SIMULATION USING R FROM WITHIN EXCEL FOR TEACHING FIRST YEAR BIOLOGISTS

Gareth Ridall
Department of Mathematics and Statistics, University of Lancaster, United Kingdom
g.ridall@lancs.ac.uk

In this paper we describe the efforts that we are making to engage and impart statistical understanding to a large group of first year degree-level biologists. By using an online assessment system we have confirmed the extent of large areas of statistical misunderstandings. In this paper we describe our strategy for breaking down these statistical understandings using software that is free and can be run within an Excel worksheet.

INTRODUCTION

At Lancaster University the Mathematics and Statistics department runs a large number of service courses that include several for the Biology department. Furthermore we run a consultancy service where researchers from applied sciences, such as biology, can request free statistical help. Both of these activities are income for the mathematics department and are useful for stimulating fruitful collaborative research. Biology itself is a subject rich with statistical applications and at its best should be stimulating rigorous research and providing applications that can be incorporated into statistics teaching. However, both biological service teaching and the financial underpinning for the statistical consultancy service have come under threat in recent times, leading to a renewed imperative of a rejuvenated relationship with the biology department, requiring a new appraisal of how service teaching and consultancy is delivered to non-statistics majors. This paper concerns the author's experiences with a five week introductory statistics course consisting of over 100 first year biologists.

We describe the initial state of the course, the changes that have already been made and finally the proposed changes for this year (2009-2010). Furthermore, we outline the motivation for using resampling methods, in particular the bootstrap, and the tools that that we intend to use to implement it. Finally, we present our conclusions and put forth the implication for future collaborations' between the maths and biology departments.

WORK DONE SO FAR

In 2008 the author inherited the service course based mainly around descriptive statistics, sampling distributions, confidence intervals, regression, correlation and ANOVA. There were several examples taken from the biological field, however these tended to be weak in relevance and artificial in nature. The practical sessions for students consisted of two, four-hour sessions each week attended by approximately 60 students. Pen, paper and calculators were used but computers were not available. During the last hour of these practical sessions a weekly assessment test would be completed and submitted. The tasks set were largely algorithmic and mechanical in nature. In their end of module electronic evaluations, students acknowledged that the bulk of their learning was during these lengthy practical sessions.

In the first instance we made several changes. The first involved removing or 'pruning' topics and tasks that were not felt to be essential to the biologists' requirements. In this vein, the aim was to make the course more coherent. Many of the mechanical calculations, such as calculating the regression, correlation coefficients and F ratios were removed, and emphasis was shifted instead to the interpretation of these statistics. ANOVA was removed entirely, on the basis that the alternative approach of simultaneously calculating and displaying means with standard errors for each data groups, was deemed more appropriate for the users in terms of simplicity and understandable logic of the non-expert.

The second set of changes to the service course related to streamlining of the assessment system, due in most part to the problems associated with large student numbers. We therefore decided to implement an online assessment approach utilising the Question Mark Perception Assessment system for which the university has a license. Assessment was divided into three

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]

categories, as follows: 1) formative assessment where a series of hints were provided before the correct answer was given, 2) summative assessment where two tries to each student were given and 3) diagnostic assessment which provided the teaching staff with the a snapshot of the level and nature of understanding of the group as a whole. We then compiled a bank of diagnostic test items to check the understanding in known problem areas (Sotos et al., 2007).

Student participation has been extremely high in the non-compulsory formative assessment. Performance has been high in the summative assessment but an alarmingly large volume of common misunderstanding has been revealed. This lead to some consideration of the *meaning* of the apparent high level of attainment in the online summative assessment system. Good results could have been achieved through rote learning and perseverance rather than through a statistical understanding – a level of understanding that may still elude some of the students. Universal confusions were noted between concepts such as the sample and the population and between the distribution of the sampling statistic and the distribution of the sample. Moreover, there was widespread confusion between p-values, type II errors and the probability that the hypothesis is true.

It is clear that despite our changes, large areas of misunderstandings persist. However, the diagnostic feedback has provided us with a motivation to make more changes. We have talked to the biologists about the need to build some computer based statistical competencies into the course using an Excel based system. We must keep in mind these failures in understandings when designing activities for practical sessions

The final change implemented to the new service course, was to split the four hour practical of 60 students into two smaller groups of 30 with two sessions of two hours each. In this way the tutorial groups became much more manageable as a function of their size allowing for more time spent with struggling students. The assessment tasks were required to be completed out of class time using the online assessment system. These changes were received well by the students with an apparent gain in academic performance.

FUTURE PLANNING

The focus on future planning for the new service course, is aimed at enhancing the practical activities within the shortened tutorial sessions. Staff from the Biology department felt there was a need to provide the student with a basic ‘statistical toolbox’ that could be downloaded onto their own laptops and used free of charge. It is hoped that such a toolkit would be used by students in later on in their degree. We looked at several Excel add on packages including the INSTAT package developed by the University of Reading. INSTAT uses an Excel type interface and provides the tools for basic exploratory data analysis. It also includes many examples and an easy to follow tutorial. However getting students to use a statistical package does not necessarily lead to a gain in understanding the principles behind it. Indeed, there is much literature to suggest that simulation, both physical simulation (eg using dice or cards) and computer based simulation, can strengthen the statistical reasoning process (Chance & Rossman, 2006; Wood, 2005; Mills, 2002). The University of Glasgow has developed an R package called RPanel which can be used for demonstrating statistical ideas through animations (Bowman et al., 2006). Several people within the Lancaster statistics department have the technical skills to write the code to construct these animations.

However, consultation with the recipient biology department highlighted a desire to maintain computer based instruction using an Excel-type spreadsheet. One possible compromise would be to embed the R code that permits these simulations within an Excel document. The technology for this is already available and can be downloaded for free from the R repository called RExcelInstaller (Baier & Neuwirth, 2007). So, theoretically R type graphics and animations can be used within Excel. We hope to take this one step further whereby the students can use their own data within Excel to conduct simulation based inference using permutation tests or the bootstrap (Effron & Tibshurani, 1993). The idea of using resampling approaches for first year students is not a new one, having been around since the 1960s with software written by Simon and Bruce (1991). However this revolutionary idea has never made it into the mainstream of statistical education. The software is still available but it is not particularly ‘polished’ or of good value for using at present. On the other hand the running of the same resampling routines within R is very simple and is free.

R also has also the ability to animate these routines so that students can have a greater insight into the process.

We feel that the goal of using animated computer animations to teach students to reason statistically is now achievable. The Simon and Bruce (1991) idea can be seen as ahead of its time in the field of statistics education, but at the time was limited by the available technology.

One possible use of simulation is the use of randomisation or permutation tests These kinds of tests can be used to illustrate the logic of significance testing and the meaning of p-values. However we feel that even if the students do pick up the idea of p-value there is still a high probability that it will be interpreted incorrectly when it comes to actual data analysis. For example it will not help correct the confusion between the notions of scientific and statistical significance. It is for this very reason that in our course we do not intend to use the permuation test. Increasingly we have tried to take the emphasis away from significance testing and toward interval based inference.

A far more useful and general teaching tool is the bootstrap. Use of the bootstrap with animation offers many learning opportunities if used intelligently. A particular strength is that it enables students to grasp the part of the classical paradigm that insists that the sample is to be considered as just one realisation of many possible samples from a population. This is poorly understood and leads to a misinterpretation of how a confidence interval should be interpreted. Another advantage of the bootstrap is that it has the potential to correct a large number of statistical conceptions about the relationship between the sample and the population inference. A further advantage, and in contrast to conventional techniques, is that little in the way of mathematical skill is needed to use it. It is not true that the bootstrap is assumption free. Like any other type of statistical reasoning the bootstrap involves the making of assumptions. In this case, assumptions about the relation of the sample to the population and in the case of the parametric bootstrap, the distribution of the sample itself. However the user of a resampling package should be given control over these assumptions and experiment with them.

CONCLUSION

The task of imparting meaning and understanding to a large group of biologists and non-mathematicians is not a trivial task. We must make use of new technology in the task of engaging large numbers of students. We have already strengthened the feedback systems by using Question Mark Perception. Unfortunately the feedback has indicated that widespread misunderstanding is widespread. In the next year we hope to address this lack of understanding by introducing both physical and computer based simulation by embedding R code within Excel and thus taking advantages of the strength of both packages.

ACKNOWLEDGEMENTS

The author wishes to thank University of Lancaster's Faculty of Mathematics and Statististics for the small CETL grant which enabled him to carry out the work whose results are discussed in this paper.

REFERENCES

Baier, T., & Neuwirth, E. (2007) Excel:: COM:: R. *Com Stats., 22*.

Bowman, A. W., Crawford, E., & Bowman, R. W. (2006). *rpanel: making graphs move with tcltk. R News*, *6*(4).

Chance, B., & Rossman, A. (2006). Using simulation to teach and learn statistics. *ICOTS-7*.

Effron, B., & Tibshurani, R. (1993). *Introduction to the Bootstrap.* New York: Chapman and Hall.

Good, P. (1984). *Permutation Tests. A practical guide to resampling methods for testing hypothesis*. New York: Spinger Verlag.

Heiberger, R., & Neuwirth, E. R. (2009). *R Through Excel*. Springer Verlag.

Manly, B. (1992). *Randomisation and Monte Carlo methods in biology*. New York: Chapman and Hall.

Mills, J D. (2002). Using Computer Simulation Methods to Teach Statistics: A Review of the Literature. *Journal of Statistical Education, 10*(3).

Simon, J. L., & Bruce (1991). A tool for every day statistical work. *Chance, 4*(1), 22-23.

Sotos, A. E. C., Vanhoof, S., Van den Noortgate, W., & Onghena, P, (2007). Students' Misconceptions of Statistical Inference: A Review of the Empirical Evidence from Research on Statistics Education. *Educational Research Review, 2*(2), 98-113.

Wood, M. (2005). The Role of Simulation Approaches in Statistics. *Journal of Statistical Education, 13*(3).