in these advanced procedures must be found. Emphasizing content, prerequisites and target groups, a summary of specialized courses offered at this level over the last two years and advertised internationally is presented. Then local four day intensive workshops on advanced topics for ecologists are described. These workshops draw on research contexts familiar to participants and use appropriate software. Menu driven packages or self written programs may be used. Participants in the workshops can bring their own data or data are chosen from their discipline. The teacher is introduced to the researchers which may result in future collaboration. Student evaluations of the workshops are reported leading to recommendations for further training.

INTRODUCTION

Statistics is a subject which becomes interesting and fully motivated for research students and others in employment when they have to use statistical procedures in their own applications and research. It is common for users of statistics in subjects other than statistics to have gaps in their knowledge even after they have completed their degrees at University. This can be caused by limited room in university course work for inclusion of statistics training beyond the elementary level as well as extensive demand on supporting mathematical theory sometimes perceived as necessary before dealing with data. But this data in the context of the student's research problem invariably provides an appreciation of the importance and relevance of the statistics and encourages the student to embark on further study.

A survey of 913 research graduates in employment with PhD and Master degrees in the biological sciences, psychology, business, economics and statistics from all New Zealand Universities (Harraway & Barker, 2005) identified gaps between techniques learned at University and techniques used in the workplace by these graduates. The respondents reported deficiencies in statistical preparation and made a series of recommendations about how to remedy the gaps. There was a consensus for focused workshops to be developed by education establishments or industry on a wide range of topics including regression and generalised linear models, multivariate methods, survey design and power analysis, clinical trials, new statistical software and Bayesian methods. Methods proposed for implementing this process included the expansion of advanced statistical service courses within universities, the development of specialist workshops in appropriate contexts for groups of postgraduate students in these areas and workplace retraining.

Some taught workshops are discussed in the next section. This includes a description of several sets of workshops provided on the web during 2009. One of the programmes involves a complete set of courses which either could lead to the equivalent of a degree in statistics or could have courses chosen selectively to support specialist subjects. Other programmes described are provided either in the home university of a particular group of students or the workplace. In the third section five workshops taught for ecologists within their university will be discussed in detail and accompanied by course evaluation comments by those participating leading to a summary of lessons learned for the future.

The courses or workshops provided outside regular university teaching which are aimed at promoting professional development for researchers appear to fall into one of three categories; those that are introductory, those provided for researchers in a wide range of applied areas where only one introductory course has been taken and thirdly those that represent an advanced specialised topic often aimed at the statistics specialist or a person in another subject who has experience in some advanced or specialised statistics activity. Generally, a mathematical background is not needed but for some of the advanced courses facility with theoretical statistics is a benefit if not essential.

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 [© 2010 ISI/IASE]

A REVIEW OF STATISTICS WORKSHOPS AND ON-LINE COURSES

This review is not an attempt to provide a complete report on statistics courses on offer by way of workshops in the two years beginning January 2009. The aim is to give a sample overview on what is on offer as a way of identifying appropriate material for inclusion in future workshops. There are differences between those courses taught to specialist groups of students within a university and those offered to a general audience through, for example, the web as part of professional development. There is usually a substantial charge for courses of both types although those aimed at students within a home university generally do not cost as much and sometimes may be built into a research programme or seminar series.

The Postgraduate Statistics Centre at Lancaster University has a list of professional development courses at different levels on offer for 2010. (www.maths.lancs.ac.uk/psc) For a general audience a series of two day courses include an *Introduction to Bayesian methods*, *Data Mining*, *Structural Equation Modelling*, *Methods for Missing Data*, *Generalized Linear Models*, *Genomics*, a variety of courses on R software, STATA and SPSS and an *Introduction to Statistics for the Life Sciences*. Three day specialist courses for statisticians cover *Pharmacological Modelling*, *Survival Analysis*, *Adaptive and Bayesian Methods in Clinical Research* and *Statistical Methods for Ordered Categorical Data*.

In Australia many Universities include workshops for target groups of researchers. For example, at least six workshops of two days duration on aspects of the package R at the introductory or intermediate level and two workshops on Bayesian methods at the intermediate level have been provided. Ten workshops at the intermediate or specialised level have been taught on areas related to epidemiology, clinical trials, survival analysis and longitudinal data ranging from one to three days. These courses have been provided by schools of population health or epidemiology at various universities such as the National Centre for Epidemiology and Public health at the Australian National University, the School of Population Health at the University of Melbourne, Deakin University, the University of Queensland, the University of Sydney and Queensland University of Technology.

The Statistical Consulting Centre at the University of Melbourne has offered two four day introductory courses on *Design and Analysis of Surveys*, *Design and Analysis of Experiments*, a six day course on *Statistics for Research Workers* and a one day course on *Producing Excellent Graphics Simply* (www.scc.ns.unimelb.edu.au/courses.html).

The University of New England at Armidale is teaching a three day course in January 2010 for postgraduate students and other professionals on *Application of evolutionary algorithms to solve complex problems in quantitative genetics and bioinformatics* and a five day course in February 2010 on *Bayesian methods in genome association studies* (jvanderw@unu.edu.au).

The Institute of Health and Biomedical Innovation at Queensland University of Technology has taught a one day course for intermediate level researchers on *Multivariate Data Analysis and SEM using SPSS and AMOS* (www.ihbi.qut.edu.au).

The New Zealand Social Statistics Network (www.nzssn.org.nz) provided in January/February 2009 and is providing again in 2010 a set of 10 five day and one four day introductory workshops in the School of Government at Victoria University of Wellington. Course titles include *Data Analysis using SPSS*, *Research Synthesis for Policy and Practice*, *Advanced Analysis of Linked Data*, *Qualitative Research Techniques*, *Introduction to Survey Design*, *Using Mixed Models in Research and Program Evaluation*, *Introduction to Structural Equation Modelling using AMOS*, *Introduction to NVIVO*, *Introduction to Case Study Design* and *Introduction to Survey Design*.

A four day advanced workshop *Modelling Patterns and Dynamics of Species Occurrence* (http://www.proteus.co.nz/home.html) was provided by Proteus Wildlife Research Consultants at the University of Otago.

Two introductory two day short courses, *Basic Statistics/Analysis of Variance* and *Simple Regression and Analysis of Covariance*, have been offered by Saville Statistical Consulting Limited (savillestat@gmail.com) at both Rotorua and Lincoln in New Zealand.

In the United States, Statistical Horizons (www.PaulDAllison.com) provided repeat presentations at different locations. Three continuing education workshops, *Longitudinal Data Analysis*, *Missing Data and Survival Analysis Using STATA*, were two days duration at an intermediate to advanced level. Two five day workshops, *Event History and Survival Analysis* and *Categorical Data Analysis* were also at an intermediate to advanced level. These courses will assist with professional development in the relevant areas of statistics use.

There are many programmes which provide distance learning through internet workshops. Details of one set of courses can be found at courses@statcourse.com At the introductory level there are *Creating Effective Graphic Presentations*, *Learn Statistics Through Applications* and *Introduction to R* each of which involve about 30 hours work over three weeks. At the intermediate level are *Power and Sample Size Determination*, *Applying Resampling Methods* and *Modelling with R* also taught for 30 hours over three weeks. A *Manager's Guide to Design and Conduct of Clinical Trials* is a four week specialised course of 40 hours.

Another extensive set of on-line workshops and courses for professional development in statistics and areas of application of statistics can be found at statistics.com. It is possible with these courses to construct a programme in advanced statistics study equivalent to a degree in statistics or just register in specific courses as desired for one's own specialty. There are over 100 courses on offer with most being four weeks duration. They range from the introductory/beginner level through intermediate level to advanced/specialist level. As well as introductory statistics and a set of advanced course on statistical method they include courses relevant to the life sciences, engineering, the social sciences, the environment, business and the statistical package R. It is stated who the courses are aimed at. For example, a course on *Structural Equation Modelling* is aimed at market researchers, education researchers, sociologists, psychologists, political scientists, economists and survey researchers, a course on *Advanced Logistic Regression* is aimed at the life sciences, business, social sciences, environmental science and engineering and a course on *Cluster Analysis* is aimed at market analysts, computational biologists, environmental scientists and IT specialists.

This brief review of courses for continuing education and professional development provides confirmation that particular methods tend to be used in different specialties (Harraway et al 2001) and consequently the review is a guide to how retraining workshops should be targeted and context developed.

THE LOCAL EXPERIENCE

The workshops now described are aimed at a group of ecologists, internal student researchers at the University of Otago rather than external students. The Ecology Research Group at Otago is a diverse group involving thesis Master and PhD students as well as staff members from a range of subjects including Botany, Zoology, Marine Science, Geography, Geology and Chemistry. Each year for the last seven years a sum of money has been available from the convenor of the Ecology Programme to help researchers in this group. Each year the students have decided to fund an intensive three or four day workshop on selected statistics topics related to procedures they are currently using or wanting to use. The topics covered have reflected closely the requested procedures for the workplace listed in Harraway and Barker (2005).

In 2002 and again in 2004 a Statistical Ecologist from outside the University of Otago was invited to teach workshops on multivariate statistics. The workshops were instructive but difficult with an emphasis on some of the underlying mathematics and limited hands on experience for the participants. The number of people attending was around 50 but response was not totally positive. The courses, based on a textbook being developed by the presenter, were scheduled out of term time and not in the summer as many of the ecologists at this time would be out in the field.

In 2005 I was instead approached to teach a course on multivariate statistics with a new group of research ecologists as those attending the previous courses had for the most part completed their study. Before agreeing to teach this workshop I outlined my plan to see if it would be acceptable. My main objective was to produce a group of researchers who would be capable of using the techniques in their own work. To achieve this I explained that half the three day workshop would involve lecture sessions of one and a half hour length and half would involve hands on laboratory sessions using data which I had accumulated from my consultations with marine science and zoology students as well as data generated by local students or other widely available data in the ecology context. I said that participants could bring their own data for analysis. I advised that SPSS would be used for analysis and even if students had never used it before they would have no problems with the package. I said the workshop would be limited to 30 participants, the capacity of an advanced computer laboratory, and the cost per person would be $250 to cover a workshop book and provision for morning and afternoon teas, but mainly tutor help in the hands on session where the staffing ratio was to be one tutor for ten students. Students were expected to have passed an introductory statistics course including an introduction to probability, hypothesis testing, confidence intervals, simple linear regression, analysis of variance and the basics of experimental design. There would be no mathematical demands other than simple algebraic manipulation. These proposals were accepted and the class was taught at the end of formal classes when a laboratory was available and before many of the researchers left for their field work.

Course content included multivariate analysis of variance, principal component analysis, factor analysis, both exploratory and confirmatory, discrimination including logistic regression if categorical predictors present, cluster analysis, scaling, canonical correlation analysis and correspondence analysis. The class was a pleasure to teach, feedback was very positive with high course ratings and another workshop was requested. Feedback included comment that the workshop was challenging and more hands on time would be desirable in a future workshop. There was a consensus that four days would be better for such a course. Participants were happy with the use of SPSS although it should be noted that this was taught in 2005 and the package R is receiving more attention now. The most encouraging comment came from a staff member in Zoology who advised a week later: “I have reanalysed one of my data sets and as a consequence of your workshop I have just submitted a new paper to a journal.”

The following year 2006 I was approached again to teach a workshop and after discussion a four day course on Generalised Linear Modelling was proposed and accepted. SPSS was used. Given the statistics backgrounds of the participants the teaching began with a short review of simple linear regression. Topics then covered were multiple linear regression, analysis of covariance, logistic and multinomial logistic regression including discussion of over dispersion for binomial data, log linear models, model selection and on the fourth day generalized linear models with appropriate link functions. Participant evaluations were again positive but the last day moved into some relatively advanced ideas only made possible by careful use of the SPSS package. At least the ideas were introduced but there was a request for more time to be spent on this last aspect of model building in future courses.

In 2007 a new ecology PhD student approached me with a request for a repeat of the workshop on multivariate statistics. But she added that her colleagues had asked for repeated measures and longitudinal data to be discussed at some point in the workshop. A slightly abbreviated version of the first multivariate workshop was prepared and on a fourth day the morning lecture and hands on session dealt with repeated measures data using the traditional analysis of variance approach including a discussion of compound symmetry with its consequential problems and correction. This helped motivate the afternoon sessions which then developed the approach through mixed models with linear fixed effects made possible by the use of SPSS. Feedback was again very encouraging but more time could have been spent on the mixed models.

In 2008 we were approached to develop a workshop which would introduce the ideas of Bayesian statistics. Another Department member with specialist interests in statistical ecology taught a three day course based on the statistical package Python. Pre requisite for the course was again an introductory statistics course with limited mathematics. On the first day lectures 1 and 2 reviewed probability and a range of probability distributions supported by use of Python in the two hands on sessions. On the second day lectures 3 and 4 reviewed model fitting using sums of squares and statistical likelihood with support from Python. The third day was devoted to Bayesian analysis, Markov Chain Monte Carlo and model selection. There was universal agreement that content was presented clearly and the hands on sessions enhanced workshop experience. All felt that the workshop helped an understanding of statistical inference. Two students found the material too advanced. There was agreement that such courses were needed and requests were made for Bayesian modelling, more multivariate analysis and model fitting in general to be topics for future workshops.

In 2009 with a new group of ecology researchers I was again asked to repeat the workshop on multivariate statistics. The course from two years earlier was updated with new examples and delivered again over a four day period. The limit was again 30 participants which meant ten students had to be turned away. On this occasion it was emphasised that students could bring their own data for analysis in the hands on sessions. This made the hands on sessions complicated but worthwhile. Several longitudinal data sets were produced on the last day and help with these was given the following week by one of the tutors building further links between our department and researchers in Zoology.

Three participants produced data from three research projects which I helped analyse during the workshop and over the following weeks. This identifies an important consequence of workshops when presented to students on campus. The statistician presenting the workshop may enter into interesting joint work with researchers in other areas similar to what can occur in a university statistical consulting unit. One of these three projects involved sex discrimination in native frogs which was the doctoral topic of a Zoology student. This project set out to classify native New Zealand frogs as male or female based on a range of measurements which would as a consequence avoid having to kill frogs in the wild to determine sex. The second project for a master degree in Botany investigated the regeneration of vegetation after various methods of land use and land clearance. The student was using the program CANOCO for ordination, principal component analysis and de-trended correspondence analysis. The third example from a project in archaeology involved the classification of 6000 year old pots of various shape and size excavated in Thailand with a view to comparing civilizations in the different parts of this region of South East Asia. Principal components and some factoring produced results which allowed comparisons. This classic example produced data similar to that reported in Manly (2007).

The course evaluation showed that 95% of the class found the teaching very effective. Four days with sessions ranging from 0900 to 1700 each day was said to be exhausting. Several requests were made for future help by way of consultation. Half the class asked for a future workshop on linear mixed models and longitudinal data. Two students asked for a workshop on the use of R and one asked for CANOCO to be used in a future workshop.

CONCLUSIONS AND RECOMMENDATIONS

There is an unsatisfied demand for carefully structured workshops on specialist and advanced statistics topics for research students in many subjects at universities. These should be taught in context because each discipline can make different demands on statistical methodology. Motivation is helped if studies related to the areas of interest of a particular group of students are used. Harraway et al. (2001) analysed 3000 research papers in five subject areas and discovered different statistics methodologies were used in different subjects confirming that context is likely to be an important issue.

Courses for internal research students in a university and external internet courses for continuing education and professional development must be distinguished. The former are discussed here as the target group of this investigation is the on campus researcher in another discipline; the earlier lists of internet courses provide ideas for an on campus programme. It is essential for on campus workshops to have a large hands-on component with interesting data sets selected from local research the students attending the course are involved with. This may not be quite so convenient with web based courses.

Student researchers in other subjects are unable to attend taught undergraduate courses over a semester due to field work or other commitments such as laboratory work. A further major problem relates to the pre requisites for these courses required for an undergraduate degree. The pre requisites will invariably require a background of mathematics, not present in the backgrounds of students in other disciplines. Some would say that if you understand the mathematics the data analysis will come naturally. Evidence from the success of the focussed workshops indicates these courses can be taught without the mathematics provided good statistical software is available.

One way round these problems which appears to be gaining some traction at our university is to develop a set of four or five workshops that can be taught at convenient times over a period of three years which is a common period of study and research for a PhD student. These workshops could be chosen to form a course on research methods which might make funding easier as it could be built into PhD scholarships. There would be no exam but assessment could be based on the submission of a portfolio of projects using the techniques taught and relevant data, maybe in some cases the student’s own data. Such a programme could be built into the teaching of a department or a consulting centre if there was funding for this available. It could even be used to justify staff appointments.

Statistics workshops for post graduate students are enjoyable to teach. But where should such programmes be located in a university? In a Department of Statistics the teaching will compete with scheduled undergraduate teaching and research demands on staff. In the research environment in universities extra teaching is seen as counterproductive to department and personal success through research. The ecology workshops described in the last section are viewed as non credit extra work. But should a Department of Statistics have an obligation to provide such teaching in a university? A positive aspect is that the statistician has contact with a wide range of applications which can be used in teaching standard courses in an undergraduate statistics programme. Consequently even introductory undergraduate teaching in statistics can be seen to be research based, a claim made about university teaching. Even more rewarding is the possibility of joint publication which will have large benefit for the statistician. Simply to be acknowledged in a paper provides little help for the statistician when it comes to promotion.

Another model is to develop the specialist workshop teaching in the framework of a consulting service within a university. This appears to be the model at Lancaster University and the University of Melbourne. This consulting activity is located hopefully as part of a Department of Statistics where staff employed in the consulting are built into the overall university statistical activity. At the University of Otago there is a Centre for the Application of Statistics and Mathematics although there is only one staff member in this unit funded by the Division of Sciences. He services the consulting needs of research students in the sciences but does no teaching. Should this Centre be expanded with staff having dual responsibility?

Feed back from over 100 students attending the ecology workshops has been positive. They have identified an on going demand for courses on specialised statistical topics to support their research. These courses must not be part of the regular undergraduate teaching of statistics and must be taught in such a way that the content can be understood with a background only of a first year introductory statistics course with no heavy demand on mathematics. These students have listed as desirable courses on mixed models for repeated measures, longitudinal data and survival analysis, Bayesian modelling and related software such as MCMC and the package R.

Under further consideration is a workshop on logistic regression, longitudinal data and survival analysis for Human Nutrition students and a course on multivariate analysis for Tourism and Social Sciences students. Both these workshops would have data relevant to the specialties; for example, alcohol in pregnancy for the first case and sustainability analysis for the second case. These courses would also be taught to fill the niche at the intermediate level rather than the beginner level, which is the assumed back ground of participants in such workshops.

REFERENCES

Harraway, J. A., Manly, B. F. J., Sutherland, H., & McRae, A. (2001). Meeting the statistical needs of researchers in the biological and health sciences. In C. Batanero (Ed.), *Training Researchers in the Use of Statistics*. Granada: International Association for Statistical Education and International Statistical Institute (pp. 177-195). Online: www.stat.auckland.ac.nz/serj.

Harraway, J. A., & 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.

Manly, B. F. J. (2005). *Multivariate Statistical Methods: a Primer* (3rd Ed.) Chapman and Hall/CRC.