TRAINING TO DEVELOP MODERN STATISTICS IN THE WORKPLACE USING R AND R COMMANDER - EXPERIENCES FROM THE NEW ZEALAND GOVERNMENT SECTOR

Ian Westbrooke[1] and Peter Ellis[2]
[1]Department of Conservation, New Zealand
[2]Ministry of Business, Innovation and Employment, New Zealand
iwestbrooke@doc.govt.nz

There is growing demand for robust, quantitative support for policy development and measurement of outcomes. In tandem, there are increasing challenges and opportunities from new ways of collecting data, often in massive quantities. Key modern statistical skills required in our workplaces include graphing and data exploration and visualisation; modelling, especially the linear model and its extensions, with emphasis on effect sizes rather than tests; designing and appropriate analysis of both small and "big" data; and application of statistical software/computing to support these. One strategy for dealing with these needs and challenges in the workplace is to increase skills of non-statisticians in the workplace. We describe our experiences in carrying training focussed on two contrasting government departments, and explain why and how we have used R and R Commander software as key elements.

INTRODUCTION

There is growing demand for more robust, quantitative support for policy development and the measurement of outcomes. This results from increasing interest in evidences-based policy and practice. At the same time, new methods of collecting data - often in massive quantities - are bringing increasing challenges and opportunities. These factors have led to an increased need for the application of modern statistics in the workplace, including data exploration and visualisation; modelling; dealing with "big" data and study design. Training in these aspects of statistics also requires the access to and training in the use of the statistical software and computing required.

We briefly review the literature on workplace training, where one approach proposed is to increase skills of non-statisticians in the workplace. We describe our experiences in carrying out training working in our organisations to enhance statistical practice. We needed to operate based on the workforce and statistical knowledge available. This involved up-skilling existing staff and training newly-recruited staff to fill in gaps in statistical knowledge. We explain why and how we have used R (R Core Team 2013) and R Commander (Fox 2005) software as key elements. We draw some conclusions, emphasising the importance of context, especially the use of real data and problems, and application to work projects; the use of tailored group and individual training and support; and moving from trainer-lead to mutual learning. We have found developing skills within the workplace a successful alternative strategy to relying on outsourcing for all tasks requiring modern statistics.

STATISTICAL TRAINING IN THE WORKPLACE

The statistics education literature focuses mainly on the formal educational sector, rather than the workplace. Most of the limited information relating to the latter is presented in the proceedings of the 4-yearly International Conferences on Teaching Statistics (ICOTS), which include the workplace as one of about nine topic areas in their programmes. However, the emphasis is mainly on how the formal education sector relates to the workplace, and so reports from within workplaces, especially in terms of training non-statisticians, are very limited. Some material on training non-statisticians in-house in the state sector is available (Hamilton 2010, Forbes et al. 2010, Westbrooke & Rohan 2014). Most other relevant literature comes from a more general perspective. Barnett (1990) looked at the requirements for meeting statistical needs in industry and suggested that either trained statisticians be employed; or the statistical skills in other staff be developed. A common thread in a number of the ICOTS papers is the importance of context for workplace training, with a need for emphasis on real data and statistical concepts and the use of projects and hands-on computing rather than on mathematical details, (e.g. Stephenson 2002, Francis & Lipson 2011). See Westbrooke & Rohan (2014) for more detail and references.

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

MODERN STATISTICS IN THE WORKPLACE

Research and technical staff in our workplaces need the following key statistical skills:

- Managing and manipulating data in ways that are efficient, reproducible and that facilitate peer review and quality control.
- Graphing and data exploration and visualisation.
- Modelling, with emphasis on effect size estimation rather than tests:
    - linear and generalised models and extensions, including mixed models;
    - tree models;
    - specialist areas, such as:
        - survival analysis;
        - seasonal adjustment;
        - animal abundance methods including mark/recapture;
    - small domain estimation.
- Dealing with increasingly large data sets and automated data collection.
- Designing and analysing observational studies and complex surveys.
- Statistical software/computing to support these.

Needs at the New Zealand Department of Conservation (DOC)

The first author became DOC's first dedicated full-time statistician in 2000 and found that by far the most effective role for a sole statistician was to spend a high proportion of time carrying out statistics education and advocacy. Science and technical staff had skill gaps in design, data collection, analysis and reporting. A wider range of staff also needed basic skills in effective data entry, management and exploration, including effective graphing. A key group of staff required training in statistical modelling skills, starting from the linear model and working through its extensions, including mixed models for handling longitudinal monitoring data. Initial emphasis was on tools to increase and improve data analysis. See Westbrooke (2011) for further information.

Needs at the New Zealand Ministry of Business, Innovation and Employment (MBIE)

The second author trained as a statistician but had worked mainly in the management of overseas aid, and became manager of tourism research at MBIE in 2011. At the time he managed a team of six to eight people who had two key roles: to produce and publish six to nine official statistics collections, and to fund three others produced by New Zealand's national statistics office; and to provide analytical support for policy teams. The previous Ministry of Tourism had recently been incorporated into a larger ministry which was soon after merged with three other agencies to become MBIE. The review leading to that merger had identified a need for improvements in tourism data, after several reviews in previous years had not led to substantive change. A five-year Tourism Data Improvement Programme was initiated in 2011 with the aim of improving the relevance and prioritisation of data collection and modernising the design, collection and analysis of tourism data. In the first two years of this programme two new major collections (Regional Tourism Indicators and Regional Tourism Estimates) were developed, based on electronic transactions data; and a major economic survey (the International Visitor Survey) was fully redeveloped with most of its data collection moved online.

To manage these processes and to increase the level of servicing for the tourism policy team and ministers, a substantial enhancement of statistical skills was required in the group. Staff were limited in their ability to carry out basic statistical tasks such as t-tests, linear regressions, and labelling points on a scatter plot. Therefore, an internal statistical seminar series was initiated to address some of these gaps. However, it then became apparent that there were major issues with the existing tools and skill set. A series of gaps emerged including in managing and manipulating data beyond one rectangle; estimating standard errors for complex surveys; taking an aggregate or summary dataset and using that as input for new analysis or plotting; and, of particular importance, producing quality graphics.

An important decision was needed on software. At the time the improvement programme was initiated, nearly all analysis was performed in Excel. Data was not stored conveniently, with several data sets being kept in a proprietary format; others kept in Excel format; and some not kept

at all. Both SAS and SPSS were part of the production workflow for processing statistics (SAS for outlier treatment, imputation and weighting of complex surveys; and SPSS for the final merging of two datasets). However, staff were reliant on programs written by specialists, so there was no capacity to troubleshoot significant issues with those programs; and the software was not used for any actual analysis or data presentation. Crucially, no staff had the skills to use SAS for appropriate estimates of standard errors in a complex survey; and the SPSS installation at the ministry did not include the (expensive) additional module that would allow the program to do this. In considering which software to use, all three of SAS, SPSS and Stata were experimented with before a commitment was made to R; mainly because of its graphic capacity, ease of performing simulations, and ability to manipulate many complex forms of data, as well as its price (free).

INCREASING SKILLS OF NON-STATISTICIANS IN THE WORKPLACE

One strategy for dealing with the increased demand for statistics in the workplace is to increase the skills of those who already have some, often limited, statistical skills. At both DOC and MBIE, we gained support for a strategy based on developing internal capacity in this way, rather than relying primarily on external support. Here, we describe our experiences, and explain why and how we have used R and R Commander software.

Increasing Skills at DOC

Westbrooke (2011) described the main steps taken to increase statistical skills at DOC, including developing courses on data handling and exploration; effective graphical visualization; and statistical modelling with two courses focusing on the linear/generalised linear model (lm/glm), and mixed models.

More recently, we have developed a course called “Designing Environmental Studies: Statistical sample design for observational studies and monitoring”, which has a strong emphasis on non-experimental studies, aiming to balance the overwhelming emphasis on experiments in almost all academic statistics courses. This course has reached a wide audience at DOC, and we have also presented it in Australia, (Sydney in 2011 and planned for Canberra in 2014). The content of this course highlights the importance of 4 Ws - the Why, What, When and Where of a study, with special emphasis on the why; and 3 Rs - Randomization, Replication and stRatification.

Another major change has been the adaptation of the first statistical modelling course to use R Commander, rather than typing code. While this 3-day course assumes that participants have been exposed to an entry level statistics course, it allows for no prior exposure to R or computer languages. Many participants found it overly challenging to grapple with both a range of new statistical concepts and to learn to type computer code, often for the first time. R Commander is an R package that allows users to create and run R code using menus that are similar to many point-and-click statistical packages. This is a useful tool for carrying out some statistical analyses, and, as intended by the package author, acts a bridge for learning to use R by writing code when the limits of R Commander are reached,. More detail, including screen shots of R Commander in use, can be found in Westbrooke &Rohan (2014).

The R community at DOC has developed considerably in recent years, with more than 70 users taking part in the most recent upgrade. Almost all general statistical work, beyond the limited functionality of Excel, takes place in R. A group of users at DOC are taking up further developments without specific training, with a wide variety of packages being used by different users. In particular, *reshape2* (Wickham, 2007) and *ggplot2* (Wickham, 2009) are being used increasingly for data manipulation and graphing respectively. A number of users have replaced the default R interface and text editor with the integrated interface provided by R Studio, however, unfortunately, it does not work well with R Commander. There are plans to look at making more user-friendly interfaces available to all R users with a likely solution being to make options of R Studio and an upgraded text editor like Notepad ++ available, as has been done at MBIE.

The years of training and support mean that there is now a community of users of statistics and of R at DOC, who can access resources and learn on their own initiative, and who interact with each other and with professional statisticians to provide mutual support and learning, in both informal and formal settings. This contrasts favourably with the situation at DOC prior to 2000,

and to that in some similar organisations that rely almost exclusively on a few specialists and external support for modern statistical input.

Increasing Skills at MBIE

MBIE provides an interesting contrast to DOC. At DOC the aim was to improve statistical skills across a broad range of science and technical staff incrementally over a decade or more, while in the MBIE tourism research team, there was a focus on transformational change in the work of one group driven by a particular urgent need in a few defined areas. The scale of the change being embarked upon was not appreciated immediately. Once the improvement plan was developed and the decision to move to using R was made, the change can be broken into 3 stages.

1. Toe in the water December 2011 to June 2012. The first stage was mostly about getting to know the new tool, R. There were material problems to solve, such as setting up R on the network; understanding the role of libraries and packages and where to store them; and learning how best to get advice from R's massive online community. Understanding how to read in data was an early obstacle. However teaching staff how to use read.csv for comma-separated variable (.csv) files provided a straightforward way to access data for example from spread sheets. During this time, training on statistical inference and analysis largely went on hold (although there some developments including how to use the survey package (Lumley 2012)) while the manager and staff struggled with the new software.

Overall this was a difficult period. Most staff were intimidated by coding and R was used as an inefficient specialist tool for only a few tasks, with lots of chopping and changing between applications. But promising signs included:

- Recognising the need for a customised consultant/trainer.
- Discovering *ggplot2* for high quality graphics. (Wickham, 2009).
- Writing our first custom functions for things like easy access to the ministry's standard palette of colours; automating the quarterly analysis of the International Visitor Survey using the *survey* package; and developing wrapper functions when using some relatively complex packages.

2. Commitment: July 2012 to April 2013. Four key developments occurred early in this second stage.

- A specialist trainer was used on a regular basis. Crucially, this led to the adoption of R Commander as an interim step to get people past the first intimidating moments with R, something that succeeded well beyond the expectations of the second author at least. While no substantive analysis was carried out by the team using R Commander, it was highly effective in encouraging staff to open up the software and learn what it could do. This was particularly important as there was little team history of successful code and limited collective knowledge and skills to draw on.
- On the trainer's advice the team adopted the *reshape2* package (Wickham 2007) for basic data manipulation. *dcast* was an instant hit, and much more understandable for people who were used to point-and-click cross-tabulation software than *apply*, *tapply*, and similar base R functions.
- A new database of electronic transactions data came into production and R -ODBC -SQL was the only way available of accessing it. This proved crucial in selling a code-based environment; and the staff members working on validating and analysing this dataset made big steps forward in their statistical skills, incorporating time series analysis and a range of basic inference tools into their work.
- The group was joined for the first time by an analyst chosen primarily for statistical computing skills. This was a significant change from the previous practice which had been to recruit tourism subject matter experts with moderate quantitative skills, and quickly contributed to a transformation of the team's work practices and raised standards.

3. Revolution: April 2013 to mid-2014. Training has moved increasingly to a mutual-learning model. The team became larger with more breadth of coverage and a new name: Sector Performance. As part of these changes more staff were recruited for their database or statistical skills, all of whom had an interest and willingness to learn. After a year of involvement, the external trainer's time reduced from 4 to 2 days per month for a further 6 months before being phased out. Monthly training sessions were split into basic - aimed those with little relevant experience and open to wider MBIE staff - and intermediate. Staff started to take turns presenting material at the intermediate sessions and at new weekly 30-45 minute standalone seminars. The weekly staff meeting included a 15 minute instant seminar and a mandatory "one thing I learned this week", each with a focus on applying R. A staff member commented that "it feels like being back at university". After 2 years, learning has become self-sustaining and the team can start to move beyond data manipulation to statistical modelling and inference. R has become the basic workbench for all nine of the team's members and a sustainable legacy of code created, but improvements and use of new techniques are still daily occurrences.

Wider Interest in NZ Government Sector

There has been wide interest in the developments at DOC and MBIE. More than 100 people expressed interest and over 60 attended a seminar that the two authors presented at an Official Statistics Seminar organised by the national statistical office (http://statisphere.govt.nz/seminars-training-forums/official-statistics-seminar-series/archived-presentations/r-language.aspx). This led directly to the formation of a local R User group in Wellington, and less directly to one in Christchurch.

CONCLUSIONS

Our experiences reinforce the emphasis in the literature on the importance of when carrying out statistical education in the workplace, with the use of real data and problems, and direct application to work projects. Training delivered in short courses outside the workplace had little success in building skills or furthering the changes in workflow needed, particularly at MBIE. Big gains come quickly when staff find that they can apply new skills and tools to carry out work tasks more easily, more quickly and often with a higher quality. This means that training needs to be customised for the workplace. It is important to integrate having participants carry out practical tasks, and as staffs' skills increase, to encourage them to share and demonstrate applications to tasks that arise in their work. In initial phases of learning, the role of a specialist trainer is likely to be very important, and learners will gain much from one-on-one support and coaching. A culture of sharing within teams can then lead naturally to a transition from trainer-led to mutual learning, with a reduced need for outside specialist input and support. Individual coaching and transition tools - such as R Commander - can play an important role in breaking down initial hurdles.

When an organisation or team identifies the need for increased statistical capacity, staff will need to change, and individuals will need to be open to learning and applying quantitative skills. We have found that some people are ready to take up these challenges, while others may choose to move to roles that require less of these skills. Changes in personnel can provide the opportunity for the recruitment of staff with the right background, aptitude and willingness, with depth of subject-matter knowledge taking a lower priority. Management need to be prepared for a capacity building programme of 2 or more years.

The combination of committed staff and management alongside appropriate approaches to cooperative team-based learning in the context of real-world tasks is a winner for increasing the uptake of modern statistics in the workplace. In preference to reliance on out-sourcing most statistical work, we have found a strategy based on statistical training and developing communities of users within our organisations, together with emphasis on recruiting staff with appropriate interests and capacity, an effective way to meet the needs for increasing modern statistical capacity in our workplaces.

ACKNOWLEDGEMENTS

We receive great support from the wider statistical community through consulting, receiving and providing specialist training, and the availability of resources such as R and its

packages (especially R Commander, *reshape2*, *plyr* and *ggplot2*); as well as more specialist software. We would like to thank the many statisticians and others who have helped with development of training, including Maheswaran Rohan at DOC.

REFERENCES

Barnett, V. (1991). Statistical trends in industry and in the social sector. In *Third International Conference on Teaching Statistics (ICOTS3): papers and abstracts* (pp. 440-445). Voorburg, The Netherlands: International Statistical Institute. http://iase-web.org/documents/papers/icots3/BOOK1/C8-3.pdf

Forbes, S., Bucknall, P., & Pihama, N. (2011). Helping make government policy analysts statistically literate. 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, http://iase-web.org/Conference_Proceedings.php?p=ICOTS_8_2010

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

Francis, G., & Lipson, K. (2011). The importance of teaching statistics in a professional context. 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. http://iase-web.org/Conference_Proceedings.php?p=ICOTS_8_2010

Hamilton, G. (2011). Statistical training for non-statistical staff at the office for national statistics. 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. http://iase-web.org/Conference_Proceedings.php?p=ICOTS_8_2010

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software, 9*(1), 1-19.

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0 www.R-project.org/

Stephenson, W. R. (2002). Experiencing statistics at a distance. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*. Voorburg, The Netherlands: International Statistical Institute. http://iase-web.org/documents/papers/icots6/4d3_step.pdf

Westbrooke, I. (2011). Statistics education in a conservation organisation—towards evidence based management. 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. http://iase-web.org/Conference_Proceedings.php?p=ICOTS_8_2010

Westbrooke, I., & Rohan, M. (2014). Statistical training in the Workplace. In MacGillivray, H., Martin, M., and Phillips, B. (Eds.), *Topics from Australian Conferences on Teaching Statistics: OZCOTS 2008-2012*. New York: Springer Science+Business Media.

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software, 21*(12), 1-20. http://www.jstatsoft.org/v21/i12/

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. New York: Springer.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software, 40*(1), 1-29. http://www.jstatsoft.org/v40/i01/