Proceedings of the 4th Applied Statistics Education and Research Collaboration (ASEARC) Conference

Edited by: Eric Beh, Laurence Park and Kenneth Russell

Parramatta, February 17-18, 2011

ASEARC

University of Wollongong

University of Western Sydney

The University of Newcastle Australia

Published by:
University of Wollongong (CSSM)
http://www.uow.edu.au/informatics/maths/research/groups/asearc/4thAnnResCon/index.html

ISBN: 978-1-74128-195-8

Credits:
Editors: Eric Beh, Laurence Park and Kenneth Russell
Cover design: Laurence Park

Printed in Wollongong by the University of Wollongong — February 2011

Welcome from the Conference Chairs

These proceedings contain the papers of the Fourth Applied Statistics Education and Research Collaboration (ASEARC) Conference hosted by the University of Western Sydney.
Of the 34 submissions, 16 were accepted as full papers and 18 were accepted as oral presentations. The full written version of each submission received one anonymous review by an independent, qualified international expert in the area. Dual submissions were explicitly prohibited.
We would like to thank the members of the program committee and the extra reviewers for their efforts. We would also like to thank The University of Western Sydney, School of Computing and Mathematics for their generous support of the event.

Conference Chairs

- Eric Beh University of Newcastle
- Laurence Park University of Western Sydney
- Ken Russell University of Wollongong/Charles Sturt University

Conference Organisation

- Stephen Hurley University of Wollongong
- Renee Mackin The Conference Team

Programme Committee

- John Best University of Newcastle
- Carole Birrell University of Wollongong
- Mark Fielding University of Wollongong
- Chandra Gulati University of Wollongong
- Robert King University of Newcastle
- Yanxia Lin University of Wollongong
- Maureen Morris University of Western Sydney
- Darfiana Nur University of Newcastle
- John Rayner University of Newcastle
- Paul Rippon University of Newcastle
- David Steel University of Wollongong
- Frank Tuyl University of Newcastle

Conference Sponsor

- Minitab `http://www.minitab.com/`

Conference Program

Day 1: February 17, 2011

Social and Economic Statistics

Education 1

Survey Methodology

Inference 1

Education 2

Day 2: February 18, 2011

Biometry

Young Statisticians

Inference 2

Statistical Consulting

Inference 3

Applications

Applying the Challenging Racism Project data to local anti-racism

Kevin Dunn

University of Western Sydney, Australia
K.Dunn@uws.edu.au

Jim Forrest

Macquarie University, Australia
jim.forrest@mq.edu.au

Abstract

Between 2001 and 2008, telephone surveys were conducted across the states of Australia to collect data on attitudes towards cultural diversity, ethnic minorities and the state of inter-ethnic relations in Australia. The survey also gathered data on the rates of experience of racism in nine different settings of everyday life. The final sample of 12,517 is a robust empirical base from which to demonstrate and analyze racism in Australia. An important finding has been on the geographic variations in the nature and prevalence of racism. Over the last three years we have been exploring the most effective means by which to use the data for the purposes of anti-racism. Given the core finding on racism as everywhere but different, our anti-racism deliverables need to be sensitive (usable) at local levels. This involves the public provision of data at regional levels. Most ambitiously we have used an entropy grouping procedure to help us provide anti-racism suggestions for types of regions. Our paper reflects on the success and limitations of our process for producing regionally-sensitive anti-racism resources, and on the applied utility of those deliverables. Key measures and issue have included: the extent to which the resources are sufficiently local; the technical language of our deliverables; the reliability of our samples (especially at local levels); the sensibility of our groupings; and the small p political sensitivity of the data.

Key words: Racism, anti-racism, statistics, entropy grouping, applied utility

Quality Assurance as a Means of Improving Statutory Mass Valuations (with special reference to Land Valuations in NSW)

John Macfarlane

University of Western Sydney, Australia
j.macfarlane@uws.edu.au

Abstract

Land and property taxes are widely applied at all levels of government as a means of generating revenue. They are an important component of a broad-based taxation regime. Most of these taxes are imposed on the basis of property values, requiring the generation of regular, cost effective, consistent and accurate estimates of value. In New South Wales, annual Land Valuations are produced for approximately 2.5 million properties. Approximately $5b in taxes is raised annually on the basis of these land valuations. This paper will describe the basic mass valuation model and discuss a variety of quality assurance measures which have been developed to improve the overall quality of land valuations. The level of improvement in the quality of land valuations will be considered and possible further developments will also be discussed.

Key words: Mass Appraisal, Statistical Modeling, Quality Assurance

Assessing and minimising the impact of non-response on survey estimates: Reviewing recent housing research in Australia.

Peter Phibbs

University of Western Sydney, Australia
p.phibbs@uws.edu.au

Abstract

A very common research method in recent Australian housing research is the use of a sample survey. Whilst many of these surveys are analysed using robust statistical techniques the issue of non-response bias is often overlooked. This paper examines survey research undertaken by the Australian Housing and Research Institute (AHURI) over the last five years and describes the survey method and the response rate. It identifies a number of surveys where the response rate has been less than 25%. The paper goes on to identify how to measure the bias in sample surveys from non-response and makes practical suggestions for when bias is most likely to be a problem in sample surveys and identifies some potential methods for reducing the impact of non-response on survey estimates.

Key words: Non response, bias, surveys, housing

Stepping into Statistics: An Online Head Start Statistics Program

Anne Porter

University of Wollongong, Australia
alp@uow.edu.au

Abstract

It makes no sense that students would enroll in a subject, pay fees and walk away with a zero, low mark and a fail grade. There are students who are inactive, do not come to class and do not submit assignments that deserve such a grade. How to engage these students has inspired innovations in an introductory statistics subject for the past fifteen years. Over that time innovations including the inclusion of real data, E-learning, aligning assessment and objectives, new assessment practices, learning design maps, a laboratory manual, video learning support resources, assignments based on real social issues have been implemented. At a graduate level the innovations with on campus and distance and international students have been highly successful with many learning outcomes including a low failure rate. The most recent innovations, a learning design map and competency based assessment have highlighted another group of students at the undergraduate level, who attend lectures and laboratory classes, are active in class answering questions, but who have difficulties with complying with the competency based approach to assessment, test and re-test after guidance. While students have almost unanimously endorsed the assessment system, approximately two-thirds of completing students indicated that they would like access to a head start program. This paper discusses one approach regarding the development of an electronic, head start module, which incorporates both learning support and discipline content. It canvasses the use of different types of resources, and different genres of video, mechanisms for facilitating communication between students, and production issues and issues in merging the head start program with the introductory statistics subject.

Key words: student class, teaching statistics, online learning

Improving Statistical Education through a Learning Design

Norhayati Baharun & Anne Porter

School of Mathematics and Applied Statistics, University of Wollongong, NSW, 2522, AUSTRALIA
nbb470@uow.edu.au

Abstract

This paper presents the results of a study examining the student learning experience of statistics within e-learning environment at the University of Wollongong. This study involves a cohort of 191 undergraduate students who enrolled in an Introductory Statistics subject in Autumn 2010 session. A learning design map was used within the subject e-learning site aiming at providing guidance which details out timing tasks and resources, and supports materials on week-by-week basis to students in learning the subject. The findings reveal the students gained benefits from the use of the map in helping them to learn and understand the subject; however they highlight some issues on the design of subject particularly within e-learning environment in terms of browser compatibility, file accessibility, map layout, and choices of design varieties. The paper concludes with a discussion on the needs of learning design in teaching practices and the learning of statistics and followed by suggestions for further research.

Keywords: learning experiences, learning design, e-learning environment

1. Introduction

Over the last decade, the use of Information and Communication Technology (ICT) in the higher education sector has been growing exponentially in supporting teaching and learning processes as well as providing benefits to students who can learn any subjects anytime and anywhere without limits [5]. As for teachers, the use of e-learning has created the potential to extend to new student markets, offering more flexible learning environments for students. E-learning provides the possibility of monitoring student progress and activity, as well as providing space for creating new and innovative learning resources. E-learning has much potential, but requires teacher commitment and many other resources. To benefit from this commitment means that the e-learning materials should not only be designed well and be student centred but also these resources along with adequate student support should be delivered appropriately [4]. It has been further suggested that an effective e-learning should involve high authenticity, high interactivity, and high collaboration [7].

With regards to the use of technology in teaching of statistics, it is evident that today, the use of the Internet, web-based courses, online discussions, collaborative tasks, electronic texts and associated assessment materials have changed the way statistic teachers work as well as what and how to teach [2].

There has been tremendous increase in research studies focussing on the utilization of technology in the teaching and learning of statistics over the past fifteen years. These studies span many technology tools used for many different purposes but they lead to one aim, the improvement of student learning of statistics. The focus in this study is on how to improve the effectiveness of e-learning so as to enhance learning outcomes for students. More specifically, it explores the role of Learning Designs as a means of more effectively delivering resources to students.

Learning design may be at the level of a whole subject, subject component or learning resources [1]. Learning design may be defined as a variety of ways of designing student learning experiences such as planning schedules, preparing or designing course or subject outlines and materials, determining assessment tasks, anticipating students' needs, implementing new learning strategies or modifying a previous course or subject by updating materials [9].

The study draws upon the experiences of a cohort of undergraduate students enrolling in Autumn 2010 session, of Introductory Statistics subject at the University of Wollongong. It examines the students' experiences on the subject design delivered and displayed via a learning design map integrated within e-learning site (*Blackboard Learning System*). The paper includes a description of the study method along

with the context and setting of the study, the results and findings, and suggestions for future research.

2. Method

In Autumn 2010 session, the subject presentation was redesigned based on the Learning Design Visual Sequence (LDVS) representation [8]. This has been developed from work by Oliver and Herrington [6] which focussed on three elements for designing an effective online learning environment that are: (i) the content or resources learners interact with, (ii) the tasks or activities learners are required to perform, and (iii) the support mechanisms provided to assist learners to engage with the tasks and resources. The LDVS extension involves illustration of the chronology of tasks, resources, and supports using symbols for each of the three elements of learning design. This together with a brief textual description summarizes the aim of the tasks and describes what learners are required to do. The LDVS represents a visual summary of a learning design from the perspective of the teacher, thus this project (Learning Designs Project; www.learningdesigns.uow.edu.au) produces generic learning design resources and tools for the purpose of helping teachers who might want to use the template for the design or modification of further teaching.

In this study, the three elements were put together and displayed as a flowchart or a learning design map illustrating the teaching and learning activities on a week-by-week basis (see figure 1). The aim of the learning design map was to provide guidance to students in learning this subject through variety of learning resources. As can be seen in figure 1, the students could access to the resources by a click on the links provided within the map. The primary resources included lecture notes, Edu-stream (audio recorded lectures); the tasks i.e. laboratory work, laboratory tests, worked solutions, data sets; other specific learning resources such as video resources to support most topics and SPSS notes; and ongoing support materials i.e. learning strategies, student forum, past exams and laboratory tests, student support advisers and learning modules.

The assessment system designed in Autumn 2010 session permitted the students to sit a mix of two in-class and two take-home tests, with a pass mark of 70%. Students who did not pass a test for the first time were required to sit a second version of the test (re-test) out-of class. Students were provided with feedback on their own work in the first test and were also provided with worked solutions. The re-test completed out of class involved a different data set, for the same topic but where possible different wording and sequencing of questions were used. A mark of 70 per cent had to be obtained on the re-test, but this was now the maximum mark possible. The tests, re-tests, sample tests, worked solutions, and feedback were provided to students as links that can be accessed through the map. This assessment system was useful it allows the lecturer to identify students who are at risk early. Of particular interest were students who did not attempt a particular test on a second occasion, or did not follow up despite direct feedback as to how to correct their answers was provided. The challenges turned out to be how to identify the type of help required and how to provide it. For example some students needed to find ways to communicate what they knew.

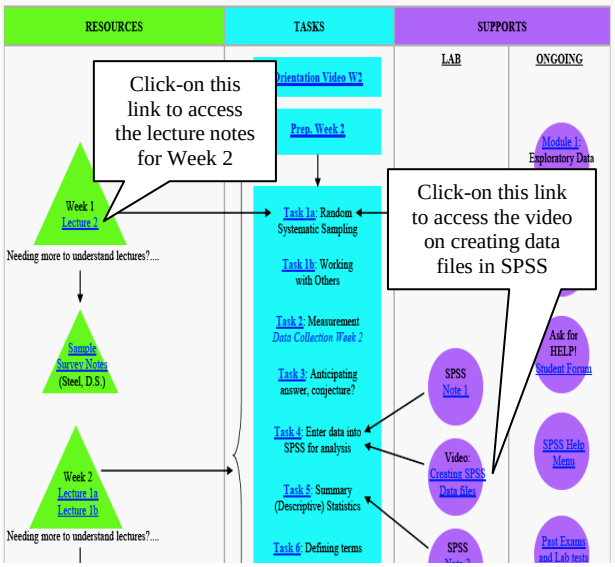


Figure 1. LDVS map for one week combining the resources, tasks and supports

A survey questionnaire was used to collect the background information on the students such as gender, their nationality either international or domestic, also included items on their expectations of the subject (anticipated grades). Other questions were on their use of the learning resources particularly the learning design map, their confidence in major topic areas, and suggestions on areas to be improved in the subject. The students were asked to complete this survey via e-learning site at the end of session (in Week 13) before their final exam.

Participants

From a cohort of 191 students, there was a large percentage of computing students (74%) enrolled in this subject. In regard to the survey, 109 students (57%) took part via the subject e-learning site. 19% of them identified themselves as being international students while the remaining 81% as domestic students. 17% of them identified themselves as females and 83% as males. In terms of grade

anticipation, 59% of them anticipated to achieve a credit or above credit at the end of session.

3. Results

Value of learning resources

In the questionnaire survey, the students were asked to rate each of the learning resources in terms of their usefulness in helping them learn and understand the subject. From previous experience, the authors expect the ratings of primary resources (lectures, assessment, laboratory tasks and worked solutions) to be high, above 80%, noting that the value of one resource changes with the improvement of another. Support resources such as student forum and other learning supports tend to be rated lower as they are not necessary learning aids for all students (see table 1).

Table 1. Rankings of learning resources in helping students to learn and understand the subject

	Moderate %	Extremely %	Total %
Worked solutions	23.4	73.0	96.4
Laboratory tests	42.3	47.7	90.0
Laboratory tasks	53.2	35.1	88.3
Lecture notes	45.5	39.1	84.6
Laboratory manual	45.9	36.9	82.8
Lecturer	56.8	24.3	81.1
Re-tests	30.3	45.9	76.2
Lectures	54.1	19.8	73.9
Tutor	38.7	33.3	72.0
Laboratory class	29.7	41.4	71.1
Learning design map	38.7	30.6	69.3
Group project	46.3	13.0	59.3
Video resources	28.8	20.7	49.5
Student forum	30.6	3.6	34.2

Specifically, the learning design map was valued high by the majority (69%) of students in helping them to learn and understand the subject. Comments indicated that some students (83%) appreciated the use of this resource because of the linking between tasks and other learning resources in the subject, the provision of references (77%), improved their ability to organize their work and update learning materials (71%), used for revision purposes (82%), and as a study checklist (62%). Examples of comments made on the use of the map were as given below.

"I used it to download the content as the map contained the links required"
"To gauge when I needed to study for the lab tests etc as they worked out to be a week or two behind"
"I used it to download the lecture notes and lab solutions and to see what parts link to where"
"As a learning compass"

However, there was room for improvement in the development of the map in the subject. Out of 83 students, a small percentage of them (17%) responded negatively (as stated below), for instance, many computing students shifted their files to another computer system and thus they dislike the multiple files used within the map.

"The materials were somewhat harder to access owing to browser issues, it would be useful if the files were also available without having to access the learning map, but rather to use the learning map as an alternate access/organizing method"
"It seemed harder to use. I prefer the folder design that has lectures in a lectures folder etc. However I do like the fact that each week has a subheading with the topics covered in that week which makes it a lot easier to revise each individual topic.."
"Technical issues make using it quite difficult. If access could be reliable I'm sure it would be good"

On the other hand, 6% of students were positive but experienced frustration or difficulties with technical issues, for example,

"The flowchart was very visually appealing but some of the links did not work properly on most computers and I had trouble accessing the information, other subjects however had a less visually appealing layout but I was able to access all the information. If the links to the information worked properly then the design map was a great idea"
"Quite good if it was integrated, using a pdf is not very effective as some browsers do not allow you to view the pdf's inside the browser and downloads it to your computer instead"

Similarly, there were also positive and negative comments made by tutors as stated below.

"As a tutor, the weekly map identifying tasks and associated resources was extremely useful. It was an incredibly useful organizer in the lab class, allowing download of the particular task either the teacher or students wanted to be worked together"
"It is better than normal teaching via e-Learning site because the teaching materials are more organized"
"It was good experience using the weekly learning design map, but it was slightly worse than normal teaching via e-learning site as it needs longer time to go through the tasks"
"I was a bit confused about the new structure used in the subject e-learning site, and I think the normal one was much easier to be used"

Confidence in the subject at the end of session

An analysis of topics indicates that majority of the students were confident in most topic areas at the end of session (see table 2). However, the students might need more resources, such as video supports and better lecture notes and worked examples particularly on the topic of normal and exponential distributions.

Table 2. Student confidence in major topic areas

	Can do %	Total % (Moderately confident & Can do)
Exploratory data	50.5	86.7
Correlation and regression	32.0	81.5
Binomial and Poisson	32.0	74.7
Confidence intervals	29.5	74.3
Using SPSS	34.9	72.6
Model fitting (Goodness of Fit Test)	30.8	71.2
Hypotheses tests	24.0	68.2
Normal and exponential	16.5	58.2

Assessment

An analysis of assessment results indicated the average marks achieved by students in Test 2 (out-of-class) and Test 4 (in class) were slightly higher than the other two tests (see table 3). These results may appear to contradict the perception of confidence in some topics highlighted above as this was observed at the end of session i.e. after the assessment has been carried out.

Table 3. Assessment marks in Autumn 2010 session

	N	Average	S. deviation
Test 1 (Exploratory data)	174	5.52	1.88
Test 2 (Correlation & Regression)	159	7.87	1.79
Test 3 (Probability & Models)	160	6.74	2.10
Test 4 (Hypothesis testing)	166	7.34	2.51

Final marks and grades

The final marks revealed an average mark of 59.73 with a standard deviation of 23.82 in Autumn 2010 session. The grade distributions were: 9% of the students achieved a high distinction (HD), 18% distinction (D), 26% credit (C), 24% pass (P), 5% pass conceded (PC), and 18% fail (F) grades. With respect to failures, the rate would have been much higher had the assessment system not allowed the identification of students at risk and the subsequent work with them to have them reach the standard required.

4. Conclusions

This study was in line with [1] principles of high quality learning, that is, effective learning designs should be based on the learners' perspective as in [8], "learning arises from what students experience from an implementation of a learning design". Whilst the learning of statistics has been associated with students having difficulties in learning and poor academic outcomes [10], this study examined the potential learning design particularly the use of learning design map within e-learning system as to improve the teaching practices and the learning of statistics.

In this paper, we were able to highlight the results on the students' experiences of the subject via the learning design map on e-learning system. The students commented that they gained benefits from the use of learning design map, however this study suggests there is a need to redesign the subject particularly within e-learning environment be more interactive, easy access, multiple browsers compatibility, choices of designs variety (folders system, learning design map, webpage, concept maps, subject content timeline) and better layout.

References

[1] D. Boud, M. Prosser, "Appraising New Technologies for Learning: A framework for development", *Educational Media International*, 39(3), 237-245, 2002.

[2] D. S. Moore, G. W. Cobb, J. Garfield, W. Q. Meeker, "Statistics education fin de siecle", *The American Statistician*, 49(3), 250-260, 1995.

[3] G. Ring, G. Mathieux, The key components of quality learning. *Paper presented at the ASTD Techknowledge 2002 Conference*, Las Vegas, 2002.

[4] M. Ally, *Foundation of Educational Theory for Online Learning. The theory and practice of online learning (2nd ed.)*, T. Anderson (Ed.), Canada, AU Press, Athabasca University, 15-44, 2008.

[5] R. A. Cole, *Issues in web-based pedagogy: A critical primer*, Westport, CT: Greenwood Press, 2000.

[6] R. Oliver, J. Herrington, *Teaching and learning online: A beginner's guide to e-learning and e-teaching in higher education*, Edith Cowan University: Western Australia, 2001.

[7] R. Oliver, A. Herrington, J. Herrington, T. Reeves, "Representing Authentic Learning Designs Supporting the Development of Online Communities of Learners", *Journal of Learning Design*, 2(2), 2007.

[8] S. Agostinho, R. Oliver, B. Harper, H. Hedberg, S. Wills, A tool to evaluate the potential for an ICT-based learning design to foster "high-quality learning". In A. Williamson, C. Gunn, A. Young, T. Clear (Eds.), *Wind of change in the sea of learning. Proceedings of the 19th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education*, Auckland, New Zealand: UNITEC Institute of Technology, 29-38, 2002.

[9] S. Bennett, S. Agostinho, L. Lockyer, L. Kosta, J. Jones, R. Koper, B. Harper, Learning Designs: Bridging the gap between theory and practice, In *ICT: Providing choices for learners and learning. Proceedings ascilite Singapore 2007*, 51-60, 2007.

[10] W. Pan, M. Tang, "Examining the effectiveness of innovative instructional methods on reducing statistics anxiety for graduate students in the social sciences", *Journal of Instructional Psychology*, 31(2), 149-159, 2004.

When the Teacher Stops Teaching: Supporting Undergraduate Statistics

Dr Maureen M. Morris

Student Learning Unit, University of Western Sydney, Campbelltown. AUSTRALIA
m.morris@uws.edu.au

Abstract

Many first year students find the study of statistics at university challenging and universities attempt to redress learning deficits by supporting students with written and online resources and, in many cases, teaching support. The question arises: what is the most effective allocation of time and resources in supporting students. Teacher experience and the literature identify assessment as a powerful driver for student learning. The teaching intent here is not to provide 'just-in-time' remediation, but rather to address more long term learning outcomes through the review and consolidation of fundamental concepts and procedures. This paper results from the development of workshop strategies and resources that target assessment outcomes and parallel classroom teaching. Teacher devised student notes and student developed summaries addressing the knowledge and skills required for impending assessment tasks underpin the action within collaborative workshops. Anecdotal evidence from the last session identifies increased engagement and positive student feedback. The impetus now is to convert the workshop-driven learning experience into engaging, self directed and online resources allowing teacher feedback. This would broaden student access and enable timely feedback when further remediation may be needed.

Keywords: statistics, remediation, support teaching, collaborative workshops

Small Area Estimation under Spatial Nonstationarity

Hukum Chandra

University of Wollongong, Australia
hchandra@uow.edu.au

Nicola Salvati

University of Pisa, Italy
salvati@ec.unipi.it

Ray Chambers

University of Wollongong, Australia
ray@uow.edu.au

Nikos Tzavidis

University of Southampton, United Kingdom
n.tzavidis@soton.ac.uk

Abstract

A geographical weighted empirical best linear unbiased predictor (GWEBLUP) of small area means is proposed and two approaches (i.e. conditional and unconditional) for its mean squared error estimation are developed. The standard empirical best linear unbiased predictor (EBLUP) under linear mixed model and its associated mean squared error estimator can be obtained as a special case of this approach. Empirical studies using both model-based and design-based simulation, with the latter based on two real data sets, show that the GWEBLUP method of small area estimation is superior to use of the EBLUP. The two proposed MSE estimators for the GWEBLUP work well. A real gain from the geographical weighted small area estimation approach is for out of sample areas, where the additional information contained in the geographical weights used in the GWEBLUP improves precision compared to standard methods.

Key words: Spatial nonstationarity, Geographical weighted regression models, Estimation for out of sample areas, Borrowing strength over space, Small area estimation

Outlier Robust Small Area Estimation

Payam Mokhtarian

Centre for Statistical and Survey Methodology, University of Wollongong, Australia
pmd952@uowmail.edu.au

Raymond Chambers

Centre for Statistical and Survey Methodology, University of Wollongong, Australia
ray@uow.edu.au

Abstract

In recent years the demand for small area statistics has increased tremendously and therefore small area estimation (SAE) has received considerable attention. The empirical unbiased linear prediction (EBLUP) estimator under linear mixed model is commonly used method of SAE (Rao, 2003). However, the EBLUP can be both biased and inefficient when data contain outliers. Sinha and Rao (2009) and Chambers et al. (2009) described outlier robust method for SAE using Huber weight function. We explore an alternative model parameters estimation for SAE in presence of outliers. Empirical results based on simulation studies show our approach leads to a more efficient small area estimates.

References

J. N. K. Rao (2003). Small Area Estimation. Wiley, New York.

S. K. Sinha & J. N. K. Rao (2009). Robust small area estimation. The Canadian Journal of Statistics, 37 (3), 381-399.

R. L. Chambers, H. Chandra, N. Salvati & N. Tzavidis (2009). Outlier Robust small area estimation. (unpublished paper)

Key words: Empirical best linear unbiased prediction, linear mixed model, robust estimation, small area estimation

Contextual Effects in Modeling for Small Domain Estimation

Mohammad-Reza Namazi-Rad

Centre for Statistical and Survey Methodology, University of Wollongong, NSW 2522, Australia
mohammad_ namazi@uow.edu.au

David Steel

Centre for Statistical and Survey Methodology, University of Wollongong, NSW 2522, Australia
david_ steel@uow.edu.au

Abstract

Many different Small Area Estimation (SAE) methods have been proposed to overcome the challenge of finding reliable estimates for small domains. Often, the required data for various research purposes are available at different levels of aggregation. Based on the available data, individual-level or aggregated-level models are used in SAE. However, parameter estimates obtained from individual and aggregated level analysis may be different, in practice. This may happen due to some substantial contextual or area-level effects in the covariates which may be misspecified in individual-level analysis. If small area models are going to be interpretable in practice, possible contextual effects should be included. Ignoring these effects leads to misleading results. In this paper, synthetic estimators and Empirical Best Linear Unbiased Predictors (EBLUPs) are evaluated in SAE based on different levels of linear mixed models. Using a numerical simulation study, the key role of contextual effects is examined for model selection in SAE.

Key words: Contextual Effect; EBLUP; Small Area Estimation; Synthetic Estimator.

1. Introduction

Sample surveys allow efficient inference about a national population when resources do not permit collecting relevant information from every member of the population. As a consequence, many sample surveys are conducted each year around the world to obtain statistical information required for policy making. In recent years, there is increasing need for statistical information at sub-national levels. Such statistics are often referred to as small area statistics, and methods for obtaining them from national surveys have become an important research topic, stimulated by demands from government agencies and businesses for data at different geographic and socio-demographic levels. In this context, Small Area Estimation (SAE) refers to statistical techniques for producing reliable estimates for geographic sub-populations (such as city, province or state) and socio-demographic sub-domains (such as age group, gender group, race group etc.) where the survey data available are insufficient for calculation of reliable direct estimates.

A fundamental property of SAE methods is that they combine related auxiliary variables with statistical models to define estimators for small area characteristics. Most statistical models used in SAE can be formulated either at the unit level or at the area level. When sample data are available at the individual or unit level, a common approach in SAE is to compute parameter estimates based on a unit-level mixed linear model. However, it is also often possible to fit this type of model at the area level, and to then compute the small area estimates based on this area-level mixed model.

In this paper we explore the relative performance of small area estimates based on area-level models with the same estimates based on unit-level models given both individual and aggregate (i.e. area level) data are available. We assume that the targets of inference are the area-level population means of a variable. A unit-level analysis is thus at a different level from which the final estimates will be calculated.

Our aim is to identify situations where aggregated-level analysis can provide more reliable estimates than unit-level analysis. This may happen due to the presence of contextual or area-level effects in the small area distribution of the target variable. Ignoring these effects in unit-level models can lead to biased estimates. However, such area-level effects are automatically included in area-level models in certain cases.

In this paper, matrices are displayed in bold type. Sample statistics are denoted by lowercase letters, with uppercase used for corresponding population statistics.

2. Area-level Approach

Fay and Herriot (1979) applied a linear regression with area random effects with unequal variances for predicting the mean value per capita income (PCI) in small geographical areas.

Throughout this paper we shall assume the target population divided into K sub-domains. In such a case, Fay-Herriot model is:

$$\hat{\bar{Y}}_k^D = \bar{Y}_k + \varepsilon_k \;\; ; \;\; k = 1, \ldots, K \tag{1}$$

where $\bar{Y}_k$ is the true population value for kth area mean for the target variable, $\hat{\bar{Y}}_k^D$ denotes its direct estimate and $\varepsilon_k | \bar{Y}_k \sim N(0, \sigma^2_{\varepsilon_k})$. $\bar{Y}_k$ is assumed in (1) to be related with P auxiliary variables as follows:

$$\bar{Y}_k = \bar{\mathbf{X}}_k' \beta + u_k \;\; ; \;\; where \;\; u_k \sim N(0, \sigma_u^2) \tag{2}$$

where $\bar{\mathbf{X}}_k$ is the vector of kth area population means for P auxiliary variables.

$$\bar{\mathbf{X}}_k' = [1 \;\; \bar{X}_{k1} \;\; \bar{X}_{k2} \;\; \ldots \;\; \bar{X}_{kP}]$$

Variance of the error term $(\sigma^2_{\varepsilon_k})$ is typically assumed to be associated with the complex sampling error for kth area and it is assumed to be known in (1). This strong assumption seems unrealistic in practice (González-Manteiga, *et. al.* (2010)). The implications of having to estimate variance components and the effectiveness of the aggregated-level approach in SAE is considered in following sections.

3. Unit-level Approach

A standard Linear Mixed Model (LMM) for individual-level population data is:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{3}$$

Supposing N to be the population size, $\mathbf{Y}$ is a column vector of N random variables, $\mathbf{X}$ is an $N \times (P+1)$ matrix of known quantities whose rows correspond to the statistical units, and β is a vector of $(P+1)$ parameters. $\mathbf{Z}$ is a $N \times K$ matrix of random-effect regressors, and finally, $\mathbf{u}$ and $\mathbf{e}$ are respectively $K \times 1$ and $N \times 1$ vectors of different random effects. Note that, $\mathbf{u}$ and $\mathbf{e}$ are assumed to be distributed independently with mean zero and covariance matrices $\mathbf{G}$ and $\mathbf{R}$, respectively.

$$E(\mathbf{u}) = \mathbf{0} \;\; \& \;\; E(\mathbf{e}) = \mathbf{0} \; ; \; Var\begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix}$$

The variance-covariance matrix for $\mathbf{Y}$ is:

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}.$$

Under the general definition of LMM, Datta and Lahiri (2000) defined the target of inference as:

$$\mu_{\bar{Y}_k} = \bar{\mathbf{X}}_k' \beta + u_k$$

The BLUP for $\mu_{\bar{Y}_k}$ is:

$$\tilde{\mu}_{\bar{Y}_k} = \bar{\mathbf{X}}_k' \tilde{\beta} + \mathbf{l}'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\tilde{\beta}) \tag{4}$$

where

$$\begin{aligned} \tilde{\beta} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \\ \mathbf{l}' &= (\underbrace{0, 0, \ldots, 0, 1}_{k}, 0, ..., 0) \end{aligned} \tag{5}$$

To calculate BLUP value for $\mu_{\bar{Y}_k}$ in (4), variance components are assumed to be known. Replacing the estimated values for the variance components in (4) and (5), a new estimator will be obtained. This estimator is presented by Harville (1991) as an "empirical BLUP" or EBLUP.

The mean value for the target variable within the kth area can be estimated based on the fitted working model through the synthetic estimation technique as:

$$\hat{\bar{Y}}_k^{Syn} = \bar{\mathbf{X}}_k' \tilde{\beta}$$

A similar approach can be used to calculate parameter estimates and consequent synthetic estimators and EBLUPs under Fay-Herriot model [Longford (2005)].

If individual-level data are available, small area estimation is usually based on models formulated at the unit level but they are ultimately used to produce estimates at the area level. Using aggregated-level analysis may cause loss of efficiency when the data is available at the individual level. When the data comes from a complex sample, it may not be very straightforward to calculate small area estimates. Therefore, a common approach is to use area-level estimates that account for the complex sampling and regression model of a form introduced in (1).

4. Contextual Models

Linear mixed models such as (3) are commonly used in SAE. However, area-level covariates can also be included in the unit-level models in order to improve the efficiency in certain cases. Supposing $\mathcal{T}_k$ to denote the vector of kth area-level covariates being included in unit-level population model, we have:

$$\begin{aligned} Y_{ik} &= (\mathbf{X}_{ik}'; \mathcal{T}_k')\beta + u_k + e_{ik} \\ & \qquad i = 1, \ldots, N_k \;\; \& \;\; k = 1, \ldots, K \\ u_k &\sim N(0, \sigma_u^2) \;\; ; \;\; e_{ik} \sim N(0, \sigma_e^2) \end{aligned} \tag{6}$$

where $\mathbf{X}_{ik}' = [1 \;\; X_{ik1} \;\; X_{ik2} \;\; \ldots \;\; X_{ikP}]$. N_k denotes the population size for kth area.

In the statistical literature, area-level covariates such as those in (6) are sometimes referred to as 'contextual effects' and model (6) is then described as a 'contextual model'. A special case of $\mathcal{T}_k$ is where the contextual effects are small area population means. Then we have:

$$\begin{aligned} Bias_\xi\left(\hat{\bar{Y}}_k^{Syn}\right) &= \bar{\mathbf{X}}_k' Bias_\xi(\tilde{\beta}) \\ Bias_\xi\left(\hat{\bar{Y}}_k^{EBLUP}\right) &\approx (1 - \gamma_k)\bar{\mathbf{X}}_k' Bias_\xi(\tilde{\beta}) \end{aligned} \tag{7}$$

where:

$$\gamma_k = \frac{\sigma_u^2}{\sigma_u^2 + \psi_k} \quad ; \quad \psi_k = Var(\bar{e}_k|\bar{Y}_k)$$

Note that, the subscript ξ denotes the bias under the assumed population model (6).

It is often the case in practice that unit-specific and area-specific coefficient estimates would have different expectations. This may happen as a result of area-level miss-specifications in individual-level analysis and can cause an error in the interpretation of statistical data.

5. Monte-Carlo Simulation

A model-assisted design-based simulation study is presented in this section to assess the empirical Mean Square Error (MSE) of resulting synthetic estimators and EBLUPs based on individual-level and aggregated-level analysis. To develop the numerical study, the gross weekly income is considered as the target variable. Available data on the length of education and training experience for different individuals aged 15 and over is then used as auxiliary information. The target variable is assumed to be related with the auxiliary variable through a linear mixed model.

Available information for 57 statistical sub-divisions in Australia about the mentioned characteristics and area population sizes is used in this study. Area means are also included in the individual-level population model for generating population data in this monte-carlo simulation.

Considering the actual area means to be the target of inference, synthetic estimates and EBLUPs are then calculated based on two working models fitted on the sample data as follows:

$$\begin{aligned} y_{ik}^{(W_1)} &= (1; x_{ik})\beta + u_k + e_{ik} \\ u_k &\sim N(0, \sigma_u^2) \;\; ; \;\; e_{ik} \sim N(0, \sigma_e^2) \\ & \qquad i = 1, \ldots, n_k \;\; \& \;\; k = 1, \ldots, K \\ \bar{y}_k^{(W_2)} &= (1; \bar{x}_k)\beta + u_k + \bar{e}_k \\ \bar{\mathbf{e}} &\sim N\Big(\mathbf{0} \,,\; diag(\tfrac{\sigma_e^2}{n_1}, \;\ldots,\; \tfrac{\sigma_e^2}{n_K})\Big) \end{aligned} \tag{8}$$

where n_k is the sample size allocated to kth area. The first working model *(W1)* can be fitted on individual-level sample data while the second working model *(W2)* uses aggregated-level sample data for estimation purposes.

This allows a comparison to be made among the performance of the models in (8) in case of having actual area means as possible contextual effects in population model. In this study, the model parameters β, σ_e^2 and σ_u^2 are empirically estimated in both unit-level and area-level models by Fisher scoring algorithm as a general method for finding maximum likelihood parameter estimates (Longford (2005)).

Figure (1) summarizes the results by giving the ratio of the MSEs for the SAEs based on unit-level and area-level models for $K = 57$ areas in the simulation.

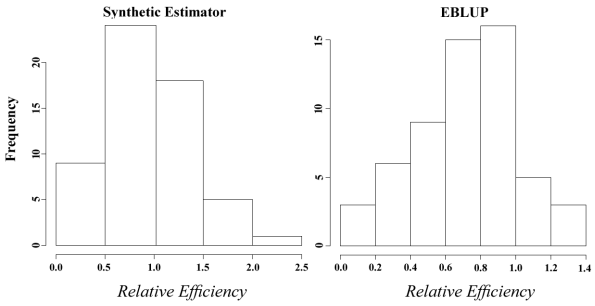


Figure 1: The Relative Efficiency of Unit-level to Area-level Model

A value less than 1 for the relative efficiency in figure (1) indicates that the unit-level approach based on *(W1)* is more precise comparing with the area-level approach based on *(W2)* in terms of MSE in each case. Using synthetic approach, it is difficult to say which model helps to obtain more precise estimates. The ratio varies below and above 1 for the synthetic estimation, while this value is generally below 1 for the EBLUP. This can be due to the effect of the shrinkage factor used in EBLUP technique (see (7)).

6. Conclusion

Individual-level analysis usually results in more stable small area estimates. However, if the unit-level working model is misspecified by exclusion of important auxiliary variables, parameter estimates obtained from the individual and aggregated level analysis will have different expectations.

In particular, if an existing contextual variable is ignored, the parameter estimates calculated from an individual-level analysis will be biased, whereas an aggregated-level analysis can lead to small area estimates with less bias. Even if contextual variables are included in an unit-level modeling, there may be an increase in the variance of parameter estimates due to increased number of variables in the working model.

References

[1] Datta, G. S., and Lahiri, P. (2000). A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems. *Statistica Sinica*. **10**, 613-627.

[2] Fay, R. E., and Herriot, R. A. (1979). Estimates of Income for Small Places: an Application of James-Stein Procedures to Census Data. *Journal of The American Statistical Association.* **74**, 269-277.

[3] González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., and Santamaría, L. (2010). Small Area Estimation under FayHerriot Models with Non-parametric Estimation of Heteroscedasticity. *Statistical Modelling*. **10**, 215-239.

[4] Harville D. A. (1991). That BLUP is a Good Thing: The Estimation of Random Effects, (Comment). *Statistical Science*. **6**, 35-39.

[5] Longford N. T. (2005). *Missing Data and Small Area Estimation*. Spring-Verlag.

Two-Sample Testing for Equality of Variances

David Allingham and J. C. W. Rayner

School of Mathematical and Physical Sciences,
The University of Newcastle, NSW 2308, Australia

David.Allingham@newcastle.edu.au and John.Rayner@newcastle.edu.au

Abstract

To test for equality of variances given two independent random samples from univariate normal populations, popular choices would be the two-sample F test and Levene's test. The latter is a nonparametric test while the former is parametric: it is the likelihood ratio test, and also a Wald test. Another Wald test of interest is based on the difference in the sample variances. We give a nonparametric analogue of this test and call it the R test. The R, F and Levene tests are compared in an indicative empirical study.

For moderate sample sizes when assuming normality the R test is nearly as powerful as the F test and nearly as robust as Levene's test. It is also an appropriate test for testing equality of variances without the assumption of normality, and so it can be strongly recommended.

Keywords: Bartlett's test; Levene's test; Wald tests.

1. Introduction

In the two-sample location problem we are given two independent random samples X_{11}, ..., X_{1m} and X_{21}, ..., X_{2n}. The pooled t-test is used to test equality of means assuming that the variances are equal and that the samples are from normal populations. Welch's test can be used when equality of variances is suspect but normality is not, and the Wilcoxon test can be used when normality is in doubt.

The corresponding dispersion problem is of interest to confirm the validity of, for example, the pooled t-test, and for its own sake. As an example, testing for reduced variability is of interest in confirming natural selection, in which reduced variability supports the hypothesis that the species evolves in a particular, albeit broad, direction. In exploratory data analysis it is sensible to test whether one population is more variable than another. If it is, the cause may be that one population is bi-modal relative to the other; the consequences of this in both the scenario and the model can then be explored in depth.

Here, a new test for equality of variances based on what might be called a nonparametric version of a very natural Wald test is introduced. In an indicative empirical study we show that, in moderately-sized samples, the new test is nearly as powerful as the F test when normality may be assumed, and is nearly as robust as Levene's test when normality is in doubt. See [3, p.519], who say that the "F test and other procedures for inference about variances are so lacking in robustness as to be of little use in practice." The new test gives a counterexample to that proposition.

We acknowledge that the usefulness of the new test is limited to moderate sample sizes of at least 25 each, a reasonable expectation in a serious study aiming at reasonable power which could not be hoped for with samples of size 10 or so.

We are aware of more expansive comparative studies such as those of [1] and [2]. Our goal here is not to emulate these studies but to merely show that the new test is competitive and interesting. Reflecting our limited study, we restrict attention to samples of equal size from both populations and a 5% level of significance.

In Section 2 the new test is introduced. In Section 3 we investigate test size. It is shown that when normality may be assumed the asymptotic χ^2 critical values may be used for moderate sample sizes, achieving test sizes 'close' to nominal. We then show that when sampling from t distributions with various

degrees of freedom, the F test is highly non-robust for small degrees of freedom, as is well-known for fat-tailed distributions. The new test is challenged somewhat for small degrees of freedom, but its performance is only slightly inferior to the Levene test.

In Section 4 it is shown that when normality holds the new test is not as powerful as the Levene test for small sample sizes, but overtakes it for moderate sample sizes of about 25. The new test is always inferior to the optimal F test, but has power that approaches that of the F test, its power being at least 95% of that of the F test throughout most of the parameter space for sample sizes of at least 80. This, in conjunction with the fact that the new test is valid when normality doesn't hold, is a strong reason for preferring the new test for moderate sample sizes.

2. Competitor Tests for the Two-Sample Dispersion Problem

We assume independent random samples of sizes m and n from normal populations, $N(\mu_i, \sigma_i^2)$ for $i = 1, 2$. We wish to test H: $\sigma_1^2 = \sigma_2^2$ against the alternative K: $\sigma_1^2 \neq \sigma_2^2$. If S_i^2, $i = 1, 2$ are the unbiased sample variances, then the so-called F test is equivalent to the likelihood ratio test and is based on the quotient of the sample variances, $S_2^2 / S_1^2 = F$, say. It is well-known, and will be confirmed yet again in Section 3, that the null distribution of F, $F_{m-1,\ n-1}$, is sensitive to departures from normality. If $F_{m-1,\ n-1}(x)$ is the cumulative distribution function of this distribution, and if c_p is such that $F_{m-1,\ n-1}(c_p) = p$, then the F test rejects H at the 100α% level when $F \leq c_{\alpha/2}$ and when $F \geq c_{1-\alpha/2}$.

Common practice when normality is in doubt is to use a nonparametric test such as the Levene or the Mood tests. In the two-sample case, Levene's test is just the pooled t-test applied to the sample residuals. There are different versions of Levene's test using different definitions of residual. The two most common versions use the group means, $|X_{ij} - \overline{X}_{i.}|$, and the group medians, $|X_{ij} - \tilde{X}_{i.}|$, in obvious notation. The latter is called the Brown-Forsythe test. The distribution of the test statistics, say L and B, that are the squares of the pooled t-test statistics using mean- and median-based residuals, respectively, is approximately $F_{1,\ m+n-2}$. Again it is well-known that the tests based on L and B are robust, in that when the population variances are equal but the populations themselves are not normal, they achieve levels 'close' to nominal. However this happens at the expense of some power. As this paper presents an indicative, rather than exhaustive, study, we will henceforth make comparisons only with the Levene test.

We now construct a new test that we will call the R test. For univariate parameters θ, a Wald test statistic for H: $\theta = \theta_0$ against the alternative K: $\theta \neq \theta_0$ is based on $\hat{\theta}$, the maximum likelihood estimator of θ, usually via the test statistic $(\hat{\theta} - \theta_0)^2 / \text{est}\,\text{var}(\hat{\theta})$, where $\text{est}\,\text{var}(\hat{\theta})$ is a consistent estimate of $\text{var}(\hat{\theta})$. This test statistic has an asymptotic χ_1^2 distribution. As well as being the likelihood ratio test, the F test is also a Wald test for testing H: $\theta = \sigma_2^2 / \sigma_1^2 = 1$ against K: $\theta \neq 1$.

A Wald test for testing H: $\theta = \sigma_2^2 - \sigma_1^2 = 0$ against K: $\theta \neq 0$ is derived in [4]. The test statistic is

$$\frac{(S_1^2 - S_2^2)^2}{2S_1^4/(n_1+1) + 2S_2^4/(n_2+1)} = W,$$

say. Being a Wald test, the asymptotic distribution of W is χ_1^2, while its exact distribution is not obvious. However, W is a one-to-one function of F, and so the two tests are equivalent. Since the exact distribution of F is known, the F test is the more convenient test.

The variances, $\text{var}(S_j^2)$, used in W are estimated optimally using the Rao-Blackwell theorem. This depends very strongly on the assumption of normality. If normality is in doubt then we can estimate $\text{var}(S_1^2 - S_2^2)$ using results in [5]. For a random sample $X_1, \ldots, X_n$ and population and sample central moments μ_r and $m_r = \sum_{j=1}^{n} (X_j - \overline{X})^r / n$, $r = 2, 3, \ldots$, [5] gives

$$\text{E}[m_r] = \mu_r + \text{O}(n^{-1}) \text{ and}$$
$$\text{var}(m_2) = (\mu_4 - \mu_2^2)/n + \text{O}(n^{-2}).$$

Applying [5, 10.5], μ_2^2 may be estimated to $\text{O}(n^{-1})$ by m_2^2, or, equivalently, $n\,m_2^2/(n-1) = S^4$, where S^2 is the unbiased sample variance. It follows that $\text{var}(m_2)$ may be estimated to order $\text{O}(n^{-2})$ by $(m_4 - m_2^2)/n$. A robust alternative to W is thus

$$\frac{(S_1^2 - S_2^2)^2}{(m_{14} - S_1^4)/n_1 + (m_{24} - S_2^4)/n_2} = R,$$

say, in which m_{i4}, i=1, 2, are the fourth central sample moments for the ith sample. We call the test based on R the R test. In large samples the denominator in R

will approximate var($S_1^2 - S_2^2$) and R will have asymptotic distribution χ_1^2.

We emphasise that the R test is a Wald test in the sense described above. Since it doesn't depend on any distributional assumptions about the data, it can be thought of as a nonparametric Wald test. It can be expected to have good properties in large samples no matter what distribution is sampled.

All the above test statistics are invariant under transformations $Y_{ij} = a(X_{ij} - b_i)$, for constants a, b_1 and b_2 and for $j = 1, ..., n_i$ and $i = 1, 2$.

3. Test Size Under Normality and Non-normality

Under the null hypothesis, the distribution of F is known exactly, that of L is known approximately, and, as above, the distribution of R is known asymptotically. When analysing data, these distributions are used to determine p-values and critical values. We now investigate their use in determining test size.

Two empirical assessments of test size, defined as the probability of rejecting the null hypothesis when it is true, will now be undertaken. The test statistics are scale invariant, and so it is sufficient under the null hypothesis to take both population variances to be one. As this is an indicative study, we take $m = n$ and the significance level to be 5%.

In the first assessment we assume normality. The extent of the error caused by using the asymptotic critical point (approximately 3.841) in the R test is shown in Figure 1, using the proportion of rejections in N = 100,000 random samples. For $m = n = 10$ and 30 these proportions are approximately 20% and 8%. Most would hopefully agree that the former is not acceptably 'close' to 5%, whilst the latter is.

For various n, we estimated the 5% critical points for each test by generating N = 100,000 pairs of random samples of size n, calculating the test statistics, ordering them and identifying the 0.95Nth percentile. The estimated critical points of R approach the χ_1^2 5% critical point. These estimated critical points will be used in the subsequent power study later to give tests with test size exactly 5%.

Even if the R test has good power, the test is of little value unless it is robust in the sense that, even when the sampled distributions are not normal, the p-values are reasonably accurate. Thus in the second assessment we estimate the proportion of rejections when the null hypothesis is true and both the populations sampled are non-normal. We consider different kurtoses via t distributions with various degrees of freedom. If the degrees of freedom are large, say 50 or more, the sampled distribution will be sufficiently normal that the proportion of rejections should be close to the nominal.

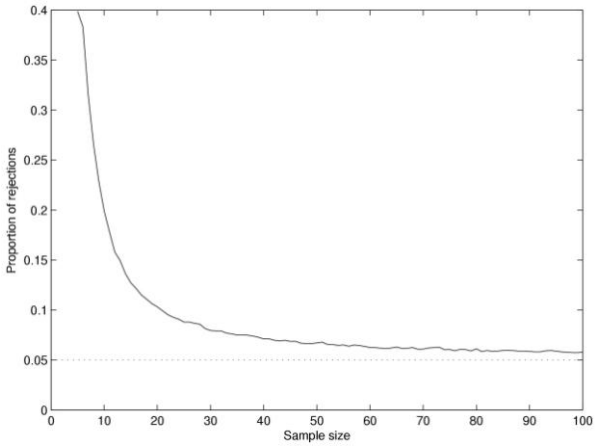


Figure 1: Proportion of rejections of the R test using the 5% critical point for sample sizes up to 100.

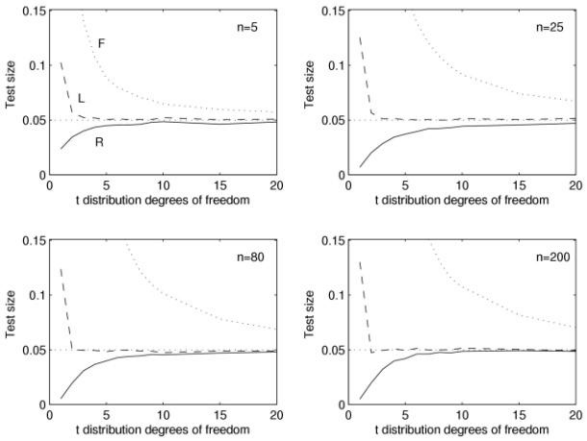


Figure 2: Test sizes for the F (dots), L (dashes) and R (solid line) tests for t distributions with varying degrees of freedom.

In Figure 2 we show the proportion of rejections for the Levene, F and R tests when sampling from t_ν distributions, for $\nu = 1, ..., 50$, with sample sizes of $m = n = 5$, 25, 80 and 200. Interestingly, the achieved test size is closer to the nominal 5% value for smaller samples, in all cases.

It is apparent that the F test performs increasingly poorly as the degrees of freedom diminish. It is also interesting to note that in this scenario the F test is always liberal (exact size greater than 5%) while the R test is always conservative (exact size less than 5%). In general, the latter is to be preferred.

The Levene test generally has exact level closer to the nominal level than the R test except for small degrees of freedom. Moreover, while the level of the R test is almost always reasonable, for very small ν the level is not as close to the exact level as perhaps would be preferable.

4. Power Under Normality

For the F, Levene and R tests we estimated the power as the proportion of rejections from N = 100,000 pairs of random samples of size n, where the first sample is from a N(0, 1) population and the second is from a N(0, σ^2) population with $\sigma^2 \geq 1$. To compare like with like, estimated critical values that give virtually exact 5% level tests were used. It is apparent that for sample sizes of about 20 the Levene test is superior to the R test; that between approximately 20 and 30 the R test takes over from the Levene test; and that thereafter the R test is always more powerful than the L test. These results are shown in Figure 3.

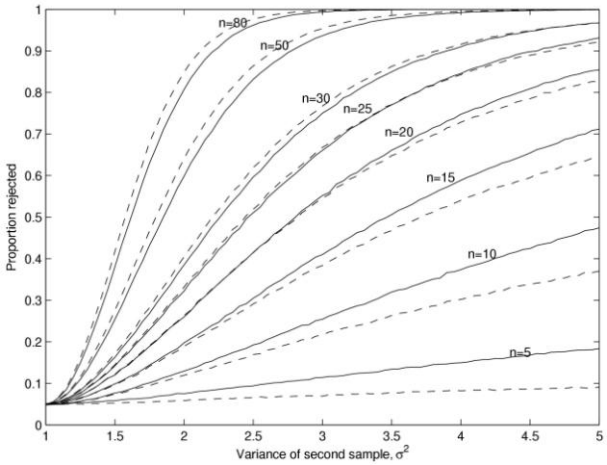


Figure 3: Power of the 5% level L test (solid line) and R test (dashed line) for various sample sizes.

When normality holds, both the Levene and R tests are always less powerful that the F test. This is explored in Figure 4, which compares the Levene test to the F test in the left-hand panel, and the R test to the F test in the right-hand panel. The figure shows a contour plot of the regions in which the ratio of the power of the stated test to the F test is either less than 95%, between 95% and 99.99%, or greater than 99.99%. The corresponding regions are far smaller for the Levene test than the R test. Moreover, it appears that for approximately $m = n > 80$, the power of the R test is always at least 95% of the power of the F test.

5. Recommendations

The R test is a nonparametric Wald test, so that when sampling from any non-normal distribution it can be expected to be at least as powerful as any competitor test in sufficiently large samples.

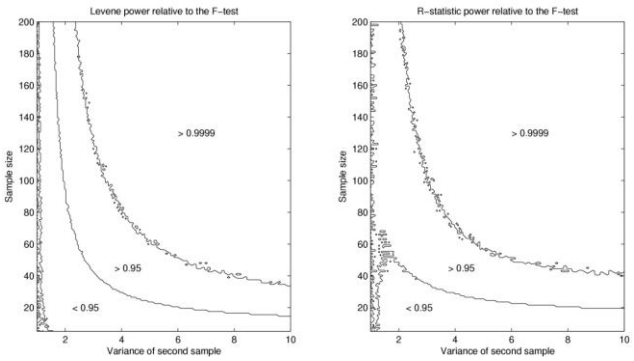


Figure 4: Contour plots of the power of the L test (left) and R test (right) relative to the F test power, showing regions in which the power ratios are less than 95%, between 95% and 99.99%, and greater than 99.99%.

If normality can be assumed then the F test is both the likelihood ratio test and a Wald test, and is the appropriate test to apply. However, if normality is doubtful then the well-known non-robustness of the F test means that tests such as the Levene test are more appropriate for small-to-moderate sample sizes. For sample sizes of at least 30, though, the R test is more powerful than the Levene test, and may be implemented using the asymptotic χ_1^2 distribution to obtain critical values and p-values.

If normality cannot be assumed, then the F test is no longer an optimal test, whereas the R test is. For moderate sample sizes of at least 30 in each sample, the R test has test size very close to the nominal and is more powerful than both the F and Levene tests. It should then be the test of choice.

Finally, we note that the R test may be extended to the multi-sample multivariate case, and that work is ongoing.

References

[1] BOOS, Dennis D. And BROWNIE, Cavell. (1989). Bootstrap methods for testing homogeneity of variances. *Technometrics*, 31, 1, 69-82.

[2] CONOVER, W.J., JOHNSON, Mark E. and JOHNSON, Myrle M. (1981). A comparative study of tests of homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, 23, 4, 351- 361.

[3] MOORE, D.S. and McCABE, G.P. (2006). *Introduction to the Practice of Statistics*. New York: W.H. Freeman.

[4] RAYNER, J.C.W. (1997). The Asymptotically Optimal Tests. *J.R.S.S., Series D (The Statistician)*, 46(3), 337-346.

[5] STUART, A. and ORD, J.K. (1994). *Kendall's Advanced Theory Of Statistics.* Vol.1: Distribution theory, 6th ed. London: Hodder Arnold.

Nonparametric Tests for Two Factor Designs with an Application to Latin Squares

J.C.W. Rayner and D.J. Best

The University of Newcastle, Callaghan, NSW, 2308, AUSTRALIA
John.Rayner@newcastle.edu.au and *John.Best@newcastle.edu.au*

Abstract

We show how to construct nonparametric tests for two factor designs. These tests depend on whether or not the levels of the factors are ordered. Pearson's X^2 statistic is decomposed into components of orders 1, 2, ... These components may be further decomposed, the decomposition depending on the design. If neither factor is ordered, the components reflect linear, quadratic etc main and interaction effects. The approach is demonstrated with reference to the latin squares design.

Keywords: Randomised block design, Randomised complete block design, orthonormal polynomials, Pearson's X^2

1. Introduction

The approach described here is based on components of the test statistic of the Pearson X^2 test of independence. The first order component utilises ranks. Tests are available of higher order components, which can be thought of as being based on generalised ranks.

In a limited empirical assessment for the latin square design we find that our first order (rank-based) test consistently gives superior power to the parametric F test and our benchmark nonparametric test, the Conover rank transform test (see [3, p.419]).

The approach generalises readily to the development of multifactor nonparametric tests.

In section 2 we construct contingency tables and show how Pearson's X^2 statistic X_P^2 may be partitioned into components that reflect, for example, linear, quadratic and other effects. The components depend on how many factors have ordered levels. In section 3 we consider no factors ordered, and in section 4 at least one factor ordered. Section 5 gives a brief empirical assessment for the latin squares design.

2. Decomposition of the Pearson Statistic into Linear, Quadratic and Other Effects

We assume that we have observations x_{ij}, $i = 1, ..., I$ and $j = 1, ..., J$, in which i and j are the levels of factors A and B respectively. All $IJ = n$ observations are ranked and we count N_{rij}, the number of times rank r is assigned to the observation at level i of factor A and level j of factor B. For simplicity we assume throughout that there are no ties.

2.1 Singly Ordered Tables: Neither Factor Ordered

Initially it is assumed that only the ranks are ordered. With no ties $\{N_{rij}\}$ defines a three-way singly ordered table of counts of zeros and ones. As in [2] and [4, section 10.2], Pearson's X^2 statistic X_P^2 may be partitioned into components Z_{uij} via

$$X_P^2 = \sum_{u=1}^{n-1}\sum_{i=1}^{I}\sum_{j=1}^{J} Z_{uij}^2$$

with $Z_{uij} = \sum_{r=1}^{n} a_u(r) N_{rij} / \sqrt{np_{.i.}p_{..j}}$, in which $\{a_u(r)\}$ is an orthonormal polynomial on $\{p_{r..}\}$ with $a_0(r) = 1$ for $r = 1, ..., n$. Here the standard dot notation has been used, so that, for example, $N_{...} = IJ = n$, the number of times a rank has been assigned. Formally X_P^2 also includes a term for Pearson's X^2 for the unordered table formed by summing over r: $\{N_{.ij}\}$. However this table has every entry one, and X^2 is zero. We also find that $N_{i.} = J$ and $N_{.j} = I$. It follows that $p_{.i.} = 1/I$ and $p_{.j} = 1/J$, giving $Z_{uij} = \sum_{r=1}^{IJ} a_u(r) N_{rij}$.

For $u = 1, 2, ..., n - 1$ define

$$SS_u = \sum_{i=1}^{I}\sum_{j=1}^{J} Z_{uij}^2$$

so that $X_P^2 = SS_1 + ... + SS_{n-1}$; the SS_u give order u assessments of factor effects.

The $\{Z_{uij}\}$ may be thought of as akin to Fourier coefficients: for each (i, j) pair Z_{uij} is the projection of x_{ij} into $[n - 1]$ dimensional 'order' space, where the first dimension reflects, roughly, location, and the second reflects, roughly, dispersion, and so. Now $Z_{1ij} = \sum_{r=1}^{n}(r-\mu)N_{rij}/\sigma$ in which $\mu = (n + 1)/2$ and $\sigma^2 = (n^2 - 1)/12$. The linear or location statistic is $SS_1 = \sum_{i,j} Z_{1ij}^2$. As in [4, section 3.4] this is of the form of a Kruskal-Wallis test.

2.2 Doubly Ordered Tables: One Factor Ordered

Now assume that the first factor is ordered. To reflect this change write N_{rsj} for the number of times rank r is assigned to the factor combination (s, i). As there are no ties $\{N_{rsj}\}$ defines a three-way doubly ordered table of counts of zeros and ones. As in [2] and [4, section 10.2], Pearson's X^2 statistic X_P^2 may be partitioned into components Z_{uvj} via

$$X_P^2 = \sum_{u=1}^{n-1}\sum_{v=1}^{I-1}\sum_{j=1}^{J} Z_{uvj}^2 + \sum_{v=1}^{I-1}\sum_{j=1}^{J} Z_{0vj}^2 + \sum_{u=1}^{n-1}\sum_{j=1}^{J} Z_{u0j}^2$$

with $Z_{uvj} = \sum_{r=1}^{n}\sum_{s=1}^{I} a_u(r)b_v(s)N_{rsj}/\sqrt{np_{..j}}$, in which $\{a_u(r)\}$ is orthonormal on $\{p_{r..}\}$ with $a_0(r) = 1$ for $r = 1, ..., n$ and $\{b_v(s)\}$ is orthonormal on $\{p_{.i.}\}$ with $b_0(s) = 1$ for $s = 1, ..., I$. We find that $N_{...} = n$, $p_{r..} = 1/n$, $p_{.i.} = 1/I$ and $p_{.j} = 1/J$, giving $Z_{uvj} = \sum_{r=1}^{n}\sum_{s=1}^{I} a_u(r)b_v(s)N_{rsj}/\sqrt{I}$. If for $u = 0, 1, 2, ..., n - 1$ and $v = 0, 1, ..., I - 1$, but not $(u, v) = (0, 0)$, $SS_{uv} = \sum_{j=1}^{J} Z_{uvj}^2$, we have $X_P^2 = \sum_{u,v} SS_{uv}$.

Analogous to [4, section 6.5] the Z_{11j} are Page test statistics at each of the levels of factor B, and the Z_{uvj} are extensions of Page's test statistic. Now $SS_{uv} = \sum_j Z_{uvj}^2$ gives an aggregate assessment over the whole table of order (u, v) effects, generalised correlations in the sense of [5]. As above, the aggregation of all these order (u, v) effects is X_P^2.

2.3 Completely Ordered Tables: Both Factors Ordered

Finally assume that both factors are ordered. To reflect this change write N_{rst} for the number of times rank r is assigned to the factor combination (s, t). With no ties $\{N_{rst}\}$ defines a three-way completely ordered table of counts of zeros and ones. As in [1] and [4, section 10.2], Pearson's X^2 statistic X_P^2 may be partitioned into components Z_{uvw} via

$$X_P^2 = \sum_{u=1}^{n-1}\sum_{v=1}^{I-1}\sum_{w=1}^{J-1} Z_{uvw}^2 + \sum_{v=1}^{I-1}\sum_{w=1}^{J-1} Z_{0vw}^2 + \sum_{u=1}^{n-1}\sum_{w=1}^{J-1} Z_{u0w}^2 + \sum_{u=1}^{n-1}\sum_{v=1}^{I-1} Z_{uv0}^2$$

with $Z_{uvw}\sqrt{n} = \sum_{r=1}^{n}\sum_{s=1}^{I}\sum_{t=1}^{J} a_u(r)b_v(s)c_w(t)N_{rst}$, in which $\{a_u(r)\}$ is orthonormal on $\{p_{r..}\}$ with $a_0(r) = 1$ for $r = 1, ..., n$, $\{b_v(s)\}$ is orthonormal on $\{p_{.s.}\}$ with $b_0(s) = 1$ for $s = 1, ..., I$ and $\{c_w(t)\}$ is orthonormal on $\{p_{..t}\}$ with $c_0(t) = 1$ for $t = 1, ..., J$.

In our previous notation $SS_{uvw} = Z_{uvw}^2$ for $u = 0, 1, 2, ..., n - 1$ and $v = 0, 1, ..., I - 1$, and $w = 0, 1, ..., J - 1$, but not $(u, v, w) = (0, 0, 0)$. Thus $X_P^2 = \sum_{u,v,w} S_{uvw}$. The SS_{uvw} may be thought of as further extensions of the Page test statistic, this time to three dimensions. The SS_{uv0}, SS_{u0w} and SS_{0vw} are the familiar two-dimensional generalised Page test statistics as, for example, in [4, section 6.5 and Chapter 8].

3. Factors Not Ordered

Recall now that in the two factor analysis of variance without replication with observations y_{ij}, $i = 1, ..., I$ and $j = 1, ..., J$, the total sum of squares $SS_{\text{Total}} = \sum_{i,j}(y_{ij} - \bar{y}_{..})^2$ may be arithmetically partitioned into sum of squares due to factor A, namely $SS_{\text{A}} = J\sum_i(\bar{y}_{i.} - \bar{y}_{..})^2$, due to factor B, namely $SS_{\text{B}} = I\sum_j(y_{.j} - \bar{y}_{..})^2$, and a residual or interaction sum of squares $SS_{\text{AB}} = \sum_{i,j}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$. Thus

$$SS_{\text{Total}} = SS_{\text{A}} + SS_{\text{B}} + SS_{\text{AB}}.$$

Here $y_{i.} = \sum_j y_{ij}$ and $\bar{y}_{i.} = y_{i.}/J$ etc as usual.

For each $u = 1, 2, ..., n - 1$ put $y_{ij} = Z_{uij} = \sum_{r=1}^{n} a_u(r)N_{rij}$ in SS_{Total}. The order u factor A sum of squares is $SS_{u\text{A}} = \sum_i Z_{ui.}^2/J - Z_{u..}^2/(IJ)$. As in [4, section 3.4], $SS_{1\text{A}}$ is the Kruskal-Wallis test statistic for factor A, and for general u the $SS_{u\text{A}}$ are the component test statistics discussed there. Clearly the $SS_{u\text{B}}$ are the parallel generalised Kruskal-Wallis test statistics for factor B, while the $SS_{u\text{AB}}$ are nonparametric tests for generalised interaction effects. For example, for $u = 2$, $SS_{2\text{AB}}$ assesses whether or not the

quadratic (dispersion) factor A effects are the same at different levels of factor B.

Examples.

The completely randomised design can be accessed either by calculating the $SS_{u\mathrm{A}}$ directly or by partitioning the $\{Z_{uij}\}$ as in the one factor ANOVA. However it is done, the usual Kruskal-Wallis test statistic and its extensions are obtained.

In the randomised block design factor A can be taken to be treatments and factor B blocks. Of course there is no interest in testing for a block effect or a treatment by block interaction effect. The treatment effect test is not the Friedman test, as observations are ranked overall, not merely on each block. From an overall ranking the ranks on each block may be derived, so there is more information assumed in this approach. This could result in more power when the test is applicable. In some situations only ranks within blocks are available.

4. At least One Factor Ordered

Suppose now that the first factor is ordered. The $Z_{uvj} = \sum_{r=1}^{n}\sum_{s=1}^{I} a_u(r)b_v(s)N_{rsj}/\sqrt{I}$, are generalised Page test statistics at each level of factor B. As in the Happiness example in [4, p. 147 and p. 188], X_P^2 may be partitioned into meaningful components. An alternative is to sum over the levels of factor B and obtain $Z_{uv.}$, generalised Page test statistics aggregating over factor B. This is appropriate when factor B is replicates, as in the completely randomised design, or blocks, as in the randomised block design

If both factors are ordered X_P^2 is partitioned by the SS_{uvw} of section 2.3. These are new extensions of the Page test, this time to three dimensions.

5. Latin Squares

The parametric analysis of the $t \times t$ latin square design partitions the total sum of squares into sum of squares of treatments, rows and columns and error. For the nonparametric analysis we assume that neither rows nor columns are ordered and investigate parallel partitions of the total sum of squares.

We count N_{rjk}, the number of times rank r is assigned to the treatment in row j and column k, with $r = 1, \ldots, t^2, j, k = 1, \ldots, t$. Note that treatment i, $i = 1, \ldots, t$, occurs in cells (j, k) specified by the design. As long as we know any two of the treatment, row and column, we know the other. Hence a latin square may be considered to be any of three two factor designs. This observation is utilised subsequently.

Throughout this section we assume that treatments are unordered, and that only the ranks are ordered. With no ties $\{N_{rjk}\}$ defines a three way singly ordered table of counts of zeroes and ones.

As in section 2, $X_P^2 = SS_1 + \ldots + SS_{t^2-1}$ in which

$$SS_u = \sum_{j=1}^{t}\sum_{k=1}^{t} Z_{ujk}^2 \text{ for all } u$$

$$\text{with } Z_{ujk} = \sum_{r=1}^{t^2} a_u(r)N_{rjk}.$$

The factor A test statistic of order $u = 1, \ldots, t^2 - 1$, can be denoted by $SS_{u\mathrm{A}}$, a generalised Kruskal-Wallis test statistic. By letting the factors be in turn rows and columns, columns and treatments, and treatments and rows, we are able to show that

$$3\,SS_u = 2\,SS_{u\text{treatments}} + 2\,SS_{u\text{rows}} + 2\,SS_{u\text{columns}} + SS_{u\text{treatments}\times\text{rows}} + SS_{u\text{treatments}\times\text{columns}} + SS_{u\text{rows}\times\text{columns}}.$$

In most applications it is enough to know that $SS_u = SS_{u\text{treatments}}$ + residual, but it is interesting to know that, parallel to the parametric partition, the residual could be used to assess rows and columns, and interactions between treatments, rows and columns. However, unlike the parametric case, this analysis applies for any order. We recognise that in most applications few users would be interested in treatment effects beyond orders two or three.

Empirical Study

We now briefly assess the power properties of some of the tests constructed. Treatments tests of orders one and two, with test statistics denoted by $SS_{1\mathrm{T}}$ and $SS_{2\mathrm{T}}$ respectively, are considered. We also consider tests formed from the table of counts $\{N_{rsi}\}$ where the second category is treatments, assumed to be ordered. Then test statistics S_{uv} are constructed from $\{N_{rs.}\}$, particularly the Page test based on S_{11} and the umbrella test based on S_{21}. These will be compared with the parametric F test (denoted by F) and the Conover rank transform test (denoted by CRT) that ranks the data and applies a parametric F test to the ranks.

Only the 5×5 Latin square is considered, and rather than use asymptotic critical values 5% critical values are found using random permutations. The critical value for $SS_{1\mathrm{T}}$ is 8.9059 while that for the CRT test was 3.3642. Compare these with the asymptotic critical values of 9.4877 using the χ_4^2 distribution for the $SS_{1\mathrm{T}}$ test and 3.2592 using the $F_{4,12}$ distribution for the CRT test. Not surprisingly these asymptotic critical values aren't practical for a

table of this size. However the critical value for the parametric F test is exact.

Table 5. Test sizes for competitor tests for various error distributions.

Error distn	CRT	SS_{1T}	F	SS_{2T}	S_{11}	S_{21}
Normal	0.050	0.049	0.050	0.050	0.051	0.049
Expon	0.050	0.050	0.040	0.049	0.050	0.051
U(0, 1)	0.052	0.052	0.055	0.049	0.052	0.051
Cauchy (t_1)	0.049	0.049	0.017	0.050	0.051	0.051
t_2	0.049	0.050	0.031	0.050	0.052	0.051
t_3	0.050	0.050	0.041	0.050	0.051	0.051
Lognormal	0.049	0.049	0.032	0.049	0.052	0.050

Table 6. Powers for competitor tests for various error distributions with linear alternatives $\alpha_i = (-1, -0.5, 0, 0.5, 1)$.

Error distn	CRT	SS_{1T}	F	SS_{2T}	S_{11}	S_{21}
Normal	0.62	0.69	0.64	0.07	0.91	0.02
Expon	0.78	0.83	0.68	0.22	0.96	0.01
Cauchy (t_1)	0.19	0.22	0.05	0.07	0.40	0.04
t_2	0.32	0.37	0.20	0.07	0.62	0.03
t_3	0.40	0.45	0.32	0.06	0.72	0.02
Lognormal	0.54	0.59	0.29	0.22	0.84	0.02

Table 7. Powers for competitor tests for various error distributions with quadratic alternatives $\alpha_i = (1, 0, -2, 0, 1)$.

Error distn	CRT	SS_{1T}	F	SS_{2T}	S_{11}	S_{21}
Normal	0.94	0.97	0.96	0.34	0.01	0.98
Expon	0.93	0.95	0.94	0.48	0.01	0.99
Cauchy (t_1)	0.34	0.38	0.10	0.11	0.03	0.52
t_2	0.59	0.66	0.44	0.16	0.03	0.77
t_3	0.71	0.78	0.65	0.19	0.02	0.86
Lognormal	0.74	0.78	0.57	0.43	0.01	0.91

Table 8. Powers for competitor tests for various error distributions with complex alternatives $\alpha_i = (0.5, -0.5, 0, 0.5, -0.5)$.

Error distn	CRT	SS_{1T}	F	SS_{2T}	S_{11}	S_{21}
Normal	0.27	0.31	0.28	0.04	0.07	0.04
Expon	0.46	0.52	0.33	0.11	0.09	0.03
Cauchy (t_1)	0.11	0.12	0.03	0.05	0.06	0.05
t_2	0.16	0.17	0.09	0.05	0.07	0.05
t_3	0.18	0.21	0.14	0.05	0.07	0.05
Lognormal	0.30	0.34	0.13	0.15	0.08	0.03

All simulations relate to 5% level tests with sample sizes of 25, and are based on 100,000 simulations. The error distributions are Normal, exponential, uniform (0, 1), Cauchy (t_1), t_2, t_3 and lognormal.

Using the simulated critical values we found the test sizes given in Table 5. They are remarkably close to the nominal significance level, as befits nonparametric tests. However the parametric F test fared less well, often having test size less than 5%. This will mean the corresponding powers will be less than if the nominal level was achieved. Nevertheless, this is how the test would be applied in practice.

The critical values used in Table 5 were also used to estimate powers in subsequent tables. These powers use the model $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + E_{ijk}$ but with $\beta_j = \gamma_k = 0$ for all j and k in this study. The uniform error distribution doesn't appear in Tables 6 to 8 as all powers are 1.00.

Even when normality holds, the test based on SS_{1T} is, for the alternatives considered, slightly superior to the parametric F test, and clearly superior when normality doesn't hold. This linear effects test is also uniformly slightly superior to the Conover rank transform test. This is not due to a size difference as can be seen from Table 5. The Page and umbrella tests perform well when the alternative is constructed to reflect their designed strengths, but both are sometimes biased: their power is less than their test size. The performance of the test based on SS_{2T} is disappointing, perhaps because powers have not been given for alternatives constructed to reflect their designed strengths.

References

[1] Beh, E. J. and Davy, P. J. (1998). Partitioning Pearson's chi-squared statistic for a completely ordered three-way contingency table. *Australian and NZ Journal of Statistics*, 40, 465-477.

[2] Beh, E. J. and Davy, P. J. (1999). Partitioning Pearson's chi-squared statistic for a partially ordered three-way contingency table. *Australian and NZ Journal of Statistics*, 41, 233-246.

[3] Conover, W.J. (1998). *Practical Nonparametric Statistics* (3rd ed.). New York: Wiley.

[4] Rayner, J.C.W. and Best, D.J. (2001). *A Contingency Table Approach to Nonparametric Testing*. Boca Raton: Chapman & Hall/CRC.

[5] Rayner, J.C.W. and Beh, Eric J. (2009). Towards a Better Understanding of Correlation. *Statistica Neerlandica*. 63(3), 324-333.

The Odds Ratio and Aggregate Data: The 2×2 Contingency Table

Eric J. Beh

The University of Newcastle, Callaghan, NSW, 2308, AUSTRALIA
eric.beh@newcastle.edu.au

Abstract

The odds ratio remains one of the simplest of measures for quantifying the association structure between two dichotomous variables. Its use is especially applicable when the cell values of a 2×2 contingency table are known. However, there are cases where this information is not known. This may be due to reasons of confidentiality or because the data was not collected at the time of the study. Therefore one must resort to considering other means of quantifying the association between the variables. One strategy is to consider the aggregate association index (AAI) proposed by [1]. This paper will explore the characteristics of the AAI when considering the odds ratio of the 2×2 contingency table.

Keywords: 2×2 contingency table, aggregate association index, aggregate data, odds ratio.

1. Introduction

Consider a single two-way contingency table where both variables are dichotomous. Suppose that n individuals/units are classified into this table such that the number classified into the (i, j)th cell is denoted by n_{ij} and the proportion of those in this cell $p_{ij} = n_{ij} / n$ for $i = 1, 2$ and $j = 1, 2$. Denote the proportion of the sample classified into the ith row and jth column by $p_{i\bullet} = p_{i1} + p_{i2}$ and $p_{\bullet j} = p_{1j} + p_{2j}$ respectively. Table 1 provides a description of the notation used in this paper.

	Column 1	Column 2	Total
Row 1	n_{11}	n_{12}	$n_{1\bullet}$
Row 2	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Table 1: Notation for a 2x2 contingency table

Typically, measuring the extent to which the row and column variables are associated is achieved by considering the Pearson chi-squared statistic calculated from the counts and margins of a contingency table. For a 2×2 table of the form described by Table 1, this statistic is

$$X^2 = n \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}.$$

The direction and magnitude of the association may be determined by considering the Pearson product moment correlation

$$r = \frac{p_{11}p_{22} - p_{12}p_{21}}{\sqrt{p_{1\bullet}p_{2\bullet}p_{\bullet 1}p_{\bullet 2}}}$$

so that $X^2 = nr^2$. The problem at hand is to obtain some information concerning the nature of the association between the two dichotomous variables when only the marginal information is provided.

This paper will examine the structure of the association between two dichotomous variables based only on the marginal information. We shall do so by considering the aggregate association index proposed by [1, 2] in terms of the odds ratio, a very common measure of association for 2×2 contingency tables. The point of our discussion though is not to make inferences about the magnitude of the odds ratio, but to use its properties and the marginal frequencies (or proportions), to explore the association structure of the variables.

2. Aggregate Association Index

Let $P_1 = n_{11}/n_{1\bullet}$ and $P_2 = n_{21}/n_{2\bullet}$. Here P_1 is the conditional probability of an individual/unit being classified into 'Column 1' given that they are classified in 'Row 1'. Similarly, P_2 is the conditional probability of an individual/unit being classified into 'Column 1' given that they are classified in 'Row 2'. The following comments apply to P_1 only but may be amended if one wishes to consider P_2.

When the cells of Table 1 are unknown, the bounds of the (1, 1)th cell frequency are well understood [4] to lie within the interval

$$\max(0, n_{\bullet 1} - n_{2\bullet}) \le n_{11} \le \min(n_{\bullet 1}, n_{1\bullet}).$$

Therefore, the bounds for P_1 are

$$L_1 = \max\left(0, \frac{n_{\bullet 1} - n_{2\bullet}}{n_{1\bullet}}\right) \le P_1 \le \min\left(\frac{n_{\bullet 1}}{n_{1\bullet}}, 1\right) = U_1. \quad (1)$$

[2] showed that when only marginal information is available the 95% confidence interval for P_1 is

$$L_\alpha = \max\left(0, p_{\bullet 1} - z_{\alpha/2} p_{2\bullet}\sqrt{\frac{1}{n}\left(\frac{p_{\bullet 1} p_{\bullet 2}}{p_{1\bullet} p_{2\bullet}}\right)}\right) < P_1 < \min\left(0, p_{\bullet 1} + z_{\alpha/2} p_{2\bullet}\sqrt{\frac{1}{n}\left(\frac{p_{\bullet 1} p_{\bullet 2}}{p_{1\bullet} p_{2\bullet}}\right)}\right) = U_\alpha.$$

If $L_\alpha < P_1 < U_\alpha$ then there is evidence that the row and column variables are independent at the α level of significance. However, if $L_1 < P_1 < L_\alpha$ or $U_\alpha < P_1 < U_1$ then there is evidence to suggest that the variables are associated. From this interval, [1] proposed the following index

$$A_\alpha = 100\left(1 - \frac{\chi^2_\alpha\left[(L_\alpha - L_1) + (U_1 - U_\alpha)\right] + Int(L_\alpha, U_\alpha)}{Int(L_1, U_1)}\right) \quad (2)$$

where

$$Int(a, b) = \int_a^b X^2\left(P_1 \mid p_{1\bullet}, p_{\bullet 1}\right) dP_1$$

and

$$X^2\left(P_1 \mid p_{1\bullet}, p_{\bullet 1}\right) = n\left(\frac{P_1 - p_{\bullet 1}}{p_{2\bullet}}\right)^2\left(\frac{p_{1\bullet} p_{2\bullet}}{p_{\bullet 1} p_{\bullet 2}}\right) \quad (3)$$

Equation (2) is termed the aggregate association index (AAI). For a given α, this index quantifies how likely there will be a statistically significant association between the two dichotomous variables, given only the marginal information. A value of A_α close to zero suggests there is no association between the two variables. On the other hand, an index value close to 100 suggests that such an association may exist. An index above 50 will highlight that it is more likely that a significant association may exist than not. We will consider that an association is very unlikely, given only the marginal information, if the index is below 25.

3 The Odds Ratio

One of the most common measures of association for a 2×2 contingency table is the odds ratio

$$\theta = \frac{p_{11} p_{22}}{p_{21} p_{12}} = \frac{p_{11}\{p_{11} - (p_{1\bullet} + p_{\bullet 1} - 1)\}}{(p_{1\bullet} - p_{11})(p_{\bullet 1} - p_{11})}. \quad (4)$$

Often the logarithm of the odds ratio (also simply referred to as the log-odds ratio) is considered as a measure of association between two dichotomous variables. When the cell frequencies are known, the $100(1 - \alpha)\%$ confidence interval for log-odds ratio is

$$\ln(\theta) \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

It is demonstrated in [9] that, based only on the marginal frequencies of a 2×2 contingency table, there is not enough information available to infer the magnitude of the odds ratio. The underlying premise of the AAI is not to infer the magnitude of a measure of association. Instead it is to determine how likely a particular set of fixed marginal variables will enable to researcher to conclude that there exists a statistically significant association between the two dichotomous variables. In this paper, we tackle the problem by considering the odds ratio.

Since p_{11} is unknown here, one may express this proportion in terms of the marginal proportions and the odds ratio. If one considers (4), p_{11} may be expressed as a quadratic function in terms of the odds ratio. By solving this quadratic expression, we get

$$p_{11} = \frac{B - \sqrt{B^2 - 4 p_{1\bullet} p_{\bullet 1}\theta(\theta - 1)}}{2(\theta - 1)}$$

where

$$B = \theta(p_{1\bullet} + p_{\bullet 1}) + (p_{2\bullet} + p_{\bullet 1})$$

This result has been long studied and was considered by, for example, [8, pg 7] and [6, section 6.6]. Therefore, $P_1(\theta \mid p_{1\bullet}, p_{\bullet 1})$ may be expressed as

$$P_1(\theta) = \frac{B - \sqrt{B^2 - 4p_{1\bullet}p_{\bullet 1}\theta(\theta-1)}}{2p_{1\bullet}(\theta-1)} \qquad (5)$$

when $p_{1\bullet} \neq 0$. By substituting (5) into (3), the chi-squared statistic can be expressed as a function of the odds ratio.

It is very difficult to directly determine the 100(1 – α)% confidence intervals for the odds ratio based only on the marginal information. Such an interval, which we will denote by $\hat{L}_\alpha < \theta < \hat{U}_\alpha$, can be derived by considering those θ that satisfy

$$X^2(\theta \mid p_{1\bullet}, p_{\bullet 1}) = n\left(\frac{P_1(\theta) - p_{\bullet 1}}{p_{2\bullet}}\right)^2 \left(\frac{p_{1\bullet}p_{2\bullet}}{p_{\bullet 1}p_{\bullet 2}}\right) < \chi^2_\alpha$$

where χ^2_α is the 100(1 − α) percentile of a chi-squared distribution with 1 degree of freedom. Calculating $\hat{L}_\alpha$ and $\hat{U}_\alpha$ is computationally difficult. Therefore, for the purposes of our discussion, we shall approximate the bounds based on a graphical inspection of $X^2(\theta \mid p_{1\bullet}, p_{\bullet 1})$ versus θ.

We shall also be exploring the use of the log-odds ratio in the context of the AAI in the following section.

4 Example – Fisher's Twin Data

Consider the 2×2 contingency table of Table 4 analysed by [1, 2, 5]. These data concern 30 criminal twins and classifies them according to whether they are a monozygotic twin or a dizygotic twin. The table also classifies whether their same sex twin has been convicted of a criminal offence. We shall, for now, overlook the problem surrounding the applicability of using the Pearson chi-squared statistic in cases where the cell frequencies are not greater than five. [6] provides an excellent review of strategies for including Yate's continuity correction [11]. However, studies have revealed that incorporating the correction is not essential (eg [3, 7]) and so we will not consider its inclusion here.

The chi-squared statistic for Table 2 is 13.032, and with a p-value of 0.0003, shows that there is a statistically significant association between the type of criminal twin and whether their same sex sibling has been convicted of a crime. The product moment correlation of r = +0.6591 indicates that this association is positive. Therefore a monozygotic twin of a convicted criminal is associated with being convicted of a crime, while a dizygotic twin of a convicted criminal tends not to be a convicted criminal.

	Convicted	Not Convicted	Total
Monozygotic	10	3	13
Dizygotic	3	15	17
Total	12	18	30

Table 2: Criminal twin data original considered by [5]

[2] considered the AAI of Table 2 in terms of P_1 and showed that $A_{0.05}$ = 61.83. Therefore, it is likely that a 2 $\times$ 2 contingency table with the marginal information of Table 2 will reflect a statistically significant association (at the 5% level) between the two dichotomous variables. Figure 1 provides a graphical inspection of the meaning of this index. It shows that the Pearson chi-squared statistic is maximised at the bounds of P_1; the local maximum chi-squared values are 15.29 and 26.15. It can also be seen that the shaded region exceeding the critical value of $\chi^2_{0.05}(df = 1) = 3.84$ but below the chi-squared curve defined by (2) is quite large. This region represents 61.83% of the area under the curve and it is this quantity that is the AAI.

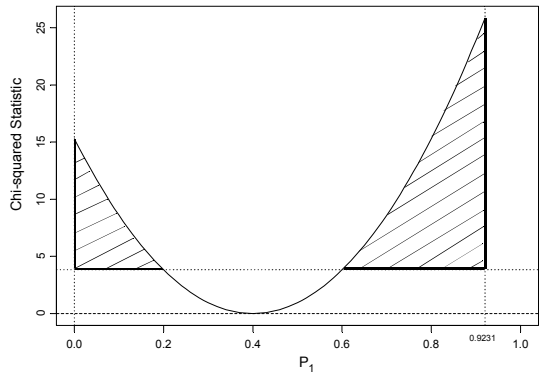


Figure 1: Plot of $X^2(P_1)$ versus P_1 for Table 1

For Table 2, $\theta = 25.00$ and the log-odds ratio of 3.22 has a 95% confidence interval of (1.26, 5.18). Thus, the 95% confidence interval for the odds ratio is (3.52, 177.48). Both these intervals indicate that there is a significant positive association between the two dichotomous variables at the 5% level of significance. This is consistent with the findings made regarding the Pearson product moment correlation. We shall now consider the case where the cell frequencies are unknown.

Despite the simplicity and popularity of the odds ratio, the issue of determining the AAI becomes a little more complicated, but equally revealing. Let us first consider the relationship between the Pearson chi-squared statistic and the odds ratio – see Figure 3. This figure graphically shows that a maximum chi-squared statistic is reached when the odds ratio approaches zero or reaches infinity.

Similarly, the chi-squared statistic achieves its minimum of zero when the odds ratio is 1.

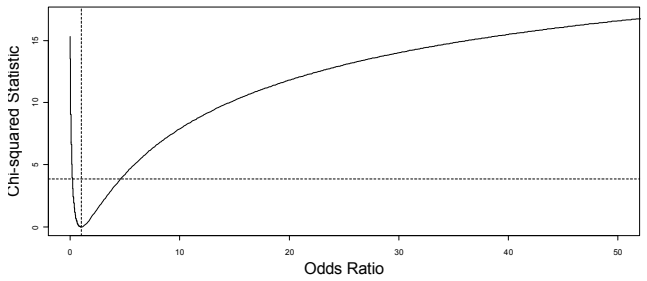


Figure 3: Plot of $X^2(\theta)$ versus θ.

Figure 3 shows the relationship between the chi-squared statistic and the odds ratio using (5). We can see that the chi-squared statistic is exceeded by the critical value of 3.84, at the 5% level of significance, when (approximately) $0.11 < \theta < 7.7$. However, since the shape of the curve is biased towards those odds ratios greater than 1, determining whether there may exist a positive or negative association using the odds ratio can produce misleading conclusions.

To overcome this problem we may also consider the log-odds ratio. Figure 4 shows the relationship between the Pearson chi-squared statistic and the log-odds ratio using (5). It reveals that when using (5) the local maximums of the Pearson chi-squared statistic (15.2941 and 26.1538) are reached as the log-odds ratio approaches negative, and positive, infinity.

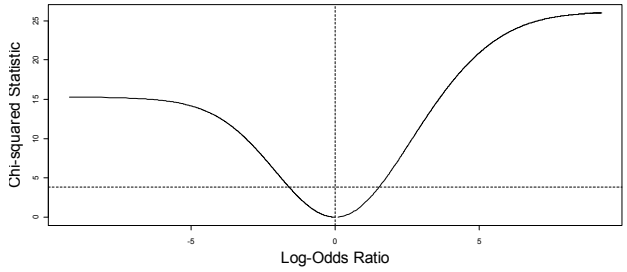


Figure 4: Plot of $X^2(\ln\theta)$ versus $\ln\theta$.

Figure 4 shows that, given only the marginal information of Table 2, there appears to be some evidence that a strong association exists. This is evident by considering the area under the curve that lies above the critical value of 3.84. In fact, by considering a log-odds ratio greater than zero, we can see that the area under the curve, using (5) is far greater than the area under the curve when the log-odds ratio is negative. This suggests that, not only is there strong evidence of a significant association between the two dichotomous variables, but that the association is more likely to be positive than negative.

5 Discussion

This paper discusses the use of the aggregate association index in terms of the odds ratio for a single 2×2 contingency table. By considering the index in this manner, we can identify how likely two categorical variables will be associated based only on the marginal frequencies using the most popular of simple measures of association. Of course, we may explore the behaviour of this index in terms of other simple measures of association, including $\beta_{11} = p_{11}/(p_{1\bullet}p_{\bullet 1})$ which is referred to as the (1, 1)th Pearson ratio.

Our focus has been concerned with the chi-squared statistic but the index may be generalised for other measures of association such as the Goodman-Kruskal tau index. Other popular measures for 2×2 contingency tables such as Yule's Q ("coefficient of association") or Yule's Y ("coefficient of colligation") may also be examined in this context. On may also consider extending this index for multiple 2×2 tables or larger sized contingency tables. We shall consider these, and other, issues in future discussions of the index.

References

[1] E.J. Beh, Correspondence analysis of aggregate data: The 2×2 table, Journal of Statistical Planning and Inference, 138, 2941 – 2952, 2008.

[2] E.J. Beh, The aggregate association index, Computational Statistics & Data Analysis, 54, 1570 – 1580, 2010.

[3] W.J. Conover, Some reasons for not using Yates continuity correction on 2×2 contingency tables (with discussion), Journal of the American Statistical Association, 69, 374 – 382, 1974.

[4] O.D. Duncan, B. Davis, An alternative to ecological correlation, American Sociological Review, 18, 665 – 666, 1953.

[5] R.A. Fisher, The logic of inductive inference (with discussion), Journal of the Royal Statistical Society (Series A), 98, 39 – 82, 1935

[6] J.L. Fleiss, B. Levin, M.C. Paik, Statistical Methods for Rates and Proportions (3rd ed), Wiley: NJ, 2003

[7] J.E. Grizzle, Continuity correction in the χ^2 for 2×2 tables, American Statistician, 21 (October), 28 – 32, 1967.

[8] F. Mosteller, Association and estimation in contingency tables, Journal of the American Statistical Association, 63, 1 – 28, 1968.

[9] R.L. Plackett, The marginal totals of a 2×2 table, Biometrika, 64, 37 – 42, 1977.

[10] J. Wakefield, Ecological inference for 2×2 tables (with discussion), Journal of the Royal Statistical Society, Series A, 167, 385 – 424, 2004.

[11] F. Yates, Contingency tables involving small numbers and the χ^2 test, Journal of the Royal Statistical Society Supplement, 1, 217 – 235, 1934.

An early ASEARC-taught subject: has it been a success?

Carole Birrell* and Ken Russell†

**Centre for Statistical and Survey Methodology, University of Wollongong, NSW, 2522, AUSTRALIA*
†School of Computing & Mathematics, Charles Sturt University, Wagga Wagga, NSW, 2678, AUSTRALIA

Abstract

The Applied Statistics Education and Research Collaboration (ASEARC) aims to "involve joint development and delivery of subjects and courses. ... There would be efficiency benefits involved in sharing subjects. There would also be significant benefits in ... students accessing subjects that would otherwise not be available to them, developed and presented by experts who would not usually be accessible. In parallel with the subject review the technological and administrative environment will also be assessed ... "

A 300-level subject covering Sample Surveys and Experimental Design has now been taught jointly to the Universities of Wollongong and Newcastle for two years, first using video-conferencing and then the Access Grid. In both years the subject was delivered by the same two lecturers.

We provide an initial review of the subject. We discuss its organisation, the use of the technology, changes in our teaching and administrative styles needed to cope with this mode of delivery, feedback from students, and our reaction to all of these. An overview of the subject results is given.

Key words: Access Grid, collaboration, statistics education, video-conference

1. Introduction

"The Applied Statistics Education and Research Collaboration (ASEARC) is a collaboration between statisticians and universities to work together exchanging information, supporting each other, and sharing loads." The collaboration includes joint development and delivery of subjects and courses [1].

A 300-level subject covering Sample Surveys and Experimental Design has now been taught jointly to the Universities of Wollongong (UoW) and Newcastle (UoN), delivered from UoW using videoconferencing in 2009 and the Access Grid (AG) in 2010.

We provide an initial review of the subject. We discuss its organisation, the use of the technology, changes in our teaching and administrative styles needed to cope with this mode of delivery, feedback from students, and our reaction to all of these. An overview of the subject results is given.

2. Subject Description

The equivalent of "STAT335 Sample Surveys and Experimental Design" at UoW is called "STAT3170 Surveys and Experiments" at UoN. In both years the subject was delivered by Ken Russell (Experimental Design component in weeks 1 - 6, 13) and Carole Birrell (Sample Surveys component in weeks 7 - 13) from UoW. The subject is aimed primarily at undergraduate students undertaking a statistics major but is open to students who have second year statistics prerequisites. A set of printed notes including lecture notes and tutorial questions is available to students at a minimal cost.

UoN has 12 teaching weeks plus a revision week (week 13) in which no new material is presented. At UoW, there are 13 teaching weeks plus a study week. In 2009 and 2010, the teaching weeks aligned and so the only change to the schedule was to incorporate the revision week in week 13 at both sites. Classes for UoN start on the hour beginning at 9 am, whereas, at UoW, classes start on the half-hour beginning at 8:30 am. Thus, students at UoN need to be aware of the different class times when choosing subjects.

Collaboration between the universities included a discussion of topics for each component, number of assessment tasks and the allocation of marks for in-term assessment and the final exam. The subject has three hours of face-to-face teaching each week. The lecturer has discretion in using classes as either lectures or tutorials. For each component, in-term assessment includes three small assignments and a project using SAS. The allocation is 40% in-term assessment and 60% exam.

Email address: cbirrell@uow.edu.au, kerussell@csu.edu.au (Carole Birrell* and Ken Russell†)

3. Modes of Delivery

In 2009, our first delivery to UoN was undertaken with videoconferencing. The AG was available but the lecturers were apprehensive about the reliability of the technology. In 2010, we trialled the AG technology.

3.1. Videoconferencing

Videoconferencing allows two-way video and audio communications between remote parties [2]. With videoconferencing, it is possible to record the class and, as such, the lecturer is recorded as well as any interaction between the sites. We found that the students did not access this very much, partly because the files of recordings were very large and downloading was time consuming and used up their download allocation. The document camera was used for hand writing a solution to a problem. This was not the best quality (not as crisp as an overhead projector). We did not have access to a smartboard with videoconferencing, although it was later discovered that it would have been possible. Although students could ask questions, it was not possible for them to write down anything for us to see.

A technician came to the room to connect the endpoints but left the room and monitored the connection from another location on campus during the lecture. It often took up to 15 - 20 minutes for everything to be connected properly. During the lecture, if the lecturer wanted to change from the presentation on the PC to the document camera, it was necessary to switch visuals using a remote control operated by the lecturer. If the connection between the universities dropped out, which occasionally happened, it was sometimes necessary for the technician to physically come back to the room. An obvious disadvantage was the loss of class time in connecting and reconnecting if necessary.

3.2. Access Grid

For ASEARC, the delivery of courses via media can be done through the AG using a room dedicated for the purpose, the Access Grid Room (AGR). "ASEARC's use of the AG is part of an international communication network that provides multimedia presentation to groups at different locations. The AG involves cameras, microphones, speakers, projectors, and other tools to support the presentation, such as the interactive 'whiteboard' to display lecture slides, otherwise called a smartboard. The board is also capable of receiving handwriting and replicating this on boards located in the other AGRs". For more information on the AG see http://www.accessgrid.org/ and [1].

In 2010, we used the AG technology for the delivery of STAT335. The UoN students received three projected images: one of the smartboard, one of the lecturer, and the other of the students at UoW. At UoW, the students were in the same room as the lecturer and the smartboard, and also saw a projected image of the students at UoN. Technicians were present in the AGRs at each end for the duration of the lecture and this proved to be worthwhile as the lecturers could then focus on the subject delivery.

There were many advantages of AG over videoconferencing. Firstly, use of the smartboard enhanced delivery of lecture and tutorial material. What was written on the smartboard could be captured and saved into a PDF file and subsequently the file could be uploaded to a subject website. Both the students and the lecturers particularly liked this feature. If a student missed a lecture, they could "fill in" the lecture notes by looking at what had been written on the smartboard. Although it did not capture the audio, it proved to be very helpful and was well utilised by students.

For consultation, it was possible for the UoN students to write on the UoW smartboard and vice versa. This was helpful when trying to work through a problem, and was used effectively by a couple of the students in Newcastle in 2010.

4. Teaching style: challenges and changes

Teaching with either videoconferencing or the AG is different to teaching just face-to-face. Many factors need to be taken into account both administratively and for pedagogical purposes.

Giving students many opportunities to interact during the lecture helps to minimise the "television viewing" mentality. Looking into the camera when addressing the students at the other end, rather than looking at the screen, gives the impression of making eye-contact and helps the students feel involved. When asking questions, we found that if we just said "Are there any questions?", it was confusing as to which group of students should answer first. Instead, it was better to address the group first by saying "This question is for the Newcastle students. Can someone tell me what the next step is in solving . . . " or even to address them by name: "Peter in Newcastle, can you tell me what the next step is in solving . . . ".

Visual variety can be achieved by changing the input (Caladine, 2008). For videoconferencing, this was done by switching between the computer and the document camera. For AG, we achieved this by changing from presentation/lecture style to writing on the smartboard where students fill in gaps left in lecture notes, or by solving tutorial problems on the smartboard with input from students.

4.1. Class Website

The subject had a website on UoW's *eLearning space* (which uses the Blackboard Learning System; see http://www.blackboard.com). This site allows documents to be stored for access by students, permits the

lodgment of assignments and their retrieval after marking, and provides a 'chat room'. It was used by the lecturers in exactly the same way as they would have used it to teach just STAT335 at UoW. The only difficulty experienced was in arranging access to this site for the UoN students. This is discussed later.

Administratively, organisation of lecture material is important since material is delivered electronically, requiring all lecture slides to be typed ahead of time. The use of eLearning space for everything means having a detailed schedule of dates for uploading of slides, tutorials, tutorial solutions, assignments questions and the return of marked assignments.

4.2. Assessments and Marking

Completed assignments had to be uploaded to the eLearning site by both UoW and UoN students. Requiring that all students adhered to the same process meant that UoN students were not disadvantaged. The lecturers were provided with a Toshiba Tablet PC and the PDF Annotator software (http://www.grahlsoftware.com). This enabled them to mark a PDF file containing a student's assignment. The lecturer had to download the assignment, mark the assignment (using PDF Annotator), and then upload the marked assignment to the eLearning site.

This procedure had definite advantages. The students kept their original assignment, having first scanned it, and then sent a soft copy. Students could send in their assignments from home if they had access to a printer/scanner. The lecturers could keep an original copy of the assignment and then use a second copy to mark and comment on. The marked assignment could also be kept for reference, which proved helpful if a student wished to discuss any comments or the marking scheme. Some students typed their assignments; however, it was not a requirement. Others typed parts and hand wrote in the equations, others hand wrote all.

The disadvantage for students was the requirement to make a PDF copy of assignments and upload to a website rather than simply handing a paper copy to the lecturer. This was especially noted by the local students, who had access to the lecturer. It is important to make the process as simple as possible by giving students access to a scanner or photocopier which can scan and email a document to the student. UoN students had access to a scanner. At UoW, administrative staff were available to scan assignments if necessary.

In 2009, the Sample Surveys component had five 'weekly' assignments and a project. This was reduced to three fortnightly assignments and a project in 2010, partly to align with the Experimental Design component but primarily to reduce the student burden, especially in regard to the scanning process.

5. Student results

The final marks in STAT335 and STAT3170 are given below for the 2009 and 2010 cohorts. The numbers of students at UoN were approximately half those at UoW although the numbers are small. In 2009, UoN had a particularly strong group of students. A Mann-Whitney test shows a near significant difference between UoW and UoN in 2009 (p=0.06). In 2010, although the means are almost equal, the variation is much greater at Wollongong, mostly due to two failures. A Mann-Whitney test shows no significant difference between the 2 groups in 2010 (p=0.67).

UoN		UoW
	3	0
	4	0
	5	2447
940	6	8
	7	013
982	8	0

Stem width:10

Site	*n*	Mean	sd.
UoW	11	59.5	15.0
UoN	6	75.7	12.5
Total	17	65.2	15.9

Table 1: 2009 Results

UoN		UoW
	2	7
	3	
	4	2
	5	
32	6	
3	7	123
0	8	56
	9	2

Stem width:10

Site	*n*	Mean	sd.
UoW	8	69.1	22.5
UoN	4	69.9	8.5
Total	12	69.4	18.5

Table 2: 2010 Results

6. Student Feedback

In both years, all students were given a short questionnaire which specifically asked about the mode of teaching delivery. The main themes and number of comments (given in parentheses) are given below.

6.1. 2009 Videoconference

From UoN students:

- Lost class time due to technical difficulties (4).
- Difficulty in asking questions (3).
- Assignments tedious to hand in (2).
- Should be a cheaper course on our HECS debt (2).

In particular, from one student "*An interesting and useful subject. It was good to have lecturers teaching material from their fields so they could give real-life examples*" and from another UoN student "*There was no problem with me to use this way of communication. This is the second course for me*".

From UoW students:

- Lost class time due to technical difficulties (7).
- Assignments tedious to hand in (3).
- Prefer to have use of whiteboard (2).
- Didn't like the split screen with the UoN projected image in one corner of the presentation (3).
- Allows students to study different area that would not otherwise be available (1).

6.2. 2010 Access Grid

From UoN students:

- The Access Grid worked well (2).
- Difficulty in asking questions (1).
- The smartboard was used effectively (2), made use of saved smartboard file on eLearning (3).
- Set up consultation time over Access Grid (1).

From UoW students:

- Improve sound (3).
- The smartboard was used effectively (6), made use of saved smartboard file on eLearning (6).
- Allows students to study different area that would not otherwise be available (1).

7. Potential difficulties

Considerable effort is required to offer the subject to more than one University. A coordinator at UoN was needed to assist with this, although the number of hours required for this was not great (particularly in the second year, as we gained experience). The two Universities have different rules for the presentation of Subject Information sheets and the cover pages of examination papers. Printed Subject Notes had to be posted to UoN. The UoN students had to be given access to the UoW computing system so that they could use eLearning space, and had to be told how to use it. Classes could be held only when the AG rooms at both campuses were available. A common time for a final examination had to be arranged. UoN staff had to post the examination papers to UoW for marking. Final marks for UoN students were only 'recommendations' until approved by UoN.

We do not wish to overemphasize these difficulties. With goodwill, all of these potential problems were dealt with, and we are very grateful to all concerned for their cooperation. Nevertheless, it is necessary to be aware of these matters.

The cost of having a technician on hand throughout a class is considerable but small relative to the cost of running STAT335 and STAT3170 separately. It is hoped that costs can be reduced as we gain experience and the technology matures.

8. Conclusion

The small number of students at both universities suggests that running a joint subject is worthwhile. The results from the two cohorts show that UoN students are not disadvantaged and perform as well as or even better than the UoW students. We can learn from our experiences from the last two years, given student feedback and experience of the lecturers.

8.1. What have we learnt?

- AG technology is much more reliable and conducive to learning than videoconferencing. (This contradicts our earlier expectation).
- Saving the output from the smartboard is particularly useful for students. Feedback mentioned that it assisted students to check notes taken in class, or to catch up if they missed a class.
- It is necessary to simplify the process of getting off-site students access to UoW eLearning, and to ensure that students are aware of having to change passwords within 90 days.

8.2. How can we improve?

- Make the process of asking questions more comfortable for students - give more opportunities.
- Set up a more formal consultation time for Newcastle students over AGR.
- Produce a short video or provide simple step-by-step instructions on how to upload an assignment. A practice session in scanning a document and uploading in first week may be useful.
- Provide a simple 'how to' document on getting into the eLearning site.
- Use the smartboard more for interaction.
- Capture each slide that appears on the smartboard, not just the ones annotated.
- Consider whether to hold two laboratory classes in the session. This would require a tutor for off-site students. One possibility is to get students to bring their laptops to the AGR and have a 'tutor' in the room at UoN. Students without laptops could look on with students with laptops.

References

[1] ASEARC, Interactive media, accessed 29/11/2010, http://www.uow.edu.au/informatics/maths/research/groups/asearc/nteractivemedia/index.html (2010).

[2] R. Caladine, Teaching and learning with videoconference, Tech. rep., Centre for Educational Development and Interactive Resources, University of Wollongong (2008).

[1]The authors would like to thank the anonymous referee for their helpful suggestions.

Experiences with clickers in an introductory statistics course

Alice M. Richardson

University of Canberra, Canberra, ACT, 2601, AUSTRALIA
Alice.Richardson@canberra.edu.au

Abstract

This paper reports on a pilot study of the use of a student response system, commonly known as clickers, in an introductory statistics course. Early results show a small but significant increase in test results following the introduction of clickers. Affective aspects of clickers are also discussed, based on the results of a survey of student attitudes.

Keywords: Statistics Concept Inventory, statistics education

1. Introduction

Clicker technologies were first used in educational settings in the 1960s [1]. Before that they had been used by businesses to collect data in meetings, and by government to collect and display votes in parliamentary settings, for example. Much academic literature makes reference to television programs that also use audience response systems, such as "Who wants to be a Millionnaire?".

An array of different disciplines have tried clickers in classes, ranging from nursing [2] and biology [3] to environmental science [4] and finance [5].

2. Educational setting

Introduction to Statistics at the University of Canberra is a service unit that consisted of three hour lectures and one hour tutorial per week. Although no extraneous mathematics was introduced with the statistical concepts, the unit contained a significant amount of theory and formulae. There was some attempt to contextualize the learning however students often failed to acquire the Statistical language appropriate for application in their education and future professions. We recognized this as a key area for improvement and following a successful bid for institutional teaching and learning development funds, we set about revamping the way in which statistics was taught, in order to make the subject more successful for the students by utilizing some language learning strategies.

2. Project description

The project was concerned with redesigning the delivery of statistics to first year undergraduates and postgraduates with no prior Statistics study. In line with recommendations [6] we initially concentrated on what it was that we felt students needed to know and be able to do following the teaching sessions. The goals of the project were to:

• ensure that students were equipped with skills in interpreting data that would enhance their future performance in their chosen field of study;
• improve students' ability to use data to inform their practice in their chosen area of study, thus arriving at greater understanding of the underlying statistical principles;
• increase the amount of learner-centeredness in appreciating the role of statistics;
• develop online materials that allowed students flexible access, and more opportunities to interact with materials at their own pace.

This project concentrates on benefits that students perceive and display over one semester only. A future longitudinal study could address benefits over longer periods of time.

Initial implementation of the strategies above took place in 2008, and are described in [7].

As well as all the strategies described above, clickers were used on a weekly basis during semester. The questions were multiple choice questions, with generally just three choices, gathered from [8]. There is no anecdotal evidence that students realised the source of the questions. If they had, they would have been able to prepare for the clicker questions, but there are 10 multiple choice questions per chapter so a lot of extra effort for little direct reward.

TurningPoint software, which connects to Microsoft Powerpoint, was employed to run the clicker sessions. One question had to be removed from the analysis when it was discovered that due to a cut-and-paste error, the three multiple-choice options were identical! This did not stop students selecting a variety of responses, which caused much laughter in class when the mistake was discovered. For more very honest and personal experiences with clickers in class, see [9].

A practice clicker session was held in week 1, with students responding to questions such as "What is your gender?" and "Which degree are you enrolled in?" During the course of the semester, clickers were handed out at the start and collected at the end of each class that they were required. None were lost during the semester, and only one clicker stopped working during the semester. Class size was small (no more than 40 on any one day) so not a great deal of time was lost in this exercise. See [10] for a discussion of the handling of clickers in larger classes.

3. The clicker experience during the semester

The number of students who answered clicker questions varied from 10 to 24, with a mean of 19 and a standard deviation of 3.

An ANOVA was conducted to test whether question type (descriptive, graphical, probability, design and inference) has a significant effect on the mean percent correct. There was no significant difference ($F = 0.193$, $p = 0.939$, $df = 4$ and 20). This suggests that clickers can be used successfully across all parts of the introductory statistics curriculum.

We identify three main patterns of response across the 24 questions that had three alternative responses.

1. Clear majority, where the percentage choosing one response (not necessarily the right one) is more than the sum of the percentages choosing the other two responses. Twenty of the 24 questions were answered in this way. Recall that these questions were typically administered at the end of a lecture on a given topic, and so it is not surprising that students knew the answer. Only once did the majority get it wrong: in a question on the stems of stem-and-leaf plots, the majority chose 00, 10, 20, …, 90 when they were supposed to choose 0, 1, 2, …, 9. It is hardly surprising that a majority of students were led astray in their selection of a response when two of them are so similar and almost come down to a typesetting issue! We also note that the questions were taken directly from [8]. The questions are "straightforward questions about basic facts from the chapter" and students are told that they "should expect all of your answers to be correct".
2. Split across two, either with percentages such as 50/45/5. Two questions went this way. One was a question about the scale invariance of the correlation, and the other involved students recognising that a lengthy description was of a completely randomised experiment.
3. Split across three e.g. 38/23/38. Three questions went this way. One involved calculating a five-number summary, one involved identifying a null hypothesis from a lengthy description and the other involved identifying an experiment that was not matched pairs from three lengthy descriptions.

From the very small number of questions that led to a split, it appears that problems can arise when a little bit of calculation is required, or when there is a fair amount of reading (70 - 80 words). Students also complained when questions were presented that required a page of output, followed by questions on the next slides. Our advice is to provide students with a handout with attached graphs and output to assist with timely comprehension of the background to the questions.

The responses to clicker questions cannot be tied to individual students, although some systems do allow for this to be done. Anonymity of response is a selling point to some students, while other lecturers use the clickers as part of formal assessment and therefore need to be able to record individual students' responses [11].

4. Results at the end

Student learning outcomes were measured through four in-class tests, a final exam, and a Statistics Concept Inventory [12].

Table 1 shows that while the means of the control group (CG) and the experimental group (EG) for test 1 and 2 are not significantly different, for test 3 the differences in means are significant at the 5% level. For test 4, the means are significantly different but EG scores lower than CG. This may be due to the fact that there were two significant errors in the EG paper, and furthermore different topics were tested in the two groups: two-sample hypothesis tests and regression in EG and regression and analysis of variance in CG.

Table 1. Results of assessment items for CG and EG.

	group	
	language (2008, CG)	clickers (2010, EG)
assessment	mean (s.d.) n = 22	mean (s.d.) n = 28
Test 1 (30)	21.8 (4.7)	22.1 (5.0)
Test 2 (30)	20.5 (4.7)	20.1 (5.8)
Test 3 (30)	19.1 (6.7)	20.0 (4.4)
Test 4 (30)	20.5 (5.7)	21.0 (5.7)
Exam (100)	61.5 (21.5)	65.2 (15.2)

Table 2 shows the distribution of grades in the final examination for 2008 and 2010, after eliminating students who passed the unit on in-term assessment and chose not to attempt to improve their grade by sitting the examination. A χ^2 test shows that there are no significant differences ($X^2 = 4.05$, $p = 0.3944$, $df = 4$) between the grade distributions for the two groups.

Table 2. Grade distribution for CG and EG.

	group	
	language (2008, CG)	Clickers (2010, EG)
grade	percent n = 22	percent n = 28
Fail	5.0	10.7
Pass	40.0	35.7
Credit	5.0	21.4
Distinction	35.0	17.8
High Distinction	15.0	14.3
Total	100.0	99.9

It is interesting to note that EG produced a higher percentage of credit grades compared to CG. Data on UAIs and GPAs for these students was too sparse to be able to confirm the effect in CG where Credit grades were under-represented compared to Pass or a Distinction grades.

Other research [13] found a correlation of 0.57 between clicker scores and final exam marks. His sample size is reported to be 24, although the accompanying scatter plot appears to have about 40 observations. We are not able to comment on the correlation in this data, as it is not possible to link clicker scores and exam marks. It is however possible to link attitudes towards clickers and exam marks, as discussed in the next section.

5. Affective aspects of clickers

Six questions asked on a five-point Likert scale, scored from 1 = strongly disagree to 5 = strongly agree:

- I used the clickers in class
- The clicker questions helped my understanding during class.
- The clicker questions helped my revision.
- I would recommend the use of clickers to other statistics classes in particular.
- I would recommend the use of clickers to other university classes in general.
- Please add any comments.

If the clicker scores are summed, the maximum possible score is 25. The 22 scores obtained were approximately Normally distributed, with a mean of 19.95 and a standard deviation of 2.87. This suggests that the recorded reaction to clickers was positive, which may however mean that only those who had enjoyed using the clickers may have bothered to respond.

So the students who responded to the clicker survey were uniformly positive about it, but did those students actually perform better than those who did not respond? Independent-samples t tests comparing test and exam marks between the 22 students who did respond to the clicker survey and the 11 students who did not show no significant difference in mean scores ($p = 0.286$, 0.226, 0.624, 0.831 and 0.290 for tests 1 – 4 and the exam respectively.) So while they did not show a significant difference in their mean test results, we can take this to indicate that a student's attitude to the clickers may not influence their exam result, but their usage of them still might. Only one student who responded to the clicker survey had used them less than "three or four times".

Another way to look at whether good students use clickers is to study the relationship between the qualitative aspects of clicker use and exam marks. The relationship is weakly positive and not statistically significant ($r = 0.337$, $p = 0.202$, $n = 16$).

Students were also invited to comment on the clickers. Eleven students took up the invitation and their comments were uniformly positive. Some reinforced the positive aspects of clickers e.g. "It can help the students review the key points and find the weakness quickly"; "Clickers give us opportunity to learn from our mistakes".

Some recommended greater use e.g. "I think we might have attempted more clickers, if we had more time during our semester."

Some recommended variations in use. "I believe the clickers are excellent for on the spot feedback [to] help direct [the] lecture. The part that is hard to tell is when further explanation is given, how to tell that it is understood? Should there be a second similar question given after the further explanation from the first, this could see if the results increase to show that students understand!" It is worth reiterating that the questions selected were revision questions from [8] and the author himself emphasises that by the end of a

chapter, students should be able to answer all these questions correctly.

"I reckon clickers are a very good tool to understanding more in lectures, however you lose out when you cannot attend lectures and listen to them online." Lectures were recorded using the audio-only system Audacity. A system such as Echo360. which captures both voice and Powerpoint, should improve the capacity of students listening asynchronously to benefit from the clicker questions

Finally, there were several general comments such as "Thanks for making statistics easy by adding activities like clickers".

6. Conclusions

A distinction should be made between retention of knowledge from the first course to subsequent ones, and capstone courses. A capstone Statistics course, using a book such as [14] or [15], aims to integrate knowledge from many previous courses, not just one. Typically such a course is case study-based, involving a mixture of basic techniques such as exploratory data analysis with advanced techniques such as generalised linear and multilevel models.

We have described above a project that brought together experts from language teaching and statistics to produce a new teaching/learning model for beginning students. We evaluated this new approach and demonstrated that we have made a statistically significant difference to students' performance in tests and exams, improved their understanding of some key concepts and their ability to view statistics in relation to their professional practice. It is this multi-sensory approach to delivery that we would recommend to others. A caveat is that the groups of students involved in this project were fairly small (around n = 25). Consequently, we need to act with caution if we are to apply these new teaching strategies to a larger group of students, say, of size 200. Admittedly, the size of the student cohort might make group work more difficult but with attention paid to management issues related to a larger size group, it is still possible to implement such group work. For some group work management ideas, see [16].

References

[1] Judson, E. and Sawada, D. (2002). Learning from past and present: electronic response systems in college lecture halls. *Journal of Computers in Mathematics and Science Teaching* 21, 167 – 181.

[2] Berry, J. (2009). Technology support in nursing education: clickers in the classroom. *Nursing Education Research*, 30, 295 – 298.

[3] Crossgrove, K. & Curran, K.L. (2008). Using clickers in nonmajors- and majors-level biology courses: student opinion, learning and long-term retention of course material. *CBE-Life Sciences Education*,7, 146 – 154.

[4] Czekanski, A.J. & Roux, D.-M. P. (2008). The use of clicker technology to evaluate short- and long-term concept retention. *Proceedings of the 2008 ASEE Zone 1 Conference*, West Point, NY.

[5] Snavely, J.C. and Chan, K.C. (2009). Do clickers 'click' in the classroom? *Journal of Financial Education*, 35, 42 – 65.

[6] Garfield, J. (1995). How students learn statistics. *International Statistical Review*, 63(1), 25-34.

[7] Richardson, A.M. (2009). Retention of knowledge between statistics courses: results of a pilot study. Proceedings of the 3rd ASEARC Research Conference, Newcastle, December 7 – 8, 2009.

[8] Moore, D.S. (2010). *The Basic Practice of Statistics*, 5th edition. New York: Freeman.

[9] Beavers, S.L. (2010). Some days, things just "click in the classroom clicker technology in the Introductory US Politics classroom. 2010 Meeting of the Western Political Science Association, April 1 – 3, 2010.

[10] Barnett, J. (2006). Implementation of personal response units in very large lecture classes: Student perceptions. *Australasian Journal of Educational Technology*, 22, 474-494.

[11] Lowery, R.C. (2006). Clickers in the classroom: a comparison of interactive student-response keypad systems. National Technology and Social Science Conference, Las Vegas, 6 April 2006.

[12] Stone, A., Allen, K., Rhoads, T.R, Murphy, T.J., Shehab, R.L. and Saha, C. (2003). The Statistics Concept Inventory: pilot study. 3rd ASEE/IEEE Frontiers in Education Conference, November 5 – 8, Boulder CO.

[13] Lucas, A. (2009). Using peer instruction and I-clickers to enhance student participation in calculus. *Primus* 19, 219 – 231.

[14] Chatfield, C. (1995). Problem Solving: A Statistician's Guide. London: Chapman and Hall.

[15] Spurrier, J.D. (2000). The Practice of Statistics: Putting it all Together. Pacific Grove, CA: Duxbury.

[16] Ebert-May, D., Brewer, C., and Allred, S. (1997). Innovation in large lectures – teaching for active learning. *BioScience*, 47(9), 601–607.

Were Clopper & Pearson (1934) too careful?

Frank Tuyl

The University of Newcastle, Australia
frank.tuyl@newcastle.edu.au

Abstract

The 'exact' interval due to Clopper & Pearson (1934) is often considered to be the gold standard for estimating the binomial parameter. However, for practical purposes it is also often considered to be too conservative, when mean rather than minimum coverage close to nominal could be more appropriate. It is argued that (1) Clopper & Pearson themselves changed between these two criteria, (2) 'approximate' intervals are preferable to 'exact' intervals, and (3) approximate intervals are well represented by Bayesian intervals based on a uniform prior.

Key words: Exact inference, confidence interval, binomial distribution, Jeffreys prior, Bayes-Laplace prior

1. Introduction

The 'gold standard' for estimating the binomial parameter is due to Clopper & Pearson (C&P) [1]. The $(1-\alpha)100\%$ C&P interval is based on the inversion of two separate hypothesis tests, the resulting lower limit θ_l being calculated from

$$\sum_{r=x}^{n} \binom{n}{r} \theta^r (1-\theta)^{n-r} = \alpha/2 \tag{1}$$

and the upper limit θ_u from

$$\sum_{r=0}^{x} \binom{n}{r} \theta^r (1-\theta)^{n-r} = \alpha/2. \tag{2}$$

Interestingly, using the relationship between binomial summations and beta integrals, the central Bayesian interval corresponding to a beta(a, b) prior follows from equating

$$I_\theta(x+a, n-x+b) = \sum_{r=x+a}^{n+a+b-1} \binom{n+a+b-1}{r} \theta^r (1-\theta)^{n+a+b-1-r} \tag{3}$$

to $\alpha/2$ to obtain the lower limit and doing the same with

$$1-I_\theta(x+a, n-x+b) = \sum_{r=0}^{x+a-1} \binom{n+a+b-1}{r} \theta^r (1-\theta)^{n+a+b-1-r} \tag{4}$$

to obtain the upper limit, where I_θ is the incomplete beta function. It follows that the C&P lower limit can be seen to correspond to a beta$(0, 1)$, and the C&P upper limit to a beta$(1, 0)$ prior. (Calculation of C&P intervals by using the inverse beta distribution, in Excel for example, is much more straightforward than using the inverse F distribution suggested by, among many others, Blaker [2].) This means that, compared with the Bayesian interval based on the (uniform) beta$(1, 1)$ or Bayes-Laplace (B-L) prior, the C&P lower limit is based on subtracting a success from the sample and the C&P upper limit on subtracting a failure. As a result, strictly speaking the C&P lower (upper) limit calculation breaks down when $x = 0\,(n)$ and is set to $0\,(1)$.

The effect of including x in the binomial summations (1) and (2) is that the C&P interval is 'exact': frequentist coverage, defined as $C(\theta) = \sum_{r=0}^{n} p(r|\theta) I(r, \theta)$, is at least equal to nominal for *any* value of θ, as illustrated in Figure 1. (Here $p(r|\theta)$ is the binomial pdf and $I(r, \theta)$ an indicator function: it is 1 when the interval corresponding to outcome r covers θ and 0 otherwise.) In fact, the mid-P interval [3, 4] is based on including *half* of $p(x|\theta)$ in (1) and (2), leading to an 'approximate' interval: this family of intervals aims for mean coverage to be close to nominal without compromising minimum coverage too much. The common Wald interval, based on the standard Normal approximation, is a poor example due to its serious below-nominal coverage, as also shown in Figure 1.

Similar to the Wald interval, the central B-L interval has zero minimum coverage (near the extremes). This is avoided by hybrid intervals that are one-sided for $x = 0\,(n)$, central otherwise [5], but a better interval is the one based on highest posterior density (HPD) and shown in Figure 1. In fact, all B-L intervals have *nominal* mean coverage, and the HPD interval performs well with respect to minimum coverage also.

Our discussion of this Bayesian interval is relevant as in Section 2 we point to an apparent contradiction in

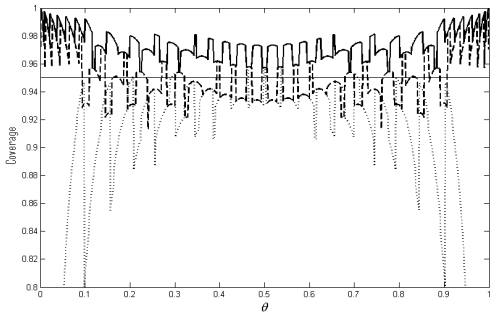


Figure 1: Coverage of the binomial parameter for $n = 30$ and $\alpha = 0.05$: Clopper & Pearson (solid), Bayes-Laplace HPD (dashed) and Wald (dotted) methods.

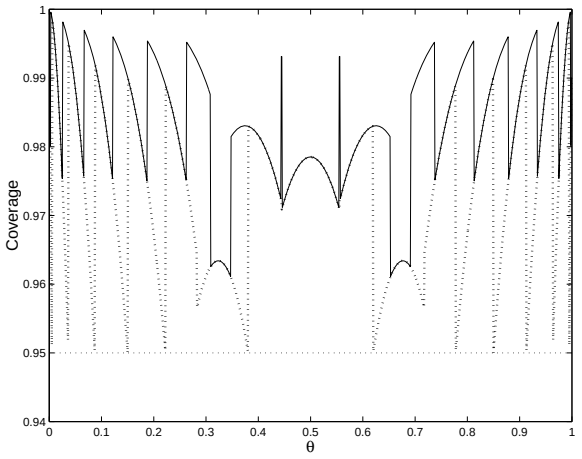


Figure 2: Coverage of the binomial parameter for $n = 10$ and $\alpha = 0.05$: Clopper & Pearson (solid) and Blaker (dotted) methods.

Clopper & Pearson's article and in Section 3 we argue that exact intervals may be seen to be more conservative than is usually shown in coverage graphs. In Section 4 we suggest the B-L HPD interval as an excellent representative of the 'approximate' family.

2. "Statistical experience"

It appears that C&P contradicted themselves with respect to which criterion, minimum or mean coverage, is more reasonable. On their first page (p.404), after a rather Bayesian reference to the *probability* of a parameter lying between two limits, they stated, "In our statistical experience it is likely that we shall meet many values of n and x; a rule must be laid down for determining θ_l and θ_u given n and x. Our confidence that θ lies within the interval (θ_l, θ_u) will depend upon the proportion of times that this prediction is correct in the long run of statistical experience, and this may be termed the confidence coefficient."

However, after showing graphically intervals for $n = 10$, C&P [1, p.406] changed the meaning of 'statistical experience': "It follows that in the long run of our statistical experience from whatever populations random samples of 10 are drawn, we may expect at least 95% of the points (x, θ) will lie inside the lozenge shaped belt, not more than 2.5% on or above the upper boundary and not more than 2.5% on or below the lower boundary." The addition of "at least" is understandable due to the discreteness of the Binomial distribution, but the "random samples of 10" phrase is crucial. We argue that in effect C&P and other exact intervals are based on the assumption, in addition to the hypothetical repeated sampling concept, that Nature is not only malicious, but omniscient as well: in effect, the exact method prepares for Nature choosing a true 'bad' value of θ based on knowledge of the sample size *and* level of confidence involved!

In fact, for the choice $n = 10$, C&P coverage is strictly above-nominal, which is improved elegantly (see Figure 2) by Blaker's method [2], based on considering confidence curves; due to its nesting property, the Blaker interval appears superior to other short exact intervals [6, 7, 8]. However, in the next section we argue that, from a practical point of view, *all* exact intervals are overly conservative.

3. "Approximate is better than exact"

The title of this section is a reference to Agresti & Coull (A&C) [9] who argued in favour of approximate intervals for most applications. We agree with their statement (p.125) that even though such intervals are technically not confidence intervals, "the operational performance of those methods is better than the exact interval in terms of how most practitioners interpret that term.". The implication is that, from a practical point of view, "narrower intervals for which the actual coverage probability could be less than .95 but is usually quite *close* to .95" are preferable.

Similarly, Brown, Cai & DasGupta (BCD) [5, p.113] considered the C&P approach "wastefully conservative and not a good choice for practical use, unless strict adherence to the prescription [coverage $\geq 1 - \alpha$] is demanded." It seems clear that A&C and BCD criticised the C&P interval because they consider mean coverage a better criterion than minimum coverage. We now show that even with respect to minimum coverage, C&P and other exact intervals are conservative.

A purist frequentist could claim that eventually a true physical constant, in the form of a proportion, could become known with substantial accuracy and that if this value were to be an "unlucky" one, exactness of previously calculated intervals would have been preferable. However, this argument is weakened if sample sizes are allowed to vary over time, as we now illustrate.

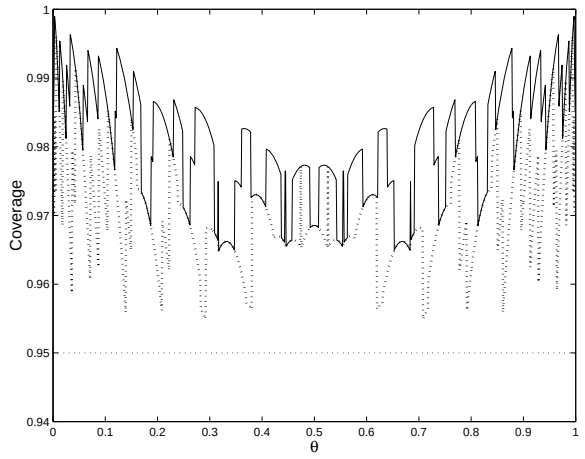


Figure 3: Coverage of the binomial parameter based on $\alpha = 0.05$ and averaging $n = 10$ and $n = 20$: Clopper & Pearson (solid) and Blaker (dotted) methods.

Supposing that the practitioner did always want to apply $\alpha = 0.05$, if in fact they considered a varying sample size (under repeated sampling), immediately Nature's scope for choosing 'bad' values of θ would be greatly reduced. Even short exact methods like Blaker's [2] turn strictly conservative as soon as repeated sampling takes place for two sample sizes (instead of one). Figure 3 is an example of this, based on adding $n = 20$ to the fixed $n = 10$ from Figure 2.

Thus, we do not even need "many values of n" to question the need for exact intervals! Finally, Figure 3 seems to lend even greater support to Stevens [10], who also argued against exact limits (p.121):

> It is the very basis of any theory of estimation, that the statistician shall be permitted to be wrong a certain proportion of times. Working within that permitted proportion, it is his job to find a pair of limits as narrow as he can possibly make them. If, however, when he presents us with his calculated limits, he says that his probability of being wrong is less than his permitted probability, we can only reply that his limits are unnecessarily wide and that he should narrow them until he is running the stipulated risk. Thus we reach the important, if at first sight paradoxical conclusion, that *it is the statistician's duty to be wrong* the stated proportion of times, and failure to reach this proportion is equivalent to using an inefficient in place of an efficient method of estimation.

In fact, Stevens showed that by taking a randomised weighted average of the two beta distributions implied by the C&P approach, the coverage probability, for *any* value of θ, equals the nominal confidence level. However, this approach leads to serious practical problems; for example, the resulting interval is typically asymmetric around $x = n/2$ if that is in fact the sample outcome, which is surely unacceptable from a practical point of view. About this interval, Lancaster [4] stated, "In certain experimental situations, this procedure would be time-consuming and even embarrassing to the statistician." It would thus appear that a 'regular' approximate interval with moderate below-nominal minimum coverage is preferable.

4. Which 'approximate' interval is best?

As stated in the introduction, intervals based on the B-L prior have nominal mean coverage, which, as far as we know, is not achieved by any 'approximate' intervals. The B-L HPD interval adds reasonable minimum coverage to this property. A more practical requirement dictates that no *individual* intervals are unreasonably short or wide, arguably satisfied by this likelihood-based interval also, but not necessarily by the approximate intervals recommended by BCD, for example.

It is unfortunate that review articles of approximate methods, such as the one by BCD, tend to consider, as representatives of the Bayesian approach, intervals based on the Jeffreys beta$(\frac{1}{2}, \frac{1}{2})$ prior only. For example, minimum coverage of the Jeffreys HPD interval converges (as $n \to \infty$) to 84.0%, as opposed to 92.7% for the B-L HPD interval. Due to the Jeffreys prior's weight near the extremes, corresponding intervals appear particularly short when $x = 0\,(n)$, simultaneously causing this low minimum coverage. Even the hybrid Jeffreys interval recommended by BCD, despite their earlier criticism of methods that are data-based [5, p.106], has limiting minimum coverage of only 89.8%. These results follow quite simply from considering Poisson intervals.

A&C derived simple approximate intervals from the Score method due to Wilson [11]. These have excellent minimum coverage, at the expense of somewhat conservative mean coverage. (Note that, in contrast, the Score interval has mean coverage closer to nominal, but limiting minimum coverage of 83.8%.) However, individual Score and A&C intervals are undesirable when they are wider than corresponding C&P intervals: when this occurs, it would seem difficult to justify such an approximate interval to a client, based on the notion that the C&P interval is 'conservative'!

It is no surprise that Score-based intervals have this undesirable property, as the Normal approximation they are based on is inadequate when x is close to 0 or n. The mid-P interval appears to be a better choice, and, in fact, is quite similar to the B-L HPD interval for such x. However, as pointed out by A&C also, the mid-P interval is conservative (on average) for small n; for $n \leq 4$, for example, coverage is strictly above-nominal.

In short, we propose the B-L HPD interval as the preferred candidate for estimation of the binomial parameter, from both Bayesian and frequentist points

of view; see also [12] and [13]. About HPD intervals, BCD stated, "The psychological resistance among some to using this interval is because of the inability to compute the endpoints at ease without software.", but implementation is straightforward, even in Excel (using its Solver function). To emphasise, in effect, this probability-based interval is derived from the normalised likelihood function and has the desirable property that no values outside it have higher likelihood than the values inside it.

5. Conclusion

We conclude that based on their original intention, which was to allow for "many values of x and n", C&P's interval is too conservative: when allowing n to vary, coverage is strictly above-nominal for all values of θ. Short exact methods only address C&P conservativeness resulting from the dual one-sided testing aspect, and lead to similar behaviour. We consider this to be an additional argument against exact methods, which appear inconsistent with C&P's "long run of statistical experience".

We suggest that the B-L HPD interval, with its mean coverage equal to nominal and minimum coverage that would seem acceptable for most practical purposes, is preferable to the approximate intervals recommended by BCD, and should be adopted by Bayesians and frequentists alike.

References

[1] C. J. Clopper, E. S. Pearson, The use of confidence or fiducial limits illustrated in the case of the binomial, Biometrika 26 (1934) 404–416.
[2] H. Blaker, Confidence curves and improved exact confidence intervals for discrete distributions, The Canadian Journal of Statistics 28 (4) (2000) 783–798.
[3] H. O. Lancaster, The combination of probabilities arising from data in discrete distributions, Biometrika 36 (3/4) (1949) 370–382.
[4] H. O. Lancaster, Significance tests in discrete distributions, Journal of the American Statistical Association 56 (294) (1961) 223–234.
[5] L. D. Brown, T. Cai, A. DasGupta, Interval estimation for a binomial proportion, Statistical Science 16 (2) (2001) 101–133 (with discussion).
[6] E. L. Crow, Confidence intervals for a proportion, Biometrika 43 (1956) 423–435.
[7] C. R. Blyth, H. A. Still, Binomial confidence intervals, Journal of the American Statistical Association 78 (381) (1983) 108–116.
[8] G. Casella, Refining binomial confidence intervals, The Canadian Journal of Statistics 14 (1986) 113–129.
[9] A. Agresti, B. A. Coull, Approximate is better than "exact" for interval estimation of binomial proportions, The American Statistician 52 (2) (1998) 119–126.
[10] W. L. Stevens, Fiducial limits of the parameter of a discontinuous distribution, Biometrika 37 (1-2) (1950) 117–129.
[11] E. B. Wilson, Probable inference, the law of succession, and statistical inference, Journal of the American Statistical Association 22 (158) (1927) 209–212.
[12] F. Tuyl, R. Gerlach, K. Mengersen, A comparison of Bayes–Laplace, Jeffreys, and other priors: the case of zero events, The American Statistician 62 (1) (2008) 40–44.
[13] F. Tuyl, R. Gerlach, K. Mengersen, The Rule of Three, its variants and extensions, International Statistical Review 77 (2) (2009) 266–275.

Principles in the design of multiphase experiments with a later laboratory phase: orthogonal designs

C.J. Brien

University of South Australia, Australia
chris.brien@unisa.edu.au

B.D. Harch

CSIRO Mathematical and Information Sciences, Australia
Bronwyn.Harch@cmis.csiro.au

R.L. Correll

CSIRO Mathematical and Information Sciences, Australia
Rho.environmetrics@bigpond.com

R.A. Bailey

Queen Mary University of London, United Kingdom
r.a.bailey@qmul.ac.uk

Abstract

It is common for the material produced from field and other experiments to be processed in a laboratory. Reasons for this include the need to measure chemical and physical attributes using equipment such as spectrometers, gas chromatographs, pH meters or wear and strength testers, or to produce processed products such as wine, bread and malt that are subsequently assessed, often by an expert panel. These experiments are multiphase. They occur widely in agriculture, food processing and pharmaceutical industries and biological and environmental sciences, although their multiphase nature is often not recognized. A systematic approach to designing the laboratory phase of such multiphase experiments, taking into account previous phases, will be described. We extend the fundamental principles of experimental design — randomization, replication and blocking — to provide general principles for designing multiphase experiments that employ orthogonal designs. In particular, the need to randomize the material produced from the first phase in the laboratory phase is emphasized. Factor-allocation diagrams are employed to depict the randomizations in a design and skeleton analysis-of-variance (ANOVA) tables to evaluate their properties. The techniques are applied to several scenarios for an athlete training experiment.

Key words: Analysis of variance, Experimental design, Laboratory experiments, Multiple randomizations, Multi-phase experiments, Multitiered experiments, Two-phase experiments

On the analysis of variety trials using mixtures of composite and individual plot samples

Brian Cullis

University of Wollongong and Centre for Mathematics, Informatics and Statistics, CSIRO, Australia
bcullis@uow.edu.au

Alison Smith

Wagga Wagga Agricultural Institute, Industry & Investment NSW, Australia
alison.smith@industry.nsw.gov.au

David Butler

Primary Industries & Fisheries, Department of Employment, Economic Development and Innovation, Toowoomba, QLD, Australia
david.butler@deedi.qld.gov.au

Robin Thompson

Rothamsted Research, Harpenden and Queen Mary College, University of London, London UK

Abstract

Field trials for evaluating plant varieties are conducted in order to obtain information on a range of traits including grain yield and grain quality. Grain quality traits are often costly to measure so that it is not possible to measure all plots individually. It has therefore become common practice to measure a single composite sample for each variety in the trial but this practice precludes the use of a valid statistical analysis. In this talk we propose an approach in which a proportion of varieties be measured using individual uncomposited samples. The remaining varieties may either be tested as composite samples or with reduced replication. This approach allows application of a valid statistical analysis using a so-called hybrid mixed model which uses a mixture of composited and uncomposited samples.

Key words: composite samples, mixed models, spatial analysis

On the design of experiments with an embedded partially replicated area

David Butler

Primary Industries & Fisheries, Department of Employment, Economic Development and Innovation, Toowoomba, QLD, Australia
david.butler@deedi.qld.gov.au

Alison Smith

Wagga Wagga Agricultural Institute, Industry & Investment NSW, Australia
alison.smith@industry.nsw.gov.au

Brian Cullis

University of Wollongong and Centre for Mathematics, Informatics and Statistics, CSIRO, Australia
bcullis@uow.edu.au

Abstract

Economically important crop varietal traits are typically measured in a field trial phase for productivity characteristics, and subsequent laboratory phases for end-product suitability. Cost, or laboratory throughput constraints can prohibit the subsequent testing of grain samples from every experimental unit in the first phase of testing. Non-genetic effects are often evident in traits measured in all phases, and it is important that a valid statistical framework exist for the analysis of all traits, including those with economic or volume constraints. A compromise sampling scheme has been proposed for expensive phase one traits, whereby a valid partially replicated design for all varieties is embedded in a contiguous area within the field experiment to reduce sample numbers, and hence measurement costs. Less expensive traits are measured on the complete experiment.

Approaches to the design of these (first phase) trials with an embedded area will be described, both in the context of individual and series of experiments across multiple locations. Optimal design provides a flexible model based framework for constructing experimental designs in these cases, and a computational approach will be discussed that extends to incorporating known fixed structures, such as genetic relationships among individuals. The utility of this design methodology in a multiphase setting, where experimental units from one phase form the ”treatments” in the next phase, and design optimality is with respect to the crop varieties, will also be considered.

Key words: optimal design, partially replicated designs, multiphase experiments, multi-environment trials

Responsibilities and issues facing young statisticians

Leo K. L. Chow

Young Statisticians Network, Statistical Society of Australia Inc., NSW Branch, Australia
leo.chow11@gmail.com

Abstract

In this talk, we would present current responsibilities and issues facing the Young Statisticians Network (YSN) of the Statistical Society of Australia Inc. (SSAI), in particular the NSW Branch of the YSN. We would first introduce the YSN and a general breakdown of the network. We would then talk about the responsibilities which go hand in hand with the current issues facing Young Statisticians such as professional development, careers and networking opportunities. What was achieved by the YSN in the past 2 years would also be presented. This would be a good opportunity for members of the YS community to provide us with feedback of the activities undertaken by the YSN and to provide suggestions for the network going forward. References: Stephen Bush (2008). Issues facing Young Statisticians.

Key words: Young statisticians, responsibilities, New South Wales

Smooth Tests of Fit for a Mixture of Two Poisson Distributions

D.J. Best, J.C.W. Rayner
The University of Newcastle, Callaghan, NSW, 2308, Australia
John.Best@newcastle.edu.au and John.Rayner@newcastle.edu.au

and O.Thas
Department of Applied Mathematics, Biometrics and Process Control, B-9000 Gent, Belgium
Olivier.Thas@Ugent.be

Abstract

In this note smooth tests of fit for a mixture of two Poisson distributions are derived and compared with a traditional Pearson chi-squared test. The tests are illustrated with a classic data set of deaths per day of women over 80 as recorded in the London Times for the years 1910 to 1912.

Keywords: Central moments, Deaths of London women data, Factorial moments, Orthonormal polynomials, Parametric bootstrap, Pearson X^2 tests

1. Introduction

A Poisson process is often used to model count data. Sometimes an underlying mechanism suggests two Poisson processes may be involved. This may be modelled by a two component Poisson mixture model. The two Poisson mixture applies generally to data more dispersed than that modelled by a single Poisson. An interesting example was given by Leroux and Puterman (1992) who fit a two Poisson mixture to fetal lamb movement data. They say the mixture model "... has a clear interpretation in terms of a ... background rate ... and an excited state." The Poisson probability function, $f(x;\ \theta)$ say, is given by

$$f(x;\ \theta) = \exp(-\theta)\theta^x/x!,\ x = 0, 1, 2, ...,$$
$$\text{in which } \theta > 0$$

and the two component Poisson mixture model has probability function

$$f^*(x;\ \theta_1,\ \theta_2,\ p) = p\ f(x;\ \theta_1) + (1 - p)\ f(x;\ \theta_2),$$
$$x = 0, 1, 2, ..., \text{ in which } \theta_1 > 0,\ \theta_2 > 0,$$
$$\theta_1 \neq \theta_2 \text{ and } 0 < p < 1.$$

A common test of fit for $f^*(x;\ \theta_1,\ \theta_2,\ p)$ is based on the well-known Pearson's X^2. If there are l classes X^2 is approximately χ^2 with $l - 4$ degrees of freedom: χ^2_{l-4}.

In section 2 we look at estimation of the parameters θ_1, θ_2 and p. Section 2 also defines the X^2 test and some smooth tests of fit. Section 3 gives a small power comparison while section 4 considers a classic data set of deaths per day of women over 80 as recorded in the London Times for the years 1910 to 1912.

2. Estimation and Test Statistics

The two most common approaches for estimating θ_1, θ_2 and p are based on moments (MOM) and maximum likelihood (ML). If we have n data points $x_1,\ x_2,\ \ldots,\ x_n$ and $\bar{x} = \sum_{i=1}^{n} x_i / n$ and $m_t = \sum_{i=1}^{n} (x_i - \bar{x})^t / n$, $t = 2, 3, ...$ the MOM estimators satisfy

$$\tilde{p} = (\bar{x} - \tilde{\theta}_2)/(\tilde{\theta}_1 - \tilde{\theta}_2),\ \tilde{\theta}_1 = (A - D)/2,$$
$$\text{and } \tilde{\theta}_2 = (A + D)/2$$

in which

$$A = 2\bar{x} + (m_3 - 3m_2 + 2\bar{x})/(m_2 - \bar{x})$$
$$\text{and } D^2 = A^2 - 4A\bar{x} + 4(m_2 + \bar{x}^2 - \bar{x}).$$

This method clearly fails if $D^2 < 0$, if any of $\tilde{\theta}_1$, $\tilde{\theta}_2$ and $\tilde{p}$ are outside their specified bounds, or if $m_2 = \bar{x}$.

Iteration is needed to find the ML estimates and given the speed of modern computers an EM type algorithm is satisfactory. This will always converge to $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{p}$ within the specified bounds if the initial estimates are also within these bounds. However convergence can be slow – occasionally more than 1,000 iterations – and a local, but not universal, maximum may be found. A grid of initial values is often worth examining. This was not done for the calculations in Table 1 because all the sizes were 0.05 suggesting universal maxima were indeed found. To check on the possibility of a local stationary point it is also useful to examine contour plots of the likelihood surface. This was done for the *Deaths of London Women* example in section 4. The following estimation equations are needed:

$$\hat{p}_k = \frac{\hat{p}_{k-1}\sum_{i=1}^{n} f(x_i, \hat{\theta}_{1,k-1})}{nf^*(x_i; \hat{\theta}_{1,k-1}, \hat{\theta}_{2,k-1}, \hat{p}_{k-1})} \text{ and}$$

$$\hat{\theta}_{r,k} = \frac{\sum_{i=1}^{n} x_i f(x_i, \hat{\theta}_{r,k-1})}{nf^*(x_i; \hat{\theta}_{1,k-1}, \hat{\theta}_{2,k-1}, \hat{p}_{k-1})}, \; r = 1, 2,$$

where $\hat{\theta}_{r,k}$ is the estimate of θ_r at the kth iteration, $\hat{p}_k$ is the estimate of p at the kth iteration and $(\hat{\theta}_{1,0}, \hat{\theta}_{2,0}, \hat{p}_0) = (\tilde{\theta}_1, \tilde{\theta}_2, \tilde{p})$ may be an admissible initial value. See, for example, Everitt and Hand (1981, p.97). Newton's method will sometimes converge to the correct values and when it does the convergence is much quicker than the above estimating equations. However, Newton's method doesn't always converge and may give estimates outside the specified bounds.

Now let O_j be the number of data points equal to j, $j = 0, 1, 2, \ldots$. Let $E_j = nf^*(j; \hat{\theta}_1, \hat{\theta}_2, \hat{p})$. Often classes are pooled in the tail until the greatest l is found such that the expectation of the classes from the lth on is at least 5. Then the Pearson test of fit statistic is

$$X^2 = \sum_{j=1}^{l} (O_j - E_j)^2 / E_j$$

and X^2 is taken to have the χ^2_{l-4} distribution.

Smooth test components V_s can be defined as

$$\hat{V}_s = \sum_{i=1}^{n} g_s(x_i; \hat{\theta}_1, \hat{\theta}_2, \hat{p}) / \sqrt{n}, \; s = 2, 3, \ldots$$

Here $\{g_s(.)\}$ is the set of orthonormal functions on the null distribution. We give formulae, in terms of the population moments $\mu, \mu_2, \ldots, \mu_6$ for the first four orthonormal functions and $\hat{V}_2$ and $\hat{V}_3$ in Appendix A. For the mixture of two Poissons these moments can be calculated from the population factorial moments $\mu'_{[t]} = p\theta_1^t + (1-p)\theta_2^t$. Smooth tests of fit are discussed in detail in Rayner et al. (2009).

Table 1. 100×powers based on 10,000 Monte Carlo samples for $n = 100$ and $\alpha = 0.05$ for a null Poisson mixture with $p = 0.5$, $\theta_1 = 2$ and $\theta_2 = 5$.

Alternative	$\hat{V}_2^2$	$\hat{V}_3^2$	$\hat{V}_4^2$	X^2
Null	5	5	5	5
NB(2, 0.4)	45	39	40	41
NB(3, 0.5)	18	20	20	18
NB(4, 0.5)	19	20	24	27
NTA(1, 2)	79	69	51	54
0.5 × NB(2, 0.4) + 0.5 × NB(2, 0.5)	33	28	30	31
0.5 × NB(2, 0.3) + 0.5 × NB(3, 0.5)	64	48	59	65
NTA(2, 2)	88	66	55	81
NTA(2, 1)	26	26	22	16
NTA(1, 3)	98	94	72	92
P(4)	37	14	13	4
P(6)	33	13	5	10

3. Indicative Size and Power Study

We consider the case $\alpha = 0.05$, $p = 0.5$, $\theta_1 = 2$, $\theta_2 = 5$. Based on 25,000 Monte Carlo samples the critical values of $\hat{V}_2^2$, $\hat{V}_3^2$ and $\hat{V}_4^2$ are 0.31, 0.91 and 0.56 respectively. We note that $\hat{V}_1 \equiv 0$ as shown in Appendix B. We use 17.5 as the X^2 critical value. Table 1 gives some powers for

- negative binomial alternatives with probability function $\binom{m+x-1}{x}\pi^m(1-\pi)^x$ for x = 0, 1, 2, ... with $m > 0$, denoted by NB(m, π),
- Neyman Type A alternatives with probability function $\frac{e^{-\lambda_1}\lambda_2^x}{x!}\sum_{j=0}^{\infty}\frac{j^x}{j!}(\lambda_1 e^{-\lambda_2})^j$ for x = 0, 1, 2, ... with $\lambda_1 > 0$ and $\lambda_2 > 0$, denoted by NTA(λ_1, λ_2) and
- Poisson alternatives $f(x; \theta)$ for x = 0, 1, 2, ... with $\theta > 0$, denoted by P(θ).

In Table 1 no one test dominates but overall perhaps that based on $\hat{V}_2^2$ does best. Double precision arithmetic was used in the Table 1 calculations. In a few cases no estimate was obtained after 10,000 iterations and these cases were discarded.

4. Example: Deaths of London Women During 1910 to 1912

A classic data set considered by a number of authors starting with Whitaker (1914) considers deaths per day of women over 80 in London during the years 1910, 1911 and 1912 as recorded in the Times newspaper. Table 2 shows the data and expected counts for $(\hat{\theta}_1, \hat{\theta}_2, \hat{p})$ = (1.257, 2.664, 0.360). Using ten classes X^2 = 1.29 with six degrees of freedom and χ^2 p-value 0.65. Also $\hat{V}_2^2$ = $(-0.077)^2$, $\hat{V}_3^2$ = $(-0.314)^2$ and $\hat{V}_4^2$ = $(-0.429)^2$, with bootstrap p-values 0.70, 0.46 and 0.55 respectively. Possibly due to different death rates in summer and winter, all tests indicate a good fit by a Poisson mixture. If a single Poisson is used to describe the data then X^2 = 27.01 with eight degrees of freedom and a χ^2 p-value is 0.001.

Table 2. Deaths per day of London women over 80 during 1910 to 1912

# deaths	0	1	2	3	4
Count	162	267	271	185	111
Mixture expected	161	271	262	191	114
Poisson expected	127	273	295	212	114

# deaths	5	6	7	8	9
Count	61	27	8	3	1
Mixture expected	58	25	9	3	1
Poisson expected	49	18	5	1	0

A plot of likelihood contours indicated the likelihood has a maximum at $(\hat{\theta}_1, \hat{\theta}_2)$ and that there are no other stationary points nearby. As $\hat{V}_1 \equiv 0$ we can give $\hat{p}$ in terms of $\hat{\theta}_1$ and $\hat{\theta}_2$ and so $\hat{p}$ does not need to be included in any likelihood contour plot.

References

[1] Everitt, B.S. and Hand, D.J.C. (1981). *Finite Mixture Distributions*. London: Chapman and Hall.

[2] Leroux, B.G. and Puterman, M.L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*, 48(2), 545-558.

[3] Rayner, J.C.W., Thas, O. and Best, D.J. (2009). *Smooth Tests of Goodness of Fit: Using R* (2nd ed.). Singapore: Wiley.

[4] Whitaker, L. (1914). On the Poisson law of small numbers. *Biometrika*, 10(1), 36-71.

Appendix A: Orthonormal Polynomials for a Poisson Mixture

Let μ be the mean and μ_t for t = 2, 3, ... the central moments, assumed to exist, of some distribution of interest. Then the first four orthonormal polynomials are, for x = 0, 1, 2, ...

$$g_0(x) = 1,\ g_1(x) = (x-\mu)/\sqrt{\mu_2},$$
$$g_2(x) = \{(x-\mu)^2 - \mu_3(x-\mu)/\mu_2 - \mu_2\}/\sqrt{d}$$
$$\text{and } g_3(x) = \{(x-\mu)^3 - a(x-\mu)^2 - b(x-\mu) - c\}/\sqrt{e}$$

where

$$d = \mu_4 - \mu_3^2/\mu_2 - \mu_2^2 \text{ and } e = \mu_6 - 2a\mu_5 + (a^2 - 2b)\mu_4 + 2(ab - c)\mu_3 + (b^2 + 2ac)\mu_2 + c^2$$

in which

$$a = (\mu_5 - \mu_3\mu_4/\mu_2 - \mu_2\mu_3)/d$$
$$b = (\mu_4^2/\mu_2 - \mu_2\mu_4 - \mu_3\mu_5/\mu_2 + \mu_3^2)/d$$
$$c = (2\mu_3\mu_4 - \mu_3^3/\mu_2 - \mu_2\mu_5)/d.$$

Again assuming they exist, for t = 2, 3, ... write $\mu'_{[t]}$ for the tth factorial moment. It now follows routinely that

$$\mu_2 = \mu'_{[2]} + \mu - \mu^2$$

$$\mu_3 = \mu'_{[3]} + 3\mu'_{[2]} + \mu - 3\mu(\mu'_{[2]} + \mu)\,\mu_2 + 2\mu^3$$

$$\mu_4 = \mu'_{[4]} + 6\mu'_{[3]} + 7\mu'_{[2]} + \mu - 4\mu(\mu'_{[3]} + 3\mu'_{[2]} + \mu) + 6\mu^2(\mu'_{[2]} + \mu) - 3\mu^4$$

$$\mu_5 = \mu'_{[5]} + 10\mu'_{[4]} + 25\mu'_{[3]} + 15\mu'_{[2]} + \mu - 5\mu(\mu'_{[4]} + 6\mu'_{[3]} + 7\mu'_{[2]} + \mu) + 10\mu^2(\mu'_{[3]} + 3\mu'_{[2]} + \mu) - 10\mu^3(\mu'_{[2]} + \mu) + 4\mu^5$$

$$\mu_6 = \mu'_{[6]} + 15\mu'_{[5]} + 65\mu'_{[4]} + 90\mu'_{[3]} + 31\mu'_{[2]} + \mu - 6\mu(\mu'_{[5]} + 10\mu'_{[4]} + 25\mu'_{[3]} + 15\mu'_{[2]} + \mu) + 15\mu^2(\mu'_{[4]} + 6\mu'_{[3]} + 7\mu'_{[2]} + \mu) - 20\mu^3(\mu'_{[3]} + 3\mu'_{[2]} + \mu) + 15\mu^4(\mu'_{[2]} + \mu) - 5\mu^6.$$

For a Poisson mixture the tth factorial moment is $\mu'_{[t]} = p\theta_1^t + (1-p)\theta_2^t$ so that, for example, $\mu = p\theta_1 + (1-p)\theta_2$. Using the ML estimators $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{p}$ and the above formulae for μ, ..., μ_6 we can calculate $\hat{V}_2$ and $\hat{V}_3$ where $\hat{V}_s = \sum_{i=1}^n g_s(x_i;\hat{\theta}_1,\hat{\theta}_2,\hat{p})/\sqrt{n}$, s = 2, 3.

Appendix B: Proof That $\hat{V}_1 \equiv 0$

Given $g_1(x)$ from Appendix A above, the first smooth component $\sum_{i=1}^n g_1(x_i;\hat{\theta}_1,\hat{\theta}_2,\hat{p})/\sqrt{n}$ is proportional to $\bar{X} - \hat{\mu}$, where $\hat{\mu} = \hat{p}\hat{\theta}_1 + (1-\hat{p})\hat{\theta}_2$ is the ML estimator of μ = E[X]. For notational convenience arguments involving θ_1, θ_2 and p are henceforth suppressed. To obtain the ML estimators of θ_1, θ_2 and p note that the likelihood is $L = \prod_{i=1}^n f^*(x_i;\theta_1,\theta_2,p)$. Taking logarithms and differentiating gives

$$\frac{\partial \log L}{\partial \theta_1} = \sum_{i=1}^n \{pf_1(x_i)[-1+\frac{x_i}{\theta_1}]\}/f^*(x_i),$$

$$\frac{\partial \log L}{\partial \theta_2} = \sum_{i=1}^n \{pf_2(x_i)[-1+\frac{x_i}{\theta_2}]\}/f^*(x_i)$$

and

$$\frac{\partial \log L}{\partial p} = \sum_{i=1}^n \{f_1(x_i) - f_2(x_i)/f^*(x_i).$$

From $\partial \log L/\partial\theta_r = 0$ for r = 1 and 2 we obtain

$$\hat{\theta}_r = \frac{\sum_{i=1}^n x_i f_r(x_i)/f^*(x_i)}{\sum_{i=1}^n f_r(x_i)/f^*(x_i)}$$

and from $\partial \log L/\partial p = 0$ we obtain

$$\sum_{i=1}^n f_1(x_i)/f^*(x_i) = \sum_{i=1}^n f_2(x_i)/f^*(x_i).$$

Using $f^*(x) = p\,f_1(x) + (1-p)\,f_2(x)$ and the equation immediately above shows that $\sum_{i=1}^n f_r(x_i)/f^*(x_i) = n$ for r = 1 and 2. It now follows that

$$\hat{\theta}_r = \sum_{i=1}^n x_i f_r(x_i)/\{nf^*(x_i)\} \text{ and}$$

$$\hat{\mu} = \hat{p}\hat{\theta}_1 + (1-\hat{p})\hat{\theta}_2 =$$

$$\hat{p}\sum_{i=1}^n x_i f_1(x_i)/f^*(x_i)/n +$$

$$(1-\hat{p})\sum_{i=1}^n x_i f_2(x_i)/f^*(x_i)/n =$$

$$\sum_{i=1}^n x_i\{\hat{p}f_1(x_i) + (1-\hat{p})f_2(x_i)\}/\{nf^*(x_i)\} =$$

$$\sum_{i=1}^n x_i/n = \bar{x}.$$

It thus follows that $\hat{V}_1 \equiv 0$.

Assessing Poisson and Logistic Regression Models Using Smooth Tests

Paul Rippon, J.C.W. Rayner

The University of Newcastle, Callaghan, NSW, 2308, AUSTRALIA
paul.rippon@newcastle.edu.au

Abstract

A smooth testing approach has been used to develop a test of the distributional assumption for generalized linear models. Application of this test to help assess Poisson and logistic regression models is discussed in this paper and power is compared to some common tests.

Key words: generalized linear models, goodness of fit, logistic regression, Poisson regression

1. Introduction

The concept of smooth testing originally proposed in [1] has been developed in [2] to provide goodness of fit tests for a wide range of distributions. In [3], these ideas have been applied to the generalized linear modelling framework, where the variables are no longer identically distributed, to derive a test of the distributional assumption. Section 2 describes the test, Section 3 comments on its application and Section 4 discusses the results of simulation studies examining the power of this test when applied to Poisson and logistic regression.

2. A Smooth Test of the Distributional Assumption in Generalized Linear Models

The generalized linear modelling structure comprises a linear combination of predictor variables related via a link function to the mean of the response distribution selected from the exponential family of distributions. In commonly used notation, independent response variables, $Y_1, \ldots, Y_n$, are distributed with density function

$$f(y_j;\theta_j) = \exp\left[\frac{y_j\theta_j - b(\theta_j)}{a(\phi_j)} + c(y_j,\phi_j)\right]$$

from an exponential family with canonical parameters θ_j to be estimated and dispersion parameters ϕ_j assumed to be known; a, b and c are known functions. Using $g(\cdot)$ to represent the link function:

$$g(\mu_j) = \eta_j = \boldsymbol{x}_j^T\boldsymbol{\beta} = x_{j1}\beta_1 + \ldots + x_{jp}\beta_p$$

where $\mu_j = E[Y_j] = b'(\theta_j)$ for $j = 1,\ldots,n$. To simplify subscripting, an explicit intercept term, β_0, is not shown. There is no loss of generality as β_1 can become an intercept term by setting all $x_{j1} = 1$.

To test the distributional assumption, the assumed response variable density, $f(y_j;\theta_j)$, is embedded within a more complex alternative density function

$$f_k(y_j;\boldsymbol{\tau},\theta_j) = C(\boldsymbol{\tau},\theta_j)\exp\left\{\sum_{i=1}^{k}\tau_i h_i(y_j;\theta_j)\right\} f(y_j;\theta_j).$$

This structure allows for 'smooth' departures from the assumed distribution controlled by the vector parameter, $\boldsymbol{\tau} = [\tau_1,\ldots,\tau_k]^T$ acting on the elements of the set, $\{h_i(y;\theta)\}$, of polynomials up to order k which are orthonormal on the assumed distribution. The normalizing constant, $C(\boldsymbol{\tau},\theta_j)$, simply ensures that $f_k(y_j;\boldsymbol{\tau},\theta_j)$ is correctly scaled to provide a valid probability density function.

When $\boldsymbol{\tau} = \mathbf{0}$, this smooth alternative collapses to the original response variable distribution. Thus a test of $H_0 : \boldsymbol{\tau} = \mathbf{0}$ against $H_A : \boldsymbol{\tau} \neq \mathbf{0}$ can reasonably be considered a test of the distributional assumption in a generalized linear model.

In [3], a score test statistic has been derived that can be expressed as a sum of squares of several contributing components:

$$\hat{S}_k = \frac{\hat{V}_1^2}{\hat{\omega}^2} + \hat{V}_2^2 + \ldots + \hat{V}_k^2$$

where

$$\hat{V}_i = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} h_i(y_j;\hat{\theta}_j).$$

The ith component involves the sum over the data of the ith order polynomial from the orthonormal sequence used in the construction of the smooth alternative distribution. The first component also contains a

term

$$\omega^2 = 1 - \frac{\mathbf{1}^T \boldsymbol{H} \mathbf{1}}{n}$$

which is related to the hat matrix, $\boldsymbol{H}$, obtained from the model estimation process.

Large values of $\hat{S}_k$ provide evidence against H_0. Asymptotically, the components $\hat{V}_1^2/\hat{\omega}^2$, $\hat{V}_2^2$, etc can each be expected to follow the $\chi^2_{(1)}$ distribution and $\hat{S}_k$ the $\chi^2_{(k)}$ distribution. In practice this has not proved a good enough approximation for common sample sizes and so a parametric bootstrap process is recommended to estimate p-values.

3. Applying the Smooth Test

In deriving this test of the distributional assumption, the linear predictor and the link function are assumed to be correctly specified. If this is not true then a large value of the test statistic may be caused by a mismatch between the data and these other components of the generalized linear model rather than an inappropriate response distribution. Similar issues arise with other tests that are used to assess generalized linear models. For example, the well-known deviance statistic is derived as a likelihood ratio test statistic comparing the fitted model with a saturated model having a linear predictor with as many parameters as there are covariate patterns. This provides the best possible fit to the observed data – assuming that the specified response distribution and link function are correct. If this is not true, then a large value of the deviance statistic may indicate a problem with the assumed distribution or link function rather than the linear predictor. Similarly, a model that 'fails' a goodness-of-link test may really have a problem with the assumed distribution or linear predictor and not the link function.

Can we ever truly diagnose the problem with a poorly fitting model? Clearly all such tests need to be carefully interpreted. There are many different ways that a model can be misspecified, some of which are very difficult to distinguish from each other. The smooth testing approach is not a panacea. In addition to providing a reliable test of the distributional assumption however, the individual components can be considered as test statistics in their own right. This can provide useful diagnostic information about the nature of any lack of fit detected.

4. Power Study

4.1. Logistic Regression

Figure 1 shows the results of a simulation study for logistic regression with a **misspecified linear predictor**. In this example, the fitted model was

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1$$

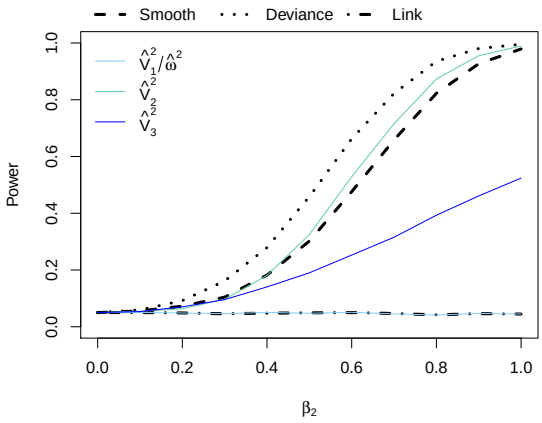


Figure 1: Power to detect a misspecified linear predictor in simulated logistic regression data.

but the true model used to simulate the data was

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

A fixed covariate pattern was used for each simulation with 25 groups corresponding to x_1 taking values $-1, -0.5, 0, 0.5, 1$ and x_2 taking values $-1.2, -0.7, -0.2, 0.3, 0.8$. There were $m = 30$ trials in each group. These two models coincide when $\beta_2 = 0$. The misspecification increases as β_2 increases (horizontal axis).

100000 simulations were conducted for $\beta_2 = 0$ to characterize the null distribution of each test statistic and 20000 simulations for each of the other β_2 values to characterize the alternative distributions. The $\alpha = 5\%$ critical value from the null distribution was used to define the rejection region and thus determine the probability of the null hypothesis being rejected (power to detect the misspecification) which is plotted on the vertical axis.

Three test statistics have been considered here: the deviance statistic, the smooth test statistic of order 3 and a link test statistic (see Appendix A). For all statistics used, the powers were based on simulated distributions and not on approximate sampling distributions. In this first example, the deviance performs best in detecting this particular kind of misspecification of the linear predictor. But the smooth test still performs reasonably well and the link test is essentially useless here. The performance of the $\hat{S}_k$ statistic is a compromise between the performance of the individual components which can also be considered separately. In this case: the first component is almost exactly matching the performance of the goodness of link test; the second component has good power and drives the performance of the overall test statistic and the third component is not particularly useful. The components correspond roughly to moments and so the second component is suggesting that the variance in the data is not

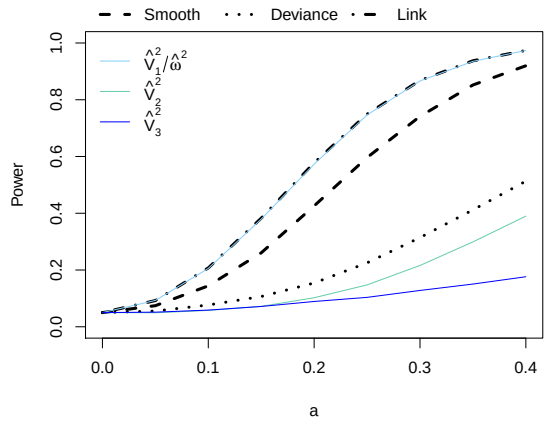


Figure 2: Power to detect a misspecified link function in simulated logistic regression data.

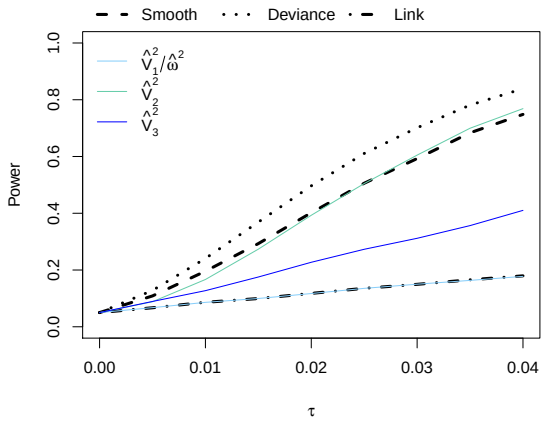


Figure 3: Power to detect a misspecified response distribution in simulated logistic regression data.

well modelled. This makes sense. A covariate is missing and so the stochastic part of the model is trying to cope with additional variation that should really have been explained by the linear predictor.

Figure 2 shows the results for a **misspecified link function** where the fitted model was

$$\pi(\eta) = \frac{e^{\eta}}{1+e^{\eta}} \qquad \log\left(\frac{\pi}{1-\pi}\right) = \eta = \beta_0 + \beta_1 x_1$$

but the data was simulated using a generalization of the logit link function (see Appendix B):

$$\pi(\eta) = \frac{e^{h(\eta;a)}}{1+e^{h(\eta;a)}}. \qquad (1)$$

The parameter a plotted along the horizontal axis controls the amount of misspecification with zero again representing no misspecification. Other simulation details are the same as in the first example.

Unsurprisingly, it is the goodness of link test that performs best here as this is the kind of problem it is designed to detect. However, the smooth test still performs well. Looking at the individual components, the first component is again matching the performance of the goodness of link test and is driving the performance of the overall test statistic in detecting this kind of misspecified model. The first component is correctly indicating that the problem is in how the mean of the data is being modelled. The second and third components aren't useful in this case.

Figure 3 shows the results for a **misspecified response distribution** where a binomial distribution is specified when fitting the model but the data was simulated using a beta-binomial distribution where the responses Y_j are $\mathrm{B}(m_j, \pi_j^*)$ for π_j^* independently distributed as beta random variables on $(0,1)$ with $E[\pi_j^*] = \pi_j$ and $\mathrm{Var}(\pi_j^*) = \tau\pi_j(1-\pi_j)$.

Again the parameter plotted along the horizontal axis, τ in this case, controls the amount of misspecification with zero representing no misspecification. The

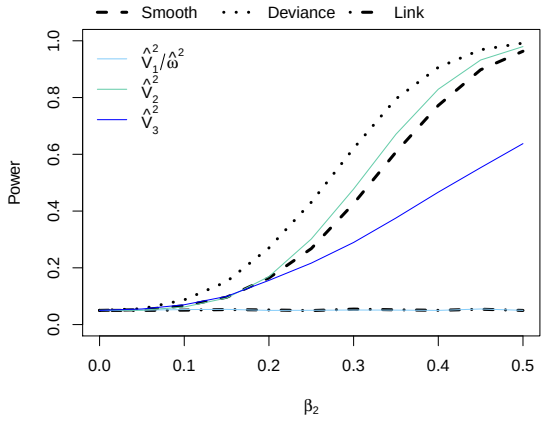


Figure 4: Power to detect a misspecified linear predictor distribution in simulated Poisson regression data.

deviance test performs best in detecting this particular type of misspecification, with the smooth test again performing reasonably well and the goodness of link test poorly. The story with the components is again similar with the first component matching the performance of the goodness of link test and the second component indicating correctly that the variance is not being modelled correctly in this example.

4.2. Poisson Regression

In Figure 4, the simulation scenario is the same as for Figure 1 except that the linear predictor is set to $\log\mu$ where $Y_j \sim \mathrm{P}(\mu_j)$. The performance of the smooth test statistic and components in detecting this type of misspecified linear predictor in Poisson regression can be seen to be very similar to that already discussed for logistic regression.

In Figure 5, a Poisson distribution is specified when fitting the model but the data was simulated using a negative binomial distribution with $\log\mu_j = \eta_j$ and variance $\mu_j + \tau\mu_j^2$. As in the similar logistic regression example, the deviance is more powerful in detecting the misspecification but the smooth test performs rea-

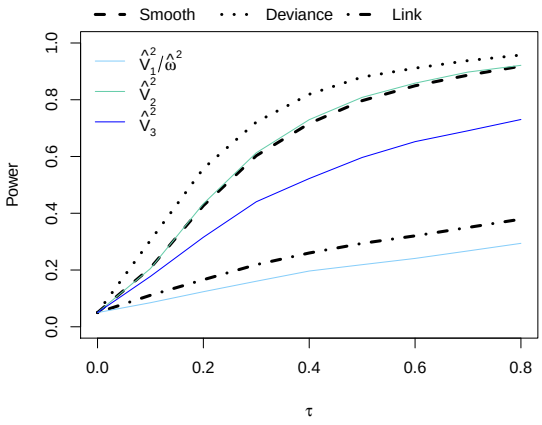


Figure 5: Power to detect a misspecified response distribution in simulated Poisson regression data.

sonably and the second component correctly indicates that the problem is in how the variance of the data is being modelled.

5. Conclusions

A smooth test for assessing the distributional assumption in generalized linear models has been derived in [3] and applied here to Poisson and logistic regression models fitted to simulated data. While not always the most powerful test, it appears to perform quite well in detecting lack of fit even when the misspecification is in the link function or the linear predictor rather than the response distribution. Interpretation of the components provides additional diagnostic information.

A. Goodness of Link Test

There are a number of tests described in the literature for testing the adequacy of the link function in a generalized linear model. Many of these are specific to a particular link function. The goodness of link test used in this paper is more generally applicable and is equivalent to the `linktest` function provided in STATA [4].

The $\hat{\eta} = \boldsymbol{X}\hat{\beta}$ term from the fitted model and a $\hat{\eta}^2$ term are used as the predictors of the original response variables in a new model. The $\hat{\eta}$ term contains all the explanatory information of the original model. If there is a misspecified link the relationship between $\hat{\eta}$ and $g(\overline{y})$ will be non-linear and the $\hat{\eta}^2$ term is likely to be significant. The difference in deviance between these two models has been used as the link test statistic in this study.

B. Generalized Logit Function

Expressed as an inverse link function, a generalization of the logit function is described by [5] in the same form as Eq. (1) but using a function $h(\eta; \alpha_1, \alpha_2)$ where the two shape parameters, α_1 and α_2, separately control the left and right tails. $\alpha_1 = \alpha_2$ gives a symmetric probability curve $\pi(\eta)$ with the logistic model as the special case $\alpha_1 = \alpha_2 = 0$. The function $h(\eta; a)$ used in Eq. 1 corresponds to $a = -\alpha_1 = \alpha_2$. This gives an asymmetric probability curve that according to [5] corresponds to a Box-Cox power transform.

References

[1] J. Neyman. Smooth tests for goodness of fit. *Skandinavisk Aktuarietidskrift*, 20:149–199, 1937.

[2] J. C. W. Rayner, O. Thas, and D. J. Best. *Smooth tests of goodness of fit: Using R*. Oxford University Press, 2nd edition, 2009.

[3] Paul Rippon. Application of smooth tests of goodness of fit to generalized linear models. *Unpublished PhD thesis.*, 2011.

[4] StataCorp. *Stata Base Reference Manual, Release 9*. Stata press, 2005.

[5] Therese A. Stukel. Generalized logistic models. *Journal of the American Statistical Association*, 83(402):426–431, 1988.

Bootstrap confidence intervals for Mean Average Precision

Laurence A. F. Park

School of Computing and Mathematics, University of Western Sydney, Australia
lapark@scm.uws.edu.au

Abstract

Due to the unconstrained nature of language, search engines (such as the Google search engine) are developed and compared by obtaining a document set, a sample set of queries and the associated relevance judgements for the queries on the document set. The de facto standard function used to measure the accuracy of each search engine on the test data is called mean Average Precision (AP). It is common practice to report mean AP scores and the results of paired significance tests against baseline search engines, but the confidence in the mean AP score is never reported. In this article, we investigate the utility of bootstrap confidence intervals for mean AP. We find that our Standardised logit bootstrap confidence intervals are very accurate for all levels of confidence examined and sample sizes.

Key words: bootstrap, confidence interval, average precision

1. Introduction

Text based search engines (such as the Google search engine), also known as text retrieval systems, have been developed for the past fifty years. During that time, many systems have been constructed based on various models. Each retrieval system is a function that takes a set of key words (the query) and returns a vector of relevance judgements, where each relevance judgement is the predicted relevance of an associated document in the systems database to the query. Rather than providing the complete list of relevance judgements to the user, the search system usually returns the ten documents with greatest associated relevance judgements (in order) to the user and provides the remaining documents if requested.

To evaluate the accuracy of a retrieval system, a sample set of queries and their associated true relevance judgements (the set of correct relevance scores for each document, for each query) must be obtained. For each query, the system computed relevance judgements and true relevance judgements are compared using an evaluation function. The most widely used retrieval system evaluation function is Average Precision (AP) [1]. AP is a function of both precision (the proportion of correct documents in the retrieved set) and recall (the proportion of correct documents retrieved). Each AP value falls within the range $[0, 1]$, 0 meaning the system has not found any relevant documents, and 1 meaning all documents predicted as relevant are relevant and all predicted as irrelevant are irrelevant.

To evaluate a system, we should obtain many queries and their associated true relevance judgements to construct the system AP distribution. Unfortunately, it is costly (in terms of time) to obtain the set of true relevance judgements for a single query, since each document must be manually judged to build the list [2], and it is common for retrieval systems to have over one million documents in their database. Therefore retrieval experiments are performed using a small sample of queries and the sample mean is reported along with paired significance test results with baseline systems.

Using this experimental method, a reader of a publication is able to identify which system has performed best in the experiment, but we are unable to compare systems across publications from other experiments unless we obtain the systems and run the experiments ourselves. To compare systems across publications, the confidence interval of the mean AP should be reported. A recent study showed that accurate confidence intervals can be produced for mean AP by fitting a t distribution to the samples, as long as the queries used were standardised using five other systems and that all authors used the same standardising systems [3]. Since there are no defined set of "standard" systems, it would be unlikely that experimental results from different authors would use the same standardising systems, and hence obtain confidence intervals that are not comparable.

In this article we will investigate the accuracy of bootstrap confidence intervals on mean Average Precision. We examine the accuracy of Percentile and Accelerated bootstrap, and we introduce the Studentised logit bootstrap, based on the analysis of the system distributions. The article will proceed as follows: Sec-

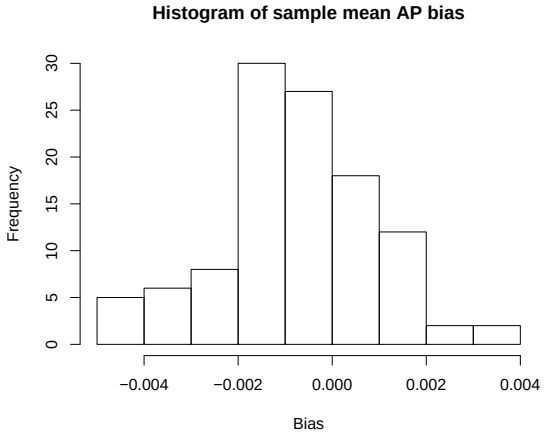


Figure 1: The distribution of sample mean AP bias. The sample mean AP bias from each system AP distribution was measured, using a sample size of 5 and the distribution of the bias across all systems is shown above.

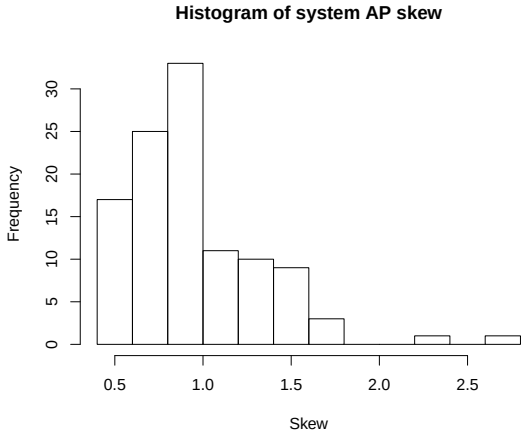


Figure 2: The distribution of system AP skewness. The skewness from each system AP distribution was measured and the distribution of the skewness across all systems is shown above.

tion 2 describes the experimental environment, section 3 examines set of system AP distributions, and Section 4 provides details of the experiments and results.

2. Experimental Environment

To conduct our experiments, we will use the set of 110 systems that participated in the TREC (`http://trec.nist.gov`, Text REtrieval Conference) 2004 Robust track. The Robust track consists of 249 queries and 528, 155 documents. We have obtained the AP of each query on each system. We will approximate the population AP distribution with the set of 249 AP values for each system. Our experiments will involve taking $1,000$ random samples of $n = 5, 10$ and 20 AP values without replacement for each system, computing the confidence interval for the mean AP and evaluating the coverage probability of the confidence interval. The bootstrap distribution is computed by taking $1,000$ random samples of size n, with replacement, from the AP sample. For each experiment, we will examine the confidence intervals at $1 - \alpha = 0.95$ to 0.50 in steps of 0.05, where $1 - \alpha$ is the proposed coverage probability.

3. System AP distribution

Before we proceed, we will examine the bias and skewness of each system AP distribution. Both bias and skewness are known to affect the accuracy of confidence intervals when computed using the bootstrap [6, 7, 8]. Bias is computed as the expected difference between the sample mean and the population mean. Using $1,000$ samples of size of $n = 5$ queries, we computed the bias for each system AP and provided the distribution in Figure 1. Given that AP ranges from 0 to 1, we can see that the bias is small and unlikely to affect our experiments.

Skewness is a measure of the asymmetry of the distribution, where a symmetric distribution has no skewness, and a asymmetric distribution can be positively or negatively skewed. We computed the skewness for each system AP population distribution and provided the distribution in Figure 2. The histogram shows that all systems are positively skewed, meaning that lower AP values are more likely than higher AP values.

To examine the skewness further, we have provided the histograms of the systems with the least, median, and greatest skewness in Figure 3. We can see that none of the system AP distributions are symmetric. The least skewed system is more likely to provide greater AP values than the other two. We can also see that the most skewed system has obtained AP values between 0 and 0.1 for most of the 249 queries, making it a poor system.

From this analysis, we have found that there is little sampling bias, but there is high skewness in each system distribution.

4. Experiments

In this section we examine the accuracy of Percentile and Accelerated bootstrap confidence intervals on our experimental environment. We also derive the novel Studentised logit bootstrap from our analysis of the system distributions.

4.1. Percentile Bootstrap

To begin our experiments, we will compute the Percentile bootstrap confidence interval of the mean AP. The percentile bootstrap confidence interval is computed as follows:

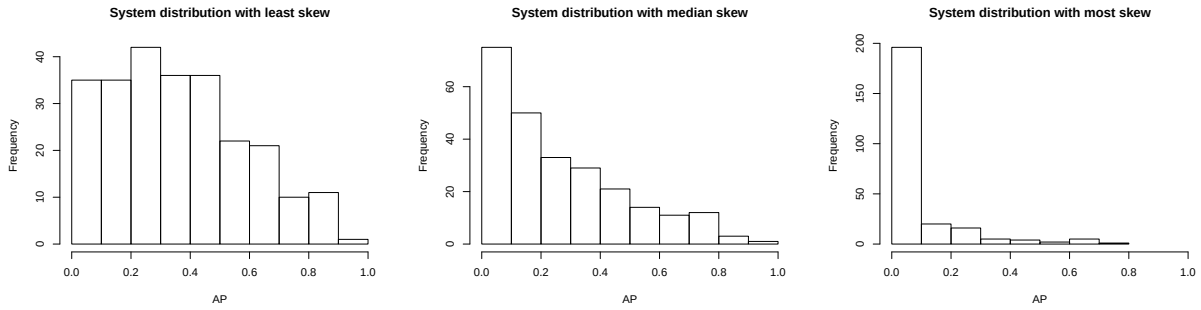


Figure 3: The AP distribution from three systems where the left-most, central and right-most distributions are associated to the systems with the least, median and most AP skewness.

$1-\alpha$	Coverage Probability		
	$n=5$	$n=10$	$n=20$
0.95	0.1814	0.1075	0.0701
0.90	0.2279	0.1576	0.1191
0.85	0.2816	0.2045	0.1672
0.80	0.3295	0.2514	0.2151
0.75	0.3658	0.2981	0.2628
0.70	0.4015	0.3435	0.3098
0.65	0.4380	0.3901	0.3587
0.60	0.4789	0.4357	0.4054
0.55	0.5166	0.4811	0.4552
0.50	0.5573	0.5269	0.5035

Table 1: Coverage probability when computing $(1-\alpha)\times 100\%$ confidence intervals of mean AP from n AP samples using the Percentile Bootstrap method.

$1-\alpha$	Coverage Probability		
	$n=5$	$n=10$	$n=20$
0.95	0.1704	0.0932	0.0592
0.90	0.2186	0.1442	0.1067
0.85	0.2613	0.1927	0.1552
0.80	0.3085	0.2392	0.2030
0.75	0.3515	0.2848	0.2500
0.70	0.3903	0.3310	0.2995
0.65	0.4278	0.3777	0.3477
0.60	0.4664	0.4242	0.3958
0.55	0.5050	0.4702	0.4445
0.50	0.5450	0.5172	0.4937

Table 2: Coverage probability when computing $(1-\alpha)\times 100\%$ confidence intervals of mean AP from n AP samples using the Accelerated Bootstrap method.

1. Compute the bootstrap distribution of the sample mean AP.
2. Use the $\alpha/2$ and $1-\alpha/2$ quantiles as the $(1-\alpha)\times 100\%$ confidence interval boundary.

It is known that the Percentile bootstrap does not provide the correct coverage when the population is skewed [4, 5]. Therefore, we will measure the accuracy of the confidence intervals and use them as a baseline. The results are provided in Table 1.

We can see from Table 1 that there is a large difference between $1-\alpha$ and the coverage probability for $n=5$ and 10. For $n=20$, we can see that the coverage probability is similar to the associated value of $1-\alpha$. This is to be expected since the distribution of the sample mean will be approximately Normal for large values of n.

4.2. Bias Corrected Accelerated Bootstrap

The bias corrected accelerated bootstrap confidence interval (BC_a) [6, 7, 8] was developed to provide good confidence intervals for a sample taking into account the bias and skewness.

The BC_a bootstrap confidence interval is intended to be a general purpose method and includes many steps to compute the confidence interval bounds, therefore we refer the reader to [6] for further information.

In this experiment we use the nonparametric form of the BC_a bootstrap (where the bias correction and acceleration statistic are derived from the sample). The results are shown in Table 2.

From Table 2, we can see that the difference between the coverage probability and $1-\alpha$ is slightly smaller when compared to the Percentile bootstrap confidence intervals, but the confidence intervals are still inaccurate for most values of $1-\alpha$ when $n=5$, and large values of $1-\alpha$ when $n=10$, but accurate for $n=20$.

4.3. Studentised logit Bootstrap

It is a concern that the AP values are constrained to the domain $[0,1]$, while this contraint is not explicitly provided when computing the confidence interval. To map the AP samples to the real domain, we can use the logit function:

$$\text{logit}(x) = \log_e\left(\frac{x}{1-x}\right)$$

The logit transform takes data from the $[0, 1]$ domain to the $(-\infty, \infty)$ domain. By observing the AP distributions in Figure 3, we can see that applying the logit transform may reduce the skewness and and provide a more Normal distribution.

Unfortunately, we can't apply the logit transform to the samples since there may be scores of 0 or 1 which are transformed to $-\infty$ and ∞ respectively. However, we are able to transform the sample mean AP. The sample mean can only be 0 or 1 if all of the samples are 0 or 1 respectively. If this is the case, then we are unable to compute a confidence interval due to the lack of variation in the sample.

To reduce the skewness, we will compute the sample mean AP bootstrap distribution and apply the logit transformation to the bootstrap distribution. Note that if a sample contains a 0 or 1, there is a chance that the a bootstrap sample mean will be 0 or 1 respectively. In this case, we remove the associated bootstrap sample. The percentile bootstrap is invariant to monotone transformations, therefore computing the percentiles gives us no benefit over the percentile bootstrap baseline.

Assuming that the skewness has been removed, we compute the mean and standard deviation of the transformed bootstrap distribution and obtain the confidence interval boundary using the t distribution.

The Studentised logit Bootstrap is computed as follows:

1. Compute the bootstrap distribution of the sample mean.
2. Reduce the distribution skewness by applying the logit transformation.
3. Obtain the maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}$ of the Normal distribution parameters μ and σ.
4. Compute the mean AP confidence interval boundary using $\hat{\mu} \pm t_{\alpha/2,n-1}\hat{\sigma}$
5. Apply the inverse logit function to the boundary to convert it back to the AP domain.

The accuracy of the confidence intervals is shown in Table 3.

We can see from Table 3 that the Coverage probability of the confidence intervals produced using Studentised logit Bootstrap is very close to the provided $1 - \alpha$ for all values of n. We can see the difference grows as $1 - \alpha$ decreases for $n = 5$, but it is most accurate for small $1 - \alpha$ (being the usual confidence range).

$1-\alpha$	Coverage Probability		
	$n = 5$	$n = 10$	$n = 20$
0.95	0.0546	0.0541	0.0466
0.90	0.1097	0.1075	0.0934
0.85	0.1646	0.1592	0.1420
0.80	0.2190	0.2101	0.1910
0.75	0.2724	0.2606	0.2406
0.70	0.3244	0.3103	0.2915
0.65	0.3742	0.3601	0.3424
0.60	0.4232	0.4089	0.3924
0.55	0.4730	0.4580	0.4431
0.50	0.5235	0.5074	0.4937

Table 3: Coverage probability when computing $(1-\alpha)\times 100\%$ confidence intervals of mean AP from n AP samples using the Studentised logit Bootstrap method.

5. Conclusion

Empirical evaluation of the accuracy of document retrieval systems is performed using a sample set of queries. The sample is usually small due to the work involved in providing manual relevance judgements for all documents for each query.

It is common place for document retrieval system evaluation to report the sample mean Average Precision (AP), but the fact that we are only working with a sample set of queries is usually ignored, making results across publications incomparable.

In this article we examined the accuracy of bootstrap confidence intervals for mean AP. We found that Percentile and Accelerated bootstrap confidence intervals had poor coverage for large $1 - \alpha$ and small number of samples (5 queries). We also found that our Standardised logit bootstrap confidence intervals were very accurate for all levels of confidence examined and sample sizes. We believe the accuracy of the method comes from the logit transform removing most of the skewness from the bootstrap distribution.

References

[1] C. Buckley, E. M. Voorhees, Evaluating evaluation measure stability, in: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '00, ACM, New York, NY, USA, 2000, pp. 33–40. doi:10.1145/345508.345543.

[2] S. Keenan, A. F. Smeaton, G. Keogh, The effect of pool depth on system evaluation in TREC, Journal of the American Society for Information Science and Technology 52 (7) (2001) 570–574. doi:10.1002/asi.1096.

[3] L. A. F. Park, Confidence intervals for information retrieval evaluation, in: A. Turpin, F. Scholer, A. Trotman (Eds.), Proceedings of the Fifteenth Australasian Document Computing Symposium (to appear), 2010.

[4] B. Efron, R. J. Tibshirani, An Introduction to the Bootstrap, CRC Monographs on Statistics and Applied Probability, Chapman and Hall, 1994.

[5] P. M. Dixon, Bootstrap resampling, in: Encyclopedia of Environmetrics, John Wiley and Sons, Ltd, 2006. doi:10.1002/9780470057339.vab028.

[6] B. Efron, Better bootstrap confidence intervals, Journal of the American Statistical Association 82 (397) (1987) 171–185. doi:10.2307/2289144.

[7] T. J. DiCiccio, B. Efron, Bootstrap confidence intervals, Statistical Science 11 (3) (1996) 189–228.

[8] F. T. Burbrink, R. A. Pyron, The Taming of the Skew: Estimating Proper Confidence Intervals for Divergence Dates, Systematic Biology 57 (2) (2008) 317–328. doi:10.1080/10635150802040605.

Wilson confidence intervals for the two-sample log-odds-ratio in stratified 2×2 contingency tables

Thomas Suesse

University of Wollongong, Australia
tsuesse@uow.edu.au

Bruce Brown

UNSW, Australia

Abstract

Large-sample Wilson-type confidence intervals (CI) are derived for a parameter of interest in many clinical trials situations: the log-odds-ratio, in a two sample experiment comparing binomial success proportions, say between cases and controls. The methods cover several scenarios: (i) results embedded in a single 2×2 contingency table, (ii) a series of K 2×2 tables with common parameter, or (iii) K tables, where the parameter may change across tables under the influence of a covariate. The calculations of the Wilson CI require only simple numerical assistance, and for example are easily carried out using Excel. The main competitor, the exact CI has two disadvantages: It requires burdensome search algorithms for the multi-table case and results in strong over-coverage associated with long confidence intervals. All the application cases are illustrated through a well-known example. A simulation study then investigates how the Wilson CI performs among several competing methods. The Wilson interval is shortest, except for very large odds ratios, while maintaining coverage similar to Wald-type intervals. An alternative to the Wald CI is the Agresti-Coull CI, calculated from Wilson and Wald CI, which has same length as Wald CI but improved coverage.

Key words: Odds Ratio, Wilson Confidence Interval, Conditional Maximum Likelihood

Statistical Consulting at the CSSM

David Steel

University of Wollongong, Australia
dsteel@uow.edu.au

Abstract

The CSSM has an active portfolio of statistical consulting involving a range of organisations. This talk will cover how CSSM works and some of the interesting projects and solutions that have been developed.

Key words: consulting, projects, solutions

A summary of methods used in statistical consulting at the University of Wollongong in 2010

Marijka Batterham

University of Wollongong, Australia
marijka@uow.edu.au

Abstract

The statistical consulting service at the University of Wollongong provides limited free advice on the design and analysis of quantitative research to staff and postgraduate research students. While most faculties have used the service, the health and science related disciplines are the key clients. Most clients required advice on general methods (ANOVA (20%), t tests (15%), regression/correlation (17%), 2/proportions (10%). Several clients were able to embrace and utilise more advanced methods (such as methods for clustered designs including mixed models and GEE (2% of consults) or mixed models for repeated or missing data (8%)). Other methods used included cluster analysis, PCA, time series, reliability/agreement, non parametric methods, structural equation modelling and survival analysis. Five percent of consults were related to basic descriptive statistics and data management. Seven percent of consults involved assistance with sample size or power calculations. While some clients were interested in discussing Bayesian methods only one client adopted a fully Bayesian approach for analysis.

Key words: statistical consulting, statistical methods, data analysis

Advice for the potential statistical consultant

K.G. Russell

University of Wollongong, NSW 2522, AUSTRALIA
Charles Sturt University, Locked Bag 588, Wagga Wagga NSW 2678, AUSTRALIA

Abstract

Many of those who have never practised statistical consulting seem to view it from one of two extremes: a trivial exercise that any statistician could do, or an extremely difficult task occasioning extreme anxiety. Of course, the truth is that it lies somewhere in between. I will consider various aspects of statistical consulting, including the characteristics of a successful consultation, the necessary knowledge, skills and attributes of the consultant, things to do and not to do, and the pleasures and frustrations that one can gain from consulting.

Key words: Consulting, knowledge, skills

1. Introduction

The provision of statistical consulting is almost as old as statistics itself. Russell [6] pointed out that a very early consultation occurred when the Chevalier de Méré sought statistical (probabilistic) advice from Fermat and Pascal on the 'problem of points'. Some private companies, research institutions and government departments have employed statisticians for many years. In Universities, the provision of formal statistical consulting has existed for some time. There was a Consulting Service at Victoria University of Wellington, New Zealand prior to 1979, and the University of Wollongong has had a Statistical Consulting Service since 1989. But, especially in Universities, there has been ambivalence about how valuable the consultant is. Advertisements for University consultants frequently offer the position at Lecturer level, or even at Research Fellow level, suggesting that the position is seen as a low-level role requiring not too much statistical maturity. Nothing could be further from the truth. While any statistical knowledge and experience will be useful, a good consultant requires many skills, of which good statistical knowledge and experience are just two. Some good academic statisticians are not suited to consulting: an ability to derive results in an idealized situation is not sufficient when that situation does not prevail.

If you discuss with imminent graduates in Statistics what they will do in the future, you will frequently encounter trepidation at having to deal with 'real' problems. When there is no hint in the description of a problem as to what assumptions can be made or what type of analysis is appropriate, and when one cannot even be confident that the description is correct, a moderate amount of resourcefulness and ingenuity is required. This is the other extreme of the view of statistical consulting.

Email address: kerussell@csu.edu.au (K.G. Russell)

2. A Successful Consultation

For a successful consultation to occur, both parties must respect the other, and the knowledge, ideas and skills that they possess. They must both be willing to listen and learn, to make active contributions to the discourse, and to accept that neither side has a monopoly on being right.

A successful consultation will leave both the client and the consultant feeling that progress has been made in solving the problem. This is not necessarily the same as both parties feeling happy at the end of the session. There may be frustration that more progress was not made, or that some misunderstanding still remains, but we can't expect everything in real life to be totally successful. The consultation will have been a failure if, at its end, the client feels that the consultant was not listening, or was not interested in trying to understand the problem, or did not know enough Statistics to solve the problem. (If the last situation is correct, the best that you can do is to express regret that you cannot help the client and redirect him to someone else.) Of course, it may be the consultant who thinks that the consultation has been a failure, because the client did not provide enough information, or wouldn't answer questions, or . . . ; however, in this relationship, it is the consultant who is the professional (who will be paid if

this is a financial transaction), and it is her responsibility to manage the discussion in a way that leads to a satisfactory outcome.

3. Statistical Knowledge

A consultant can never have too much statistical knowledge. This is not to say that it will all be needed, but one can never foretell what a client may ask, and a wide knowledge reduces the risk that one has to devise a new technique. Very few consultations give completely standard problems, and some improvisation will probably be necessary. However, the client, and journal editors and referees, will always be happier if the techniques you suggest come with suitable references. Consult Boen & Zahn [1], or Cabrera & McDougall [2], for lists of the techniques that should be at the forefront of your statistical toolkit.

However, the mere fact that you know a sophisticated method for analysing some data does not justify using it. It is very good practice to use only as much sophistication as is needed. If the client can't understand what you're doing, or why you're doing it, and the referees have never heard of the technique, it is likely that the analysis will be questioned. Indulging in the latest technique you have learnt about (or developed) when something simpler will do, is an unacceptable indulgence for a statistical consultant. (Save such things for your conference or journal publications!)

I believe that you can never know too much experimental design. Comparisons of the means of two, or k, independent populations, and the analysis of two-factor factorial experiments, may well be common, but many clients will present you with data from designs that contain nested factors, factors that are not completely crossed with one another, 'replicates' that are not true replicates, ... Your ability to modify the analysis produced by a statistical package to reflect accurately what actually happened in the data collection will probably depend on how well you understand the principles of experimental design. You need to be able to determine what the experimental units really are, as opposed to what the client thinks they are. Similarly, if your clients tend to conduct elaborate surveys, a strong knowledge of survey methodology will stand you in good stead. Should you be fortunate to have a client consult you *before* the data are collected, a thorough knowledge of the underlying principles of good design will stand you in good stead. Possibly the most common question that clients ask is 'How large a sample do I need?' Without a good design background, you can't answer this question effectively. You need to understand what a 'control' treatment is, and when and why it is necessary.

4. Computing

It is essential that you be able to use several statistical packages to perform statistical analyses. I do not believe that any one package is the 'best' for every analysis, and you will benefit if you can choose the one that is most appropriate for your particular task. Some packages are good for producing graphics, some shine in particular areas of analysis (for example, I consider GenStat to be the leader in the analysis of designed experiments). You can save time, and frustration, if you can select the most appropriate package for a particular task, rather than be forced to use just one.

I believe that it is also necessary to be able to write computer programs, for use in those situations where a standard analysis will not work. For example, the client may have brought you some data that cannot be transformed to Normality, and cannot be analyzed by a Generalized Linear Model routine. You are unlikely to be able to find an appropriate randomization test in a conventional statistical package, so you will need to write your own program. This may be done in Fortran (perhaps by modifying a paper in Edgington & Onghena [4]) or C++, or you can write a program using R or another malleable statistical package. Without the ability to do this, your capacity to deal with nonstandard statistical analyses will be limited.

5. Interpersonal skills

Perhaps even more important than technical skills are the interpersonal skills that will ensure that a consultation with a client goes well. Without these, the client may be actively resisting your attempts to solve the problem, and you may not even know what the problem really is.

5.1. Communication

An indication of how important communication skills are to statistical consulting is that there exists at least one book on consulting that considers only communication skills. See Der [3], which contains absolutely no discussion of the types of statistical designs, or analyses, that you might need. When you first meet a client, it is absolutely vital that you establish rapport with that person. Boen & Zahn [1] list various categories of clients, some of whom do not have autonomy, or lack confidence in their ability — or your ability. It never hurts to ask about the client's statistical and non-statistical background. Be clear that you don't expect the client to be an expert statistician (and mean this!).

Once (partial) rapport has been established, you will probably need to establish some 'ground rules' for the consultation. These will probably vary according to whether you work in a commercial environment (fees will need to be discussed) or a noncommercial one (are there limits on the time that your client can ask

of you?). These rules established, you can proceed towards finding out what is the problem with which the client requires assistance. Get the client to explain the problem in his own words, take lots of notes (you *won't* remember it all afterwards without notes), ask lots of questions (preferably without interrupting the client's flow of thought), paraphrase what you think the client is saying in as many ways as you can, and don't volunteer a 'solution' until you are confident that you and the client both understand the situation. Never place unquestioning faith in what the client tells you about the problem, especially when they start to use statistical jargon; always check. (This presents a significant difficulty for beginners, who have to learn to discard their assumption that the 'question' is always correct.)

I thought that I was good at checking that the client and I are 'on the same wavelength', but Russell [6] gives an example where it was just a chance remark by the client that made me realise that the labels we were applying to the units in the experiment did not agree. With another client, I soon learnt that every reference to a 'multivariate analysis' meant multiple regression. Many clients talk about examining correlations when they really want a 'linear regression'. Questions such as 'What do you hope to achieve with this analysis?' may reveal that the client does not mean what he has said.

A client may do a statistical analysis instead of you. You need to know how the data should be entered into the computer package, and a thorough description of how this is done should be given to the client. You should check that this has been understood. I often ask a client to send me a subset of the data (perhaps after the results from just a few experimental units have been entered), so that I can check that this has been done correctly.

5.2. Interest

Most clients are highly intelligent people; they are just not experts in Statistics. They can quickly tell if a consultant is not interested in a problem, or doesn't understand it, or is trying to get through the consultation as fast as possible. This usually guarantees a consultation that does not succeed. The consultant should try to learn something from every client, be curious about the background to the problem, and want to know why the client is doing the research. Such interest shines through, and helps build rapport with the client.

5.3. Commitment

Some academic statisticians have told me that they do consulting, but further discussion reveals that this means that they chat to someone for a few minutes and then turn that person away if the problem is not directly of interest to their research. To my mind, this is not consulting. A real consultant doesn't have the liberty to turn away a client unless the problem is beyond his or her technical expertise, there is no time available to attempt the problem, or an ethical conflict arises. So true statistical consulting requires a commitment to take on all problems that come, except for the reasons in the preceding sentence.

5.4. Alertness and Common Sense

It is essential that a consultant actively *listens* to what the client is saying. Only by doing this will remarks that seem inconsistent (and which often carry the greatest information) be detected. This is part of being alert. It is also necessary to maintain eye contact with the client or, at least, to watch his or her face. This will enable you to detect when a remark you have made has not been understood, or has caused confusion.

I like to think of Statistics as common sense in the presence of variability. You must have common sense to be able to detect what is achievable when told of the client's aims and ambitions. I have frequently been told 'Wow, that's a great idea!' and have thought to myself 'But it's just common sense.'

6. Dos and Don'ts

Much of this is common sense, or has been covered previously.

Do treat the client with respect. Listen carefully to what he says, clarify anything that you don't understand, and ask questions to check your understanding. Whether providing an experimental design or an analysis of data, offer something that is no more complicated than it needs to be. Explain the assumptions that you are making, check that these seem reasonable to the client, and be sure that you provide an explanation of the conclusions that can be drawn from the data. Check that the client understands these conclusions, by getting him to paraphrase them for you. Offer to read the relevant section(s) of the report or paper that the client is writing, to check that it has been worded satisfactorily. (Statistical jargon can be dangerous for a nonstatistician.) If you feel that your contribution to a paper warrants joint authorship, discuss this with the client *early* in the process. You do a disservice to yourself and the profession if you allow your work not to be acknowledged. Regrettably, some clients will resist this suggestion, even when the paper could not have come into existence without your contribution. But many clients are reasonable, and will accede to a reasonable request.

Do maintain your integrity and ethical standards. Don't accede to requests to perform data analyses that you know are incorrect, or to ignore some factors that are likely to affect the results. If necessary, decline to continue the relationship with the client.

Don't expect more of the client than you expect of yourself. Recognise that he may be under pressure from elsewhere. Don't use excessive jargon in your

speech, and do seek to clarify any jargon that is presented to you.

7. Pleasures and Frustrations

It can be very satisfying to help a client perform an investigation that comes to a successful end, even if perhaps that end is not what the client had hoped for. The client may have gained knowledge that will let him do better on the next investigation. At times, you may see that the results that have been discovered are of benefit to mankind, the Earth, or a subsection thereof. You may meet some wonderfully intelligent or pleasant people. I am proud to have made some good friends from my clients.

It would be naive to suggest that consulting does not bring frustrations. Leaving aside those frustrations that arise because a problem is proving difficult to solve, you will encounter clients, or their problems, that seem uninspiring, there will be personal interactions that are not as satisfactory as you would have hoped, and there may even be occasions when you feel that your time is being wasted. (Perhaps these latter occasions are less common if you are paid for your consultations.)

8. Conclusions

A brief paper such as this cannot do justice to a very broad topic. The references are all worth reading, and offer different perspectives on the many attributes required of a good statistical consultant. Statistical consulting is not a trivial exercise: it requires personal and intellectual skills not always recognised by those who have not practised it. But nor is it impossibly difficult. A willingness to learn, to be flexible in one's thinking, and to believe that good statistical practice will benefit the world, are good attributes to start with.

References

[1] Boen, J.R. & Zahn, D.A. (1982). *The Human Side of Statistical Consulting.* Belmont, CA: Wadsworth.

[2] Cabrera, J. & McDougall, A. (2002). *Statistical Consulting.* New York: Springer-Verlag.

[3] Der, J. (2000). *Statistical Consulting: a guide to effective communication.* Pacific Grove, CA: Brooks/Cole.

[4] Edgington, E.S. & Onghena, P. (2007). *Randomization Tests* 4th edn. Boca Raton, FL: Champman & Hall/CRC.

[5] Hand, D.J. & Everitt, B.S. (1987). *The Statistical Consultant In Action.* Cambridge: Cambridge University Press.

[6] Russell, K.G. (2001). The teaching of statistical consulting. *J. Appl. Prob.* **38A**, 20–26.

Statistical Consulting under ASEARC

Kim Colyvas

University of Newcastle, Australia
kim.colyvas@newcastle.edu.au

Trevor Moffiet

University of Newcastle, Australia
trevor.moffiet@newcastle.edu.au

Abstract

It was intended that the formation of ASEARC would create coordinated programs in undergraduate and postgraduate courses, research, research training, and consulting. This talk focuses on a proposal that is aimed at developing the research statistical training and consulting aspect of that collaboration. While the UoW and UoN have had active dedicated statistical consultants as part of their consulting units for 21 and 9 years respectively. UWS does not have dedicated consultants, although members of their academic staff do engage in consulting work. Newcastle's Statistical Support Service (UoN SSS) proposes that in collaboration with UWS and UoW, the need for the provision/sharing of dedicated statistical consulting resources with UWS be first established. This presentation should be seen as the beginning of a process towards this goal. As a possible starting point UoN SSS is prepared to provide an initial training course How to Analyse my Data for staff and research students who are enrolled in their research masters or PhDs. We have run this course for the past 4 years at Newcastle and staff and students find it helpful. Key elements of the design of this course will be described. Enrolment numbers will serve as an indicator of interest in the research training aspect. The benefit of the course and the need for follow-up consulting support will be assessed by the UWS course attendees responses to an end-of-course survey.

Key words: Statistical consulting, Research training, ASEARC

Generalised extreme value additive model analysis via variational Bayes

Sarah Neville

University of Wollongong, Australia
sen045@uow.edu.au

Matt Wand

University of Wollongong, Australia
mwand@uow.edu.au

Mark Palmer

Commonwealth Scientific and Industrial Research Organisation, Australia
Mark.Palmer@csiro.au

Abstract

Analysis of sample extremes is becoming more prominent, largely driven by increasing interest in climate change research. In the past decade, additive models for sample extremes have been developed, ranging between Bayesian and non-Bayesian approaches, with various methods of fitting employed. We devise a variational Bayes algorithm for fast approximate inference in Generalised Extreme Value additive model analysis. Such models are useful for flexibility assessing the impact of continuous predictor variables on sample extremes. The crux of the methodology is variational Bayes inference for elaborate distributions. It builds on the work of Wand, Ormerod, Padoan & Fruhwirth (2010) , in which the notion of auxiliary mixture sampling (e.g. Fruhwirth-Schnatter and Wagner, 2006) is used to handle troublesome response distributions such as GEV. The new methodology, whilst approximate, allows large Bayesian models to be fitted and assessed without the significant computing costs of Monte Carlo methods. A much faster analysis results, with little loss in accuracy. We include an illustration of the variational Bayes methodology using maximum rainfall data from Sydney and surrounding regions. This work has been done jointly with Mark Palmer (Commonwealth Scientific and Industrial Organisation) and Matt Wand (University of Wollongong).

References:

Wand, Ormerod, Padoan & Fruhwirth (2010). Variational Bayes for Elaborate Distributions. Unpublished manuscript.

Fruhwirth-Schnatter & Wagner(2006). Auxiliary mixture sampling for parameter driven models of time series of counts with applications to state space modelling. Biometrika, 93, 827841.

Key words: Auxiliary mixture sampling, Bayesian inference, generalized additive models, sample extremes, variational approximation

Prior Sensitivity Analysis for a Hierarchical Model

Junaidi, E. Stojanovski , D. Nur

The University of Newcastle, Callaghan, NSW, 2308, AUSTRALIA
Junaidi@newcastle.edu.au, Elizabeth.Stojanovski@newcastle.edu.au, Darfiana.Nur@newcastle.edu.au

Abstract

Meta-analysis can be presented in the Frequentist or Bayesian framework. Based on the model of DuMouchel, a simulation study is conducted which fixes the overall mean and variance-covariance matrix to generate estimates of the true mean effect. These estimates will be compared to the true effect to assess bias. A sensitivity analysis, to measure the robustness of results to the selection of prior distributions, is conducted by employing Uniform and Pareto distributions for the variance components, the *t*-distribution for the overall mean component and a combination of priors for both variance and mean components respectively. Results were more sensitive when the prior was changed only on the overall mean component.

Keywords: Sensitivity analysis, hierarchical Bayesian model, meta-analysis

1. Introduction

Meta analysis is a statistical method used to obtain an overall estimate by combining results from several individual related studies [10]. Combining results of comparable studies to obtain an overall estimate of treatment effect (e.g. odds ratio, relative risk, risk ratio) can reduce uncertainty and can be useful when the sample size used in each study is small in an attempt to increase power [16].

Meta-analyses can be presented in the Frequentist or Bayesian framework. Within the Frequentist framework, hypotheses are based on information presented within studies and results are often presented in term of 95% confidence intervals to estimate parameters [7]. Weighted averages tend to be used as the overall treatment effect from individual study estimates. One of the more common models used is the inverse of the within-study variances ([2], [11]).

Bayesian methods combine prior probability distributions that reflect a prior belief of the possible values, with the (likelihood) distributions based on the observed data, to produce the posterior probability distributions. The methods are based on the Bayesian rule for probability and can be considered an alternative approach to statistical inference. By multiplying the prior probability with the likelihood, information about the parameters which come from the observed data can be combined with information from the prior distribution that is external to the data ([2], [3]). The posterior distribution can be explained in terms of probabilities and can be considered as borrowing strength from the other studies.

2. Methods

Hierarchical Bayesian Model

A variety of Bayesian methods have been developed in meta-analysis which include those developed by DuMouchel ([13], [14], [15]). The standard hierarchical Bayesian model proposed by DuMouchel [13] provides the following distributional assumptions:

Level

$$Y_i \sim N(\theta_i, \sigma_Y{}^2 W_Y) \qquad i = 1, 2, \ldots, n \qquad \textbf{(1)}$$

$$\sigma_Y^2 \sim \frac{\chi^2_{\upsilon_Y}}{\upsilon_Y}$$

$$\theta_i \sim N(\mu, \sigma_\theta{}^2 W_\theta) \qquad \textbf{(2)}$$

$$\sigma_\theta^2 \sim \frac{\chi^2_{\upsilon_\theta}}{\upsilon_\theta}$$

$$\mu \sim N(0, D \to \infty)$$

The Model has 2 levels, level one indicates data from the studies, and the next level refers to study-specific parameters.

Level 1: $Y_i \sim N(\theta_i, \sigma_Y{}^2 W_Y)$

In the Model, n denotes the number of studies (i = 1, 2,…, n). Y_i indicates the observed statistics following the normal distribution with mean (θ_i.) and covariance matrix ($\sigma_Y^2 W_Y$). Furthermore, W_Y indicates the observed precision matrices (inverse observed variance-covariance matrix) describing within-study variation. If studies are assumed independent, W_Y is to be a diagonal matrix with the individual estimates of the variance of Y_i on the diagonal. σ_Y^2 indicates the degree of uncertainty around the observed precision matrix, as expressed through the respective degree of freedom V_Y which denotes set to the average number of cases of studies (df = n-1). The chi-square distribution is defined by parameter V_Y to denote how well known the variance structure W_Y.

Level 2: $\theta_i \sim N(\mu, \sigma_\theta{}^2 W_\theta)$

θ_i denotes study-specific parameters following the normal distribution with mean (μ = an overall mean) and covariance matrix ($\sigma_\theta^2 W_\theta$). W_θ is the prior precision matrix describing between-study variation. Independence is assumed between studies, so the precision matrices are all diagonal. σ_θ^2 indicates the degree of uncertainty around the prior precision matrix, as expressed through the respective degree of freedom V_θ which denotes set to equal to the number of studies (df = n -1). The chi-square distribution is defined by parameter V_θ to denote how well known the variance structure W_θ. μ is an overall mean following the normal distribution with mean (0) and variance ($D \to \infty$). $D \to \infty$ indicates that elements of D are very large and tend to infinity.

Prior Sensitivity Analysis

The prior distribution plays a crucial role in Bayesian analysis. The conclusion obtained using the Bayesian approach is dependent on the prior distributions. The choice of prior(s) distribution must be determined with care, particularly when the likelihood doesn't dominate the prior. The prior distribution will be less important when the number of studies is large. The non-informative prior distribution will be very useful in the situation when prior information, expectations and beliefs are minimal or not available. In a multi-parameter setting, the specification or elicitation of prior beliefs is not an easy task. Uniform priors or Jeffrey's prior are assumed non-informative priors. The use of vague priors can be problematic due to small amount of data. Hence choosing a vague prior distribution is heavily dependent on the situation ([5], [12]).

DuMouchel [15] stated that results can be affected by different specifications of prior distributions. Sensitivity analysis, to measure the robustness of results regarding selection of prior distributions, should always be carried out. The final results in terms of posterior distributions in meta-analysis will be more robust if the results obtained are unchanged via a sensitivity analysis [12].

Using different prior distributions, for variance components for the within and between studies standard deviation σ, were specified on the Model. However it should be realized that specification of a prior distribution on the standard deviation scale, implies a distribution on the variance and precision scales. The parameterisations for the different prior distributions are described in WinBugs. The prior distributions used on the model are presented in Table 2.1.

Variance components	$\sigma^2 \sim$ Uniform(1/1000, 1000) $1/\sigma^2 \sim$ Pareto(1, 0.25)
Overall mean	$\mu \sim t$-distribution
Combination	$\sigma^2 \sim$ Uniform(1/1000, 1000) $1/\sigma^2 \sim$ Pareto(1, 0.25) $\mu \sim t$-distribution

Table 2.1. Prior distributions used on the model due to sensitivity analysis.

3. Results

A simulation study for the model is presented. By employing 1,000 random samples in each of 30 studies, the R code program was created to simulate from the multivariate normal distribution. By fixing the overall mean and variance-covariance matrix, we generate estimates of the true mean effect. These estimates will be compared to the true effect to assess the bias. Steps used for the simulation study will be as follow.

Step 1. We fix the value of an overall mean (μ).

Step 2. We generate θ_i based on the μ (where n be the number of studies, i = 1, 2, …, n); Σ_θ indicates symmetric, positive definite nxn variance-covariance matrix.

Step 3. The value of observed statistics (Y_i) will be obtained based on θ_i. Σ_Y denotes a symmetric, positive definite nxn variance-covariance matrix.

Furthermore, by using the risk ratios Y_1, Y_2, …, Y_{30} wobtained from simulation and weighted matrix (inverse of variance), we calculate the value an overall mean in WinBugs based on the DuMouchel model presented in Section 2. We obtained an overall mean value of 2.554 associated with credible interval 2.233-2.878 which was close to the fixed true effect (2.560) confirming the setup of the simulation study.

Sensitivity analysis, to measure the robustness of results to the selection of prior distributions, is conducted. The DuMouchel model utilised the Chi-square distribution on the variance parameters, σ_Y^2 and σ_θ^2. Based on Lambert [12], the Uniform and Pareto distribution will be used here for the variance parameters. The Normal distribution will be utilised initially for the overall mean (μ), consistent with that used by Dumouchel. Furthermore, the normal distribution was imposed for an overall mean on the DuMouchel model will be changed by *t*-distribution without changing the value of variances. In addition, combining the priors on variances and an overall mean will be done. However, some changing on prior cannot be done potentially due to range of distribution not fit. For example, when we change both variances using Uniform (1/1000,1000). Simulation data based on 1,000 random samples for 30 studies here will be generated. The true overall mean (μ) used for this demonstration is 2.560.

Prior distribution for variance components

Spieghelter [6] investigates the uniform prior distribution on the variance.

$$\sigma^2 \sim Uniform\ (1/1000, 1000)$$

By using this distribution for parameters σ_Y^2 as well as σ_θ^2, burn-in for 10,000 iterations on the model, the results show an estimated overall mean are 2.558 and 2.564, respectively. These are close to the true effect (2.560). From this preliminary analysis, the use of these others prior distributions on the model do not appear to have a substantial effect on the true study estimate.

Risks Ratios	Mean	S.D	2.5%	97.5%
$\sigma_Y^2 \sim$ Uniform (1/1000, 1000)				
μ	2.558	0.144	2.283	2.826
$\sigma_\theta^2 \sim$ Uniform (1/1000, 1000)				
μ	2.564	0.089	2.417	2.714
$1/\sigma_\theta^2 \sim$ Pareto (1, 0.25)				
μ	2.459	0.539	1.406	3.497
$1/\sigma_Y^2 \sim$ Pareto(1, 0.25) & $1/\sigma_\theta^2 \sim$ Pareto(1,0.25)				
μ	2.412	0.624	1.173	3.615

Table 3.1. Summary statistics for overall mean (μ) by changing the prior variance components using Uniform and Pareto distributions on the model.

$$1/\sigma^2 \sim Pareto\ (1, 0.25)$$

This is equivalent to a uniform prior distribution for variance in the range (0, 4). By changing parameter variance at σ_θ^2, we obtained the overall mean 2.459 (1.406 – 3.497). The overall mean 2.412 (1.173 – 3.615) was obtained when σ_Y^2 and σ_θ^2 changed using this distribution. Table 3.1 shows summary statistics by changing prior distribution variance components on the Model.

Prior distribution for the overall mean

$\mu \sim t$-distribution (0, k= df)

The *t*-distribution will be employed for the overall mean of the model. Density of the *t*-distribution for degree of freedom 2, 3, 5, 10, 30 and 50 will be compared to the normal distribution (μ = 2.554). The overall estimated mean using the *t*-distribution is presented in Table 3.2. This shows the results of overall mean to be very close to the true parameter value.

Risks ratios	Mean	S.D	2.5%	97.5%
df = 2				
μ	2.556	0.166	2.232	2.875
df = 3				
μ	2.554	0.166	2.226	2.871
df = 5				
μ	2.555	0.165	2.233	2.874
df = 10				
μ	2.554	0.168	2.224	2.874
df = 30				
μ	2.552	0.167	2.228	2.881
df = 50				
μ	2.555	0.167	2.234	2.881

Table 3.2. Summary statistics the overall mean (μ) by changing the prior of mean using *t*-distribution (0, *k*=df) on the model.

Prior distribution for both variance and overall mean

Prior distributions for both the variance (Uniform and Pareto) and overall mean (*t*-distribution) simultaneously were employed for the model. By changing the overall mean using *t*-distribution (0, k=2), $\sigma_Y^2 \sim$ Uniform (1/1000,1000) obtained the overall mean is 2.455 (1.402 – 3.489). When the overall mean was changed using *t*-distribution (0, k=2), $1/\sigma_Y^2$and $1/\sigma_\theta^2 \sim$ Pareto (1, 0.25) the result was 2.406 (1.164 – 3.598). These show the overall mean to be reasonably close to the true effect. Summary results by changing the priors on the variances and mean can be seen in Table 3.3.

Risks Ratios	Mean	S.D	2.5%	97.5%
$\mu \sim t$-distribution (0, k= 2) $\sigma_Y^2 \sim$ Uniform (1/1000, 1000) $\sigma_\theta^2 \sim \chi^2$				
μ	2.455	0.530	1.402	3.489
$\mu \sim t$-distribution (0, k= 2) $1/\sigma_Y^2 \sim$ Pareto (1, 0.25) $1/\sigma_\theta^2 \sim$ Pareto (1, 0.25)				
μ	2.406	0.618	1.1647	3.5944

Table 3.3. Summary statistics for μ by changing the prior on mean using *t*-distribution and variance components using Uniform/Pareto.

4. Conclusions

The simulation study on the model showed the overall estimated mean to be close to the true effect, indicating the estimator as consistent and unbiased. While the prior distribution was imposed on the overall mean only, a change in prior showed results to be consistent. A change in prior on the variance components only and on the combination of variance and mean components simultaneously are more sensitive compared to when modifying the prior on only the mean component.

References

[1] A.J. Sutton, K.R. Abrams, D.R. Jones, "Methods for meta-analysis in medical research", 2004.

[2] A.J. Sutton, K.R. Abrams, "Bayesian methods in meta-analysis and evidence synthesis", *statistical methods in medical research,* 10 277-303, 2001.

[3] C. DiMaggio, S. Galea, "substance use and misuse in the aftermath of terrorism, A Bayesian meta-analysis", *Addiction,* 104 (6) 894-904, 2009.

[4] D.A. Berry, D.K. Stangl, "Meta-analysis in medicine and health policy", Marcel Dekker, New York, 2000.

[5] D. Gamerman, "Markov chain monte carlo : sthocastics simulation for Bayesian inference", 1997.

[6] D.J. Spiegelhalter, K.R. Abrams," Bayesian approach to clinical trials and health-care evaluation", 2004.

[7] F. Bullard, "A brief introduction to Bayesian statistics", 2001.

[8] K. Honoki, E. Stojanovski, "Prognostic significance of P16INK4a alteration for ewing sarcoma", *American cancer society*, 1351-1360, 2007.

[9] K.R. Abrams, C.L. Gillies, P.C. Lambert, "Meta-analysis of heterogeneously reported trials assessing change from baseline", *Statistics in Medicine,* 24 3823-3844, 2005.

[10]M. Borenstein, L.V. Hedges, J.P.T. Higgins, "Introduction to meta-analysis", 2009.

[11]M.J. Batterham, "Investigating heterogeneity in studies of resting energy expenditure in persons with HIV/AIDS: a meta-analysis", *American Journal of Clinical Nutrition*, 81 (3) 702-713, 2005.

[12]P.C. Lambert, A.J. Sutton, "How vague is vague? A simulation study of the impact of the use of vague prior distribution in MCMC using WinBUGS", *Statistics in Medicine,* 24 2401-2428, 2005.

[13]W. DuMouchel, "Predictive cross-validation of Bayesian meta-analysis", *Bayesian Statistics,* 5 107-128, 1996.

[14]W DuMouchel, D.A. Berry, "Meta-analysis for response models", *Statistics in Medicine,* 14 679-685, 1995.

[15]W. DuMouchel, "Bayesian meta-analysis", *Statistical Methodology in the Pharmaceutical Sciences*, 509-529, 1990.

[16]Y.Y. Jung, "Identifying differentially expressed genes in meta analysis via Bayesian model clustering", *Biometrical Journal*, 48 (3) 435-450, 2006.

Is the Basis of a Stock Index Futures Market Nonlinear?

Heni Puspaningrum*, Yan-Xia Lin and Chandra Gulati

University of Wollongong, Wollongong, NSW, 2500, AUSTRALIA

Abstract

A considerable amount of papers use a cost-carry model in modelling the relationship between future contract and spot index prices. The cost-carry model defines basis, $b_{t,T}$, as $b_{t,T} = f_{t,T} - s_t = r(T-t)$, where $f_{t,T}$ denotes the log of future contract price at time t and maturity date at time T, s_t denotes the log of spot index price at time t and r denotes the difference between interest rate and dividend rate. Using daily time series data on future contracts of the S&P 500 index and the FTSE 100 index, as well as the price levels of the corresponding underlying cash indices over the sample period from January 1, 1988 to December 31, 1998, Monoyios and Sarno (The Journal of Future Markets, Vol.22, No. 4, 2002, page: 285–314) argued that there is significant nonlinearity in the dynamics of the basis due to the existence of transaction costs or agents heterogeneity. They found that the basis follows a nonlinear stationary ESTAR (Exponential Smooth Transition Autoregressive) model. However, based on the study with the S&P 500 data series from January 1, 1998 to December 31, 2009, we conclude that there is no significant difference between a linear AR(p) model and a nonlinear STAR model in fitting the data.

Key words: autoregressive model, smooth transition autoregressive model, unit root test, cointegration

1. Introduction

[1] analysed the mean reversion of future basis of S&P 500 and FTSE 100 with daily data spanned from January 1, 1988 to December 31, 1998. In constructing the basis, the spot price is paired up with the future contract price for the nearest maturity. Thus, with this construction, [1] defined basis as $b_t = f_t - s_t$ where f_t denotes the log of future contract price at time t for the nearest maturity date and s_t denotes the log of spot index price at time t. On maturity date, f_t is rolled over for the next contract maturity date. [1] concluded that the two basis follow ESTAR (Exponential Smooth Transition Autoregressive) models.

A STAR model can be written as follow:

$$b_t = \theta_{10} + \sum_{j=1}^{p} \theta_{1j} b_{t-j} + \left[\theta_{20} + \sum_{j=1}^{p} \theta_{2j} b_{t-j} \right] G(\theta, e, b_{t-d}) + \epsilon_t \quad (1)$$

where $\{b_t\}$ is a stationary and ergodic process; $\epsilon_t \sim iid(0, \sigma_\epsilon^2)$; $d \geq 1$ is a delay parameter; $e > 0$ is a constant defining the equilibrium of the STAR model for b_t; θ is a constant determining the speed of adjustment to the equilibrium e. Two simple transition functions suggested by [2] and [3] are logistic and exponential functions:

$$G(\theta, e, b_{t-d}) = \frac{1}{1 + \exp\{-\theta(b_{t-d} - e)\}} - \frac{1}{2}, \quad (2)$$

$$G(\theta, e, b_{t-d}) = 1 - \exp\{-\theta^2 (b_{t-d} - e)^2\}. \quad (3)$$

If the transition function $G(\theta, e, b_{t-d})$ is given by (2), (1) is called a logistic smooth transition autoregressive (LSTAR) model. If the transition function $G(\theta, e, b_{t-d})$ is given by (1), (1) is called an exponential smooth transition autoregressive (ESTAR) model.

[1] argued that an ESTAR model is more appropriate for modelling basis movement than a LSTAR model due to symmetric adjustment of the basis. Furthermore, there is fairly convincing evidence that distribution of the basis is symmetric, for example the evidence provided by [4] using both parametric and nonparametric tests of symmetry applied to data for the S&P 500 index. However, [1] also tested for nonlinearities arising from the LSTAR formulation, then make conclusion confirming that the ESTAR model is more appropriate for modelling basis movement than a LSTAR model.

Using current available data, we would like to know whether the basis of S&P 500 follows an ESTAR model as [1] suggested.

*Correspondence author, email: hp261@uow.edu.au

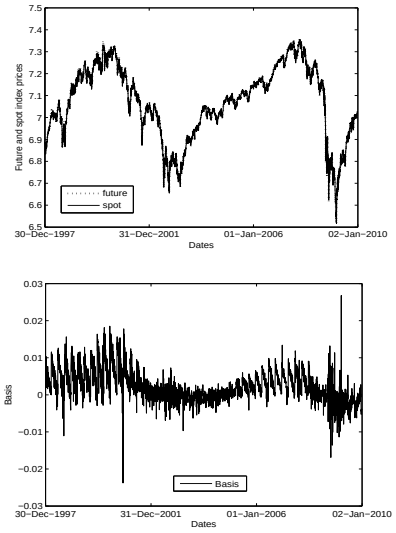


Figure 1: Top: Plot of f_t and s_t; Bottom: Plot of b_t .

2. Empirical Analysis

Using daily closing prices data of future contract and spot index prices of the S&P 500 from January 1, 1998 to December 31, 2009, the procedures in [1] are followed. Figure 1(a) shows the plots of f_t and s_t while Figure 1(b) shows the plot of b_t.

From Figure 1, the plots of f_t and s_t are almost similar indicating the basis b_t which is the difference between f_t and s_t is not large. During the data period, there are 2 major financial crises. The first is in 1999-2002 due to the South American economic crisis in Argentina, Brazil and Uruguay as well as the Dot-com bubble crisis. The second is the financial crisis of 2007 to the present triggered by the US subprime mortgage crisis. Both financial crises are reflected in the fall of the future and spot index prices. The crises are also reflected in the basis value where the basis usually has a negative value during the crisis periods.

2.1. Preliminary Statistics

Table 1 shows some summary statistics for the future prices f_t, the spot index prices s_t, the basis b_t and the demeaned basis mb_t. The PACF plots (not shown in this paper) suggest that both the future and spot index prices show significant spikes at the first 3 lags, but the first spikes is very strong. The PACF plot of the basis displays a slower decay of the PACF with significant spikes at the first five lags, lag 7 , lag 10 and lag 19. Box-Ljung autocorrelation tests statistics for AR(3) residuals using 20 lags for f_t and s_t are 30.0231 [0.0694] and 29.3014 [0.0820], respectively, where the figures in the parentheses are the p-values. Thus, we can accept the null hypothesis of no autocorrelation in residuals for f_t and s_t using AR(3) models and then use $p = 3$ for unit root tests. Box-Ljung autocorrelation tests statistics using 20 lags on mb_t for AR(5), AR(7) AR(10) and AR(19) residuals are 58.0468 [0.0000], 41.2758 [0.0034], 27.1426 [0.1313], 2.7141 [1.0000], respectively. From these results, $p = 10$ is enough to make the residuals become unautocorrelated for mb_t.

Table 1: Summary Statistics

	f_t	s_t	b_t	mb_t
Minimum	6.5160	6.5169	-0.0237	-0.0260
Maximum	7.3627	7.3557	0.0267	0.0244
Mean	7.0711	7.0688	0.0023	-6.60E-06
Variance	0.0279	0.0272	1.74E-05	1.74E-05

Notes: f_t, s_t, b_t and mb_t denote the log of the future prices, the log of the spot index prices, the basis and the demeaned basis, respectively. The demeaned basis is defined as $mb_t = b_t - \bar{b}$, where $\bar{b}$ is the mean of the basis so that the mean of mb_t is zero.

Table 2: Unit Root Tests for S&P 500

Future prices	$f_t^{(c)}$	Lags	Δf_t	Lags
	-2.1139	2	-44.072**	1
Spot Index	$s_t^{(c)}$	Lags	Δs_t	Lags
prices	-2.1255	2	-43.824**	1
Demeaned	mb_t	Lags	Δmb_t	Lags
basis	-7.1598**	9	-25.312**	8

Notes: The statistics are augmented Dickey-Fuller test statistics for the null hypothesis of a unit root process; (c) superscripts indicate that a constant was included in the augmented Dickey-Fuller regression; "Lags " in the fourth column are the lags used in the augmented Dickey-Fuller regression for f_t, s_t, and mb_t while the last column denotes the lags used for Δf_t, Δs_t, and Δb_t; * and ** superscripts indicate significance at 5% and 1%, respectively, based on critical values in [5].

The standard augmented Dickey-Fuller (ADF) unit root tests reported in Table 2 shows that both f_t and s_t are I(1) while mb_t is I(0). Using other lags do not change the conclusions.

Johansen cointegration test (see [6], [7]) is employed and reported in Table 3. The test uses a maximum likelihood procedure in a vector autoregression comprising f_t and s_t, with a lag length of 2 and an unrestricted constant term. We use a lag length of 2 because $p = 3$ is the common lag for f_t and s_t, so that in the vector autoregression, the lag length is $p - 1 = 2$. We also try for other lags such as $p - 1 = 6, 9, 18$, but they do not change the conclusions. Both Johansen likelihood ratio (LR) test statistics clearly suggest that there are 2 cointegrating relationships between f_t and s_t, but the first cointegrating relationship shows much more significant than the second one. Financial theory based on the cost-carry model suggests that the cointegrating parameter equals unity, i.e. in this case means one unit price of f_t is cointegrated with one unit price of s_t or the first cointegrating vector β in the Johansen cointegration test results is [1,-1]. However, from Table 3, the first cointegrating vector, i.e. the first row of β',

Table 3: Johansen Maximum Likelihood Cointegration Results for S&P 500

H_0	H_1	LR
Maximum Eigenvalue LR Test		
$r = 0$	$r = 1$	296.1**
$r \leq 1$	$r = 2$	5.105*
Trace LR Test		
$r = 0$	$r \geq 1$	301.2**
$r \leq 1$	$r = 2$	5.105*
Eigenvalues	Standardized β' eigenvectors	
	f_t	s_t
0.0902856	1.0000	-1.0124
0.00163014	0.37218	1.0000
LR-test restriction = $\chi^2(1)$		83.801 [0.0000] **

in the Johansen cointegration test results for the data is [1,-1.0124]. Imposing the restriction of the first row of β' equals [1,-1] produces the χ^2 statistic reported in the last row of Table 3. It concludes that there is not enough support for the restriction. It is quite different conclusion compared to [1] where they concluded that only one cointegrating relationship exists and the cointegrating relationship with the restriction of $[1, -1]$ can be supported.

2.2. Linearity Tests

Table 4 reports linearity tests results. The first linearity test employed is a RESET test (see [8]) of the null hypothesis of linearity of the residuals from an AR(10) for mb_t against the alternative hypothesis of general model misspecification involving a higher-order polynomial to represent a different functional form. Under the null hypothesis, the statistics is distributed as $\chi^2(q)$ with q is equal to the number of higher-order terms in alternative model. Table 4 reports the result from executing RESET test statistic where the alternative model with a quadratic and a cubic terms are included. The null hypothesis is very strongly rejected considered with the p-value of virtually zero, suggesting that a linear AR(10) process for mb_t is misspecified.

The second linearity tests are based on [3]. The tests can also be used to discriminate between ESTAR or LSTAR models since the third-order terms disappear in the Taylor series expansion of the ESTAR transition function. The artificial regression of (1) is estimated as follow:

$$\begin{aligned} b_t &= \phi_{00} + \sum_{j=1}^{p} \Big(\phi_{0j} b_{t-j} + \phi_{1j} b_{t-j} b_{t-d} + \\ &\quad \phi_{2j} b_{t-j} b_{t-d}^2 + \phi_{3j} b_{t-j} b_{t-d}^3\Big) + \\ &\quad \phi_4 b_{t-d}^2 + \phi_5 b_{t-d}^3 + errors \end{aligned} \tag{4}$$

where ϕ_4 and ϕ_5 become zero if $d \leq p$. Keeping the delay parameter d fixed, testing the null hypothesis

$$H_0 : \phi_{1j} = \phi_{2j} = \phi_{3j} = \phi_4 = \phi_5 = 0,$$

Table 4: Linearity tests on the demeaned basis mb_t

RESET Test			11.8 [0.00]**
Lags Used			10
d	LM^G	LM^3	LM^E
p =7			
1	15.6 [0.00]**	7.1 [0.00]**	19.6 [0.00]**
2	15.6 [0.00]**	7.1 [0.00]**	19.6 [0.00]**
p =10			
1	11.0 [0.00]**	4.6 [0.00]**	14.1 [0.00]**
2	11.0 [0.00]**	4.6 [0.00]**	14.1 [0.00]**

Notes: RESET test statistic is computed considering a linear AR(p) regression with 10 lags without a constant as the constant is not significant at 5% significant level against an alternative model with a quadratic and a cubic term. The F-statistics forms are used for the RESET test, LM^G , LM^3 and LM^E and the values in parentheses are the p-values. * and ** superscripts indicate significance at 5% and 1%, respectively.

$\forall j \in \{1, \cdots, p\}$ against its complement is a general test (LM^G) of the hypothesis of linearity against smooth transition nonlinearity. Given that the ESTAR model implies no cubic terms in the artificial regression (i.e.: $\phi_{3j} = \phi_5 = 0$ if the true model is an ESTAR model, but $\phi_{3j} \neq \phi_5 \neq 0$ if the true model is an LSTAR), thus, testing the null hypothesis that

$$H_0 : \phi_{3j} = \phi_5 = 0,$$

$\forall j \in \{1, \cdots, p\}$ provides a test (LM^3) of ESTAR nonlinearity against LSTAR-type nonlinearity. Moreover, if the restrictions $\phi_{3j} = \phi_5 = 0$ cannot be rejected at the chosen significance level, then a more powerful test (LM^E) for linearity against ESTAR-type nonlinearity is obtained by testing the null hypothesis

$$H_0 : \phi_{1j} = \phi_{2j} = \phi_4 = 0 | \phi_{3j} = \phi_5 = 0,$$

$\forall j \in \{1, \cdots, p\}$.

A lag length of 7 and 10 are considered for executing the linearity tests for mb_t using the artificial regression in (4). Table 4 shows values of the test statistics LM^G, LM^3 and LM^E. The delay parameter $d \in \{1, 2, 3, 4, 5\}$ are considered. We only report for $d = 1$ and $d = 2$ as the values are the same for other d. However, the test statistics LM^G, LM^3 and LM^E show that different values of d do not affect the results. From Table 4, the p-values from LM^G, LM^3 and LM^E statistics are virtually zero for both $p = 7$ and $p = 10$. From the LM^G statistics, we can conclude that linearity is strongly rejected. From the LM^3 and LM^E statistics, we can conclude that a LSTAR model is much more strongly supported than an ESTAR model. It is quite different conclusion compared to [1] where they concluded the opposite one, i.e. an ESTAR model is more favoured than a LSTAR model. From Table 4, the LM^G, LM^3 and LM^E statistics for $p = 7$ are higher than those for $p = 10$. Therefore, we choose a LSTAR model with $p = 7$ and $d = 1$ for model estimation.

Table 5: Estimation results for the demeaned basis mb_t

	LSTAR(7)	AR(7)
$\hat{\theta}_{11}(=-\hat{\theta}_{21})$	0.3642 (0.0177)	0.3672 (0.0179)
$\hat{\theta}_{12}(=-\hat{\theta}_{22})$	0.1885 (0.0189)	0.1864 (0.0190)
$\hat{\theta}_{13}(=-\hat{\theta}_{23})$	0.0856 (0.0193)	0.0902 (0.0193)
$\hat{\theta}_{14}(=-\hat{\theta}_{24})$	0.1086 (0.0192)	0.1095 (0.0192)
$\hat{\theta}_{15}(=-\hat{\theta}_{25})$	0.0556 (0.0193)	0.0604 (0.0193)
$\hat{\theta}_{16}(=-\hat{\theta}_{26})$	0.0084 (0.0189)	0.0131 (0.0190)
$\hat{\theta}_{17}(=-\hat{\theta}_{27})$	0.0713 (0.0176)	0.0708 (0.0178)
θ	-26.0070 (8.7306)	
SSE	0.0214	0.0214
LR	0.000 [1.0000]	0.000 [1.0000]
SW	0.891 [0.0000]**	0.892 [0.0000]**
BL (20)	40.287 [0.0046]**	41.372 [0.0033]**

Notes: Figures in parentheses beside coefficient estimates denote the estimated standard errors. SSE is sum square error; LR is a likelihood ratio statistic for parameter restrictions; SW is a Shapiro-Wilk normality test for residuals; BL is a Box-Ljung autocorrelation test for residuals using 20 lags; the figures in parentheses denote the p-values.

2.3. Estimation Results

Table 5 reports comparison of model estimation results for a nonlinear LSTAR model with $p = 7$ and $d = 1$ and a linear AR(7) model. The nonlinear LSTAR model estimation uses a nonlinear least squares method in the form of (1) and (2) for mb_t. As the mean of mb_t is zero, theoretically, $\theta_{10} = \theta_{20} = e = 0$. Further restriction of $\theta_{2j} = -\theta_{1j}$ for $j = 1, \cdots, 7$ produces the likelihood ratio statistic, LR, in Table 5 concluding that the restrictions can not be rejected at the conventional 5% significance level. A linear AR(7) is also estimated as a comparison. The LR statistic comparing the LSTAR model and the AR(7) model concludes that there is no significant difference between the two models. Furthermore, the parameter estimates of θ_{1j}, $j = 1, \cdots, 7$, for the two models are quite similar. Other statistics such as Shapiro-Wilk normality test and Box-Ljung autocorrelation test for residuals are also similar for the two models. [1] did not make model comparison and they concluded that a nonlinear ESTAR model quite fits with the data they have.

3. Conclusion

Using current available data, from January 1, 1998 to December 31, 2009, we examine the basis of S&P 500 following procedures in [1]. Even though we can conclude that there is possibility nonlinearity in the basis, there is no significant difference between a nonlinear LSTAR model and a linear autoregressive model in fitting the data. It is a different conclusion compared to [1] concluding that a nonlinear ESTAR model quite fits with the data they have.

Our data has two major financial crises while the data used by [1] does not have a major financial crisis. This different data characteristic may lead to different conclusions.

We also have a concern in the way the basis is constructed. By pairing up the spot price with the future contract with the nearest maturity, it may produce artificial jumps at the time of maturity. The longer the time to maturity, the higher the difference between the future price and the spot price. For example for S&P 500, it has 4 maturity times during a year which are the third Friday in March, June, September and December. We find that at those times, there are jumps in the basis. Figure 2 shows the plot of b_t from January 1, 1998 to October 19, 1998 with jumps on the third Friday in March, June, September 1998. [1] did not discuss this issue. [9] argued that it may create volatility and bias in the parameter estimates. Therefore, the next step of this research will examine the cointegration of f_t and s_t with a time trend for each future contract.

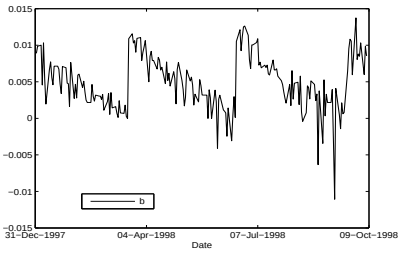


Figure 2: Plot of b_t from January 1, 1998 to October 19, 1998.

References

[1] M. Monoyios, L. Sarno, Mean reversion in stock index futures markets: A nonlinear analysis, The Journal of Futures Markets 22 (4) (2002) 285–314.

[2] C. Granger, T. Terasvirta, Modelling Nonlinear Economic Relationship, Oxford University Press, Oxford, 1993.

[3] T. Terasvirta, Specification, estimation and evaluation of smooth transition autoregressive models, Journal of the American Statistical Association 89 (1994) 208–218.

[4] G. Dwyer, P. Locke, W. Yu, Index arbitrage and nonlinear dynamics between the S&P 500 future and cash, Review of Financial Studies 9 (1996) 301–332.

[5] W. Fuller, Introduction to Statistical Time Series, John Wiley, New York, 1976.

[6] S. Johansen, Statistical analysis of cointegrating vectors, Journal of Economic Dynamics and Control 12 (2/3) (1988) 231–254.

[7] S. Johansen, Estimation and hypothesis testing of cointegartion vectors in gaussian vector autoregressive models, Ecnometrica 59 (6) (1991) 1551–1580.

[8] J. Ramsey, Tests for specification errors in classical least-squares regression analysis, Journal of the Royal Statistical Analysis, Series B, 13 (1969) 350–371.

[9] C. Ma, J. Mercer, M. Walker, Rolling over futures contracts: A note, The Journal of Futures Markets 12 (1992) 203–217.

Threshold Autoregressive Models in Finance: A Comparative Approach

David Gibson

The University of Newcastle, Callaghan, NSW, 2308, AUSTRALIA

Abstract

Financial instruments are known to exhibit abrupt and dramatic changes in behaviour. This paper investigates the relative efficacy of two-regime threshold autoregressive (TAR) models and smooth threshold autoregressive (STAR) models , applied successfully to econometric dynamics, in the finance domain. The nature of this class of models is explored in relation to the conventional linear modeling approach.

Key words: Threshold, Nonlinear, Autoregressive, STAR

1. Introduction

Autoregressive models have been applied across diverse fields of endeavour. Yule (1927) applied the first autoregressive model to the understanding of Wolfer's sunspot numbers over time, but authors such as Pesaran and Timmerman (1995) have extended the autoregressive model into the financial domain.

Practitioners in many fields are increasingly faced with real data possessing nonlinear attributes. It is known that stationary Gaussian autoregressive models are structurally determined by their first two moments. Consequently, linear autoregressive models must be time reversible. Many real datasets are time irreversible, suggesting that the underlying process is nonlinear. Indeed, in Tong's seminal paper on threshold models, he would argue that no linear Gaussian model could explain the cyclical dynamics observed in sections of the lynx data (Tong and Lim, 1980). Furthermore, he argued that characteristics of nonlinear models, such as time irreversibility and limit cycles, mandated the development of practical nonlinear models to help resolve ongoing difficulties in real data. Tong's explanation and application of locally linear threshold models introduced striking opportunities for model building strategies.

2. Threshold Autoregressive (TAR) Models

The *Threshold Autoregressive* (TAR) family proposed and explained by Tong (1983) are contained within the state-dependent (regime-switching) model family, along with the bilinear and exponential autoregressive (EAR) models.

The simplest class of TAR models is the *Self Exciting Threshold Autoregressive* (SETAR) models of order p introduced by Tong (1983) and specified by the following equation:

$$Y_t = \begin{cases} a_0 + \sum_{j=1}^{p} a_j Y_{t-j} + \epsilon_t & \text{if } Y_{t-d} \leq r \\ (a_0 + b_0) + \sum_{j=1}^{p} (a_j + b_j) Y_{t-j} + \epsilon_t & \text{if } Y_{t-d} > r \end{cases} \tag{1}$$

TAR models are piecewise linear. The threshold process divides one dimensional Euclidean space into k regimes, with a linear autoregressive model in each regime. Such a process makes the model nonlinear for at least two regimes, but remains locally linear (Tsay, 1989). One of the simplest of TAR models equates the state determining variable with the lagged response, producing what is known as a *Self-Exciting Threshold Autoregressive* (SETAR) model.

A comparatively recent development is the *Smooth Transition Autoregressive* (STAR) model, developed by Terasvirta and Anderson (1992). The STAR model of order p model is defined by

$$\begin{aligned} Y_t =& a_0 + a_1 Y_{t-1} + ... + a_p Y_{t-p} + \\ & (b_0 + b_1 Y_{t-1} + ... b_p Y_{t-p}) G\left(\frac{Y_{t-d} - r}{z}\right) + \varepsilon_t, \end{aligned} \tag{2}$$

where d, p, r, $\{\varepsilon_t\}$ are as defined above, z is a smoothing parameter $z \in \Re^+$ and G is a known distribution function which is assumed to be continuous. Transitions are now possible along a continuous scale, making the regime-switching process 'smooth'. This helps overcome the abrupt switch in parameter values characteristic of simpler TAR models.

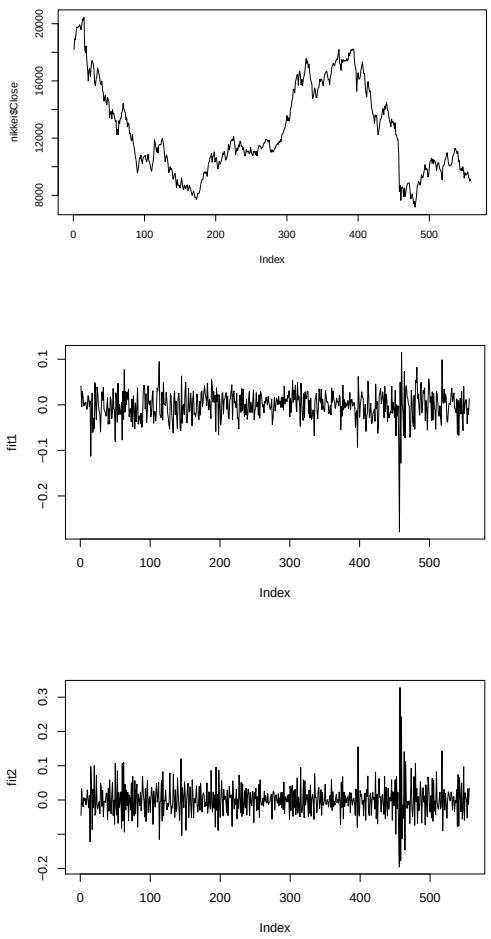


Figure 1: Sequence plot for the Nikkei-225 Index and First difference of the logged Nikkei-225 data

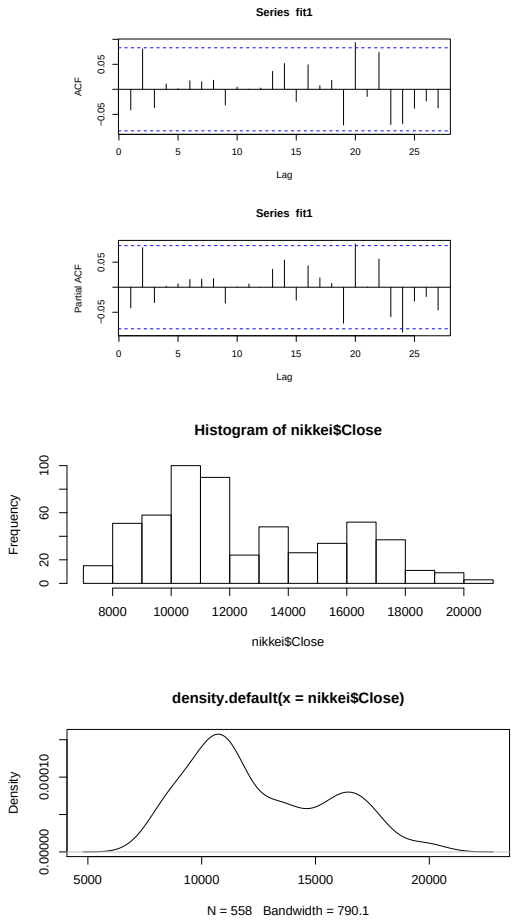


Figure 2: Autocorrelation function (ACF) and partial autocorrelation function (PACF) of the logged and differenced series and Histogram and density plot for the Nikkei-225 data

3. NIKKEI-225 Index Case Study

Weekly closing value data was obtained from the NIKKEI 225 Index, an average of 225 stocks traded on the Tokyo Stock Exchange, from January 2000 to September 2010. Characteristics of a nonlinear process seem to be present (as shown in Figure 2). Irregular amplitude in the peaks and troughs suggest time irreversibility and asymmetry. Peaks rise and fall dramatically, and there appears to be a very slight downward trend. The first difference of the logged data (Figure 2) revealed an irregular spike around the 450th data point.

3.0.1. Sample Correlation

The logged and differenced series (Figure 3), both the ACF and PACF reveal no significant values until the 20th lag. Non-significant lags can be evidence towards nonlinearity. A histogram and density plot of the data suggest a bimodal distribution, a characteristic common to nonlinear processes.

3.0.2. Linearity Testing

Test	Tar-F Test	LR Test	Keenan Test
Statistic	1.472	35.913	10.398
p	< 0.01	0.143	< 0.01

Disagreement is noted between the tests over the nature of the process. Tsay's TAR-F test successfully rejects the null hypothesis of a linear process, while Chan's Likelihood Ratio test fails to do so. A possible reason for this is that Chan's test retains the greatest power when the alternative is the "true" model under consideration.

3.0.3. Identification and Estimation

Iterative linear model building strategies with varying autoregressive and moving-average parameters met with mixed results. After many attempts, an appropriate linear model was specified for baseline comparison.

```
Call:
arima(x = fit2, order = c(4, 0, 0))

Coefficients:
          ar1      ar2      ar3      ar4  intercept
```

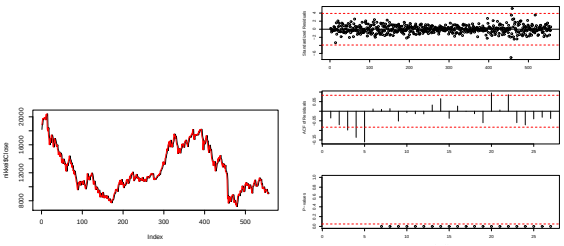


Figure 3: Fitted values for the AR(4) model and Residuals for the AR(4) model

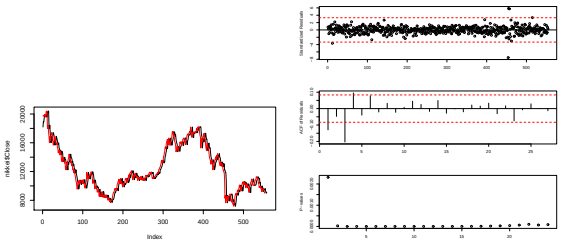


Figure 4: Fitted values for the SETAR model and Diagnostics for the SETAR model

```
coef    -8.394919e-01 -5.584072e-01 -4.074595e-01 -2.018603e-01 -4.997183e-05
p-value  1.053899e-90  5.544701e-27  3.859383e-15  1.148524e-06  9.203625e-01
```

All coefficients are strongly significant here in the AR(4) base model. The final chosen SETAR model had a threshold delay d of 0, and autoregressive order 2 in the lower and upper regimes. This gives:

```
SETAR(2, 2 , 2 ) model delay = 0

                Estimate Std.Err  t-value Pr(>|t|)
intercept-fit2  -0.0265  0.0016 -16.2387   0.0000
lag1-fit2       -0.3437  0.0412  -8.3466   0.0000
lag2-fit2       -0.1290  0.0391  -3.2986   0.0011

                Estimate Std.Err t-value Pr(>|t|)
intercept-fit2   0.0258  0.0022 11.9718    0e+00
lag1-fit2       -0.5495  0.0568 -9.6697    0e+00
lag2-fit2       -0.1768  0.0513 -3.4490    7e-04
```

All parameter estimates are significant in this model, with a considerable decrease in AIC. The SETAR (2,2,2) model obtained in the output above is:

$$Y_t = \begin{cases} -0.0265 - 0.3437Y_{t-1} - \\ \quad 0.1290Y_{t-2} + 0.0257\epsilon_t & Y_{t-1} \leq 0.00087 \\ \\ 0.0258 - 0.5495Y_{t-1} - \\ \quad 0.1768Y_{t-2} + 0.0309\epsilon_t & Y_t > -0.00087 \end{cases} \quad (3)$$

An additional procedure for automatic STAR model estimation was employed. This produced a satisfactory model, but integrated tests for regime addition beyond the first threshold were rejected.

```
Testing linearity...   p-Value =  1.160231e-09
The series is nonlinear. Incremental building procedure:
Building a 2 regime STAR.
Performing grid search for starting values...
Starting values fixed: gamma =  34 , th =  -1.108075 ; SSE =  330.1924
Optimization algorithm converged
Regime  1 :
Linear parameters: -10.838266, -4.2711662, -0.022057
Regime  2 :
Linear parameters: 10.8208907, 3.6067803, -0.2918321
Non-linear parameters:
33.9990364, -2.9458302
```

3.0.4. Diagnostics

Residual plots of the linear model (Figure 4) reveal non-random residual scatter with distinct point compression near the centre, and autocorrelation from the 3rd to the 5th lag. Introducing the threshold value has improved model fit statistics (Figure 5), but there is little appreciable improvement in the fitted values relative to the original process. Standardised residuals

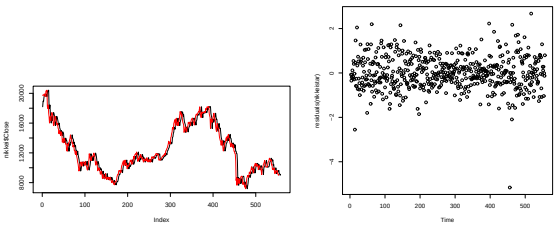


Figure 5: Fitted values for the STAR model and Residuals for the STAR model

demonstrate point dispersal at the two extremes. The sample ACF reveals significant lags 1, 3 and 4. The failure to reject the null hypothesis in a test for the inclusion of additional regimes has reverted the STAR model in this case to the simpler SETAR form.

3.0.5. Forecasting

In-sample forecasting involves subtraction of the final few data points and using the selected model for prediction of these known values. Five values were removed as shown below.

```
[1] 9642.12  9253.46  9179.38  8991.06  9114.13
```

These values can be compared with those generated by the chosen linear and nonlinear models, of which the following correspond with the plots in Figures 7 and 8.

```
AR(4)
[1] 9445.729 9429.617 9413.498 9397.382 9381.266
SETAR
```

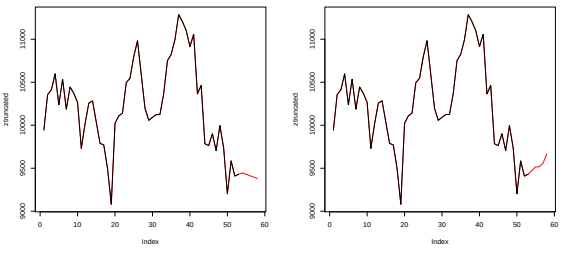


Figure 6: Restricted in-sample forecast values for the AR(4) model and Restricted in-sample forecast values for the SETAR model

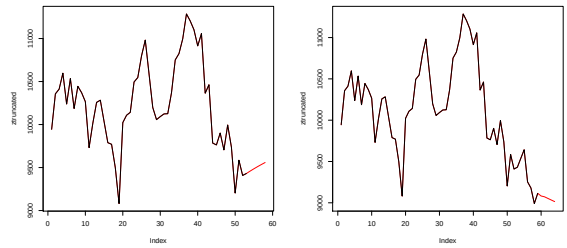


Figure 7: Restricted in-sample forecast values for the STAR model and Restricted out-of-sample forecast values for the AR(4) model

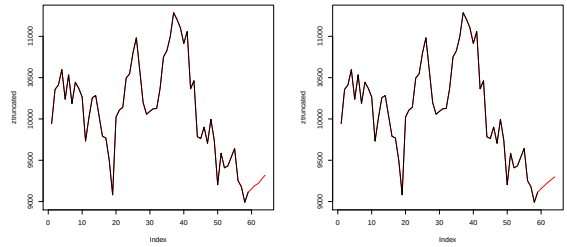


Figure 8: Restricted out-of-sample forecast values for the SETAR model and Restricted out-of-sample forecast values for the STAR model

```
[1] 9472.965 9515.371 9516.460 9562.410 9666.390
STAR
[1] 9458.754 9485.459 9510.953 9535.287 9558.515
```

The fitted values for the AR(4) model fall centrally within the peak and trough of the original data, and is indicative of a reasonable model fit (see Figure 7). The SETAR model appears to be having difficulty in accurately predicting the final few values. A similar outcome is noted for the STAR model (Figure 8), with the predictions unable to suitably account for the drop in returns in the final few values. Calculations for the out-of-sample forecasts are shown below.

```
AR(4)
[1] 9082.537 9072.741 9053.255 9033.856 9014.585
SETAR
[1] 9152.935 9195.608 9219.340 9271.747 9319.628
STAR
[1] 9156.223 9194.343 9230.701 9265.431 9298.609
```

The out-of-sample predictions from the linear model reveal a slow downward trend as the series progresses. The SETAR model forecasts could be interpreted as an improvement on the ARIMA model, as shown in Figure 9. Rising values might indicate an attempt by the model to more effectively capture the volatility in the series, and reflect overall movements in the process. STAR model prediction values resemble strongly those seen in the SETAR model.

4. Conclusion

The extension of the autoregressive model to the regime-switching class is a natural progression, but the characteristics of nonlinearity are not always immediately detectable in visual plots such as the sequence chart. Within the two-regime Self-Exciting Threshold Autoregressive (SETAR) model simulation series these attributes are brought to the forefront and allow for the relative strengths of this class of models to be quantified. Exploratory analyses reveal with great speed the violation of linear modeling assumptions and the difficulties inherent to fitting this type of model. Data asymmetry and time irreversibility are traditional indicators of nonlinearity, while the rejection of normality in the data makes applying ARIMA models hazardous.

SETAR model building strategies are, conversely, ideal for this type of process. Model fitting summaries and diagnostic procedures reveal clear preferences, while both in-sample and out-of-sample forecasts are improved in the threshold model case. An unsurprising result, it suggests that real data sets demonstrating similar behaviour may benefit from the application of this model form. Empirical results from the application of the nonlinear models highlight the improvements in out-of-sample forecasting. Similar performance was also noted between the SETAR and STAR models in applying process dynamics to future data points. This outcome can be interpreted, however, as the result of smooth models with finite thresholds. In-sample forecast remains problematic, as in several cases the models were unable to properly replicate the observed model behaviour.

References

[1] Gibson, David., (2010) "Threshold Autoregressive Models in Finance: A Comparative Approach", Master's Thesis, The University of Newcastle

[2] Pesaran, M Hashem, Timmerman, Allan., (1995) "Predictability of stock returns: Robustness and economic significance", The Journal of Finance, Vol. 50, pp 1201 - 1228

[3] Terasvirta, T, Anderson, H M., (1992) "Characterizing Nonlinearities in Business Cycles Using Smooth Transition Autoregressive Models", John Wiley & Sons, Vol. 7

[4] Tong, Howell., (1983) "Threshold Models in Non-linear Time Series Analysis", Springer-Verlag New York Inc.

[5] Tong, H., Lim, K.S., (1980) "Threshold autoregression, limit cycles, and cyclical data", Journal of the Royal Statistical Society, Vol. 42, pp 245 - 292

[6] Tsay, Ruey S., (1989) "Testing and modeling threshold autoregressive processes", Journal of the American Statistical Association, Vol. 84, pp 231 - 240

[7] Yule, Udny., (1927) "On a method for investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers", Philosophical Transactions of the Royal Society of London, Vol. 226, pp 267 - 298

The Analysis of Marine Mammal Take Data from the Hawaiian Deep-Set Longline Fishery

Bryan Manly

Western EcoSystems Technology Inc., Australia
bmanly@west-inc.com

Abstract

The analysis of data on interactions (”takes”) between marine mammals and fishing vessel is important because these interactions are viewed very seriously, particularly if they result in a death or serious injury. For the Hawaiian deep-set longline fishery the data available for analysis for a marine mammal species in the whole fishery or part of it consists of a list of the trips by vessels with a federal government observer on board, with information for each trip on the year when the trip ended, the quarter of the year when the trip ended, the probability that the trip was selected to be observed, and the number of marine mammals takes that occurred. Approximately 30% of trips have an observer on board, and information on the total number of fishing trips is also available. In this talk I will outline the methods used for (a) comparing the number of takes in the different quarters of the year (with randomization tests), (b) estimating take rates and total take numbers (with Horvitz Thompson estimators), (c) estimating standard errors for take rates and total numbers (using parametric bootstrapping), and (d) finding 95% confidence limits for take rates and numbers (based on the binomial distribution assuming a maximum of one take per trip).

Key words: randomisation test, bootstrapping, Horvitz Thompson estimators

Systems theory and improving healthcare

Peter P. Howley[1], Sheuwen W. Chuang[2]

[1]The University of Newcastle, Callaghan, NSW, 2308, AUSTRALIA
Peter.Howley@newcastle.edu.au
[2] Taipei Medical University, Taipei, Taiwan
Sheuwen@tmu.edu.tw

Abstract

The accreditation and quality measurement and reporting systems in health care organisations are believed to influence patient safety and quality of care. In order to gain knowledge about the effects of these two systems, an holistic healthcare systems relationship model was recently constructed and a series of adaptive-control studies developed to explore relationships between segments within the systems relationship model.

This paper describes where we've been, where we are and where we're headed: the studies, the models, the supporting research and the systems theory-based approach encompassing the current direction.

Keywords: clinical indicators, control relationship, quality measurement and reporting, accreditation, systems theory, feedback, Bayesian, posterior predictive

1. Introduction

Clinical indicators (CIs) are essentially measures of performance in a clinical setting. They are intended to be used as screening tools that can identify possible problems or opportunities to improve processes and outcomes in health care organisations (HCOs). Since 1993, Australian HCOs preparing for accreditation, or re-accreditation, with the Australian Council on Healthcare Standards (ACHS) have submitted data on sets of CIs. The ACHS routinely collates this information in 6-month periods and generates CI reports that are provided to the HCOs and accreditation surveyors. In 2009 the ACHS received data from 671 Australian and New Zealand HCOs on 370 CIs across 23 specialties [1].

Individual HCOs who have contributed their CI data to the ACHS receive two kinds of CI reports. The reports are personalised for the HCO and report on the CIs on which they provided data. One of these two reports is a bi-annual 'six-monthly' report which provides aggregated results across all contributing HCOs as well as a comparison with the individual HCO's 'peer' organisations. The second is an annual 'trend' report, which shows comparative information for the period covered, starting from as early as 2001. The six-monthly report provides more simplistic statistical comparisons, is less descriptive and shorter than the trend report.

Additionally, an annual report on the healthcare system's performance is provided which does not identify specific HCOs but instead reports upon performances across the entire healthcare system by clinical area. The report, currently in its 11th edition, is named the "Australasian Clinical Indicator Report 2001-2009: Determining the Potential to improve quality of care: 11th Edition", or DPI report. It is designed for use by governments and medical colleges for directing policy and assisting with resource allocation requirements within clinical areas. The three reports stem from the existing quality and measurement reporting (QMR) system.

Chuang and Inder (2009) [2] identified an holistic healthcare systems relationship model. The model provides the platform for a series of adaptive-control studies into the control and communication relationships between the accreditation system, the QMR system and the HCO-level system. They are referred to as adaptive since the studies can be adapted to any similar accreditation and QMR system despite being conducted in the ACHS setting.

This paper summarises the key developments over the past decade, the current research and the future direction in the improvement of health care via the

QMR and accreditation systems in Australia. The ACHS is the chosen setting.

2. Developments over the past decade

The QMR System

The development of methods for measuring and reporting on the ACHS clinical indicators has been continuing over the past decade. Major improvements during this period have included the introduction of Bayesian hierarchical models, the estimation of potential gains [3] and trend reports.

For a given CI, the i^{th} HCO provides the observed number of patients who incur the 'event of interest' (O_i) and the number of patients at risk of the event (D_i). The CI data can be reported as the observed proportions (O_i/D_i), however, they encompass the between-HCO, or systematic, variation as well as within-HCO, or sampling, variation. Thus, the observed CI proportions for individual HCOs will vary from the 'true underlying proportions' due to sampling variation. In the Bayesian paradigm, a two-stage hierarchical model is used to represent the distributions for the two sources of variation. The first level corresponds to the distribution of the CI proportions across all hospitals, thus representing the systematic variation. The second stage corresponds to the sampling distribution of the O_i.

Bayesian hierarchical models and shrinkage estimators were introduced to account for the effects of sampling variation on the estimated CI proportions For the beta-binomial two-stage hierarchical model, the individual HCO's proportion of admissions having the event of interest, θ_i, is assumed to be drawn from a beta distribution with parameters π and M, where π represents the mean CI proportion and the *spread parameter*, M, indicates the spread of proportions among the hospitals and is inversely related to the variance of the proportions between HCOs, $\sigma^2 = \pi(1-\pi)/(1+M)$. Thus $\theta_i \sim \text{Beta}(\pi, M)$. The O_i, is assumed to follow a binomial distribution, $O_i \sim \text{binomial}(D_i, \theta_i)$ [3].

For each CI, a measure of the potential gains (or reduction in the number of undesirable events) that could be achieved if the mean proportion was shifted to the 20^{th} centile was introduced. Its calculation is based on the amount of variation in the system (represented by the difference in the mean, π, and 20^{th} centile, p_{20}, of the rates across all HCOs) and the impact upon the system, or volume effect, (represented by the summed D_i across all HCOs providing data for the CI) as shown in expression (1).

$$(\pi - p_{20})\sum_{i=1}^{n} D_i \qquad (1)$$

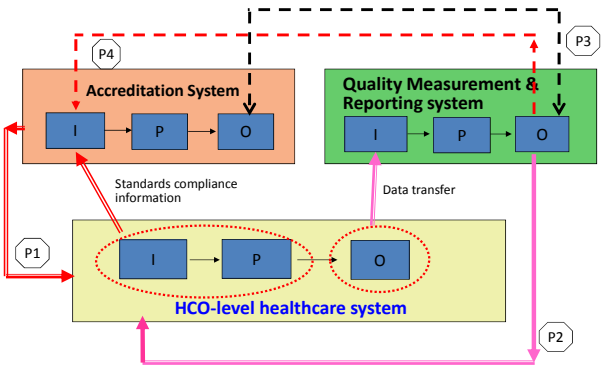
It facilitates and motivates scientific investigation within clinical areas. Smaller variation and smaller potential for system impact (in terms of potential for events occurring, represented by ΣD_i) is reflected in a smaller value for the potential gains. Reported as part of the DPI report, this measure enables comparisons of clinical areas for improvement activity rather than allocating responsibility solely to individual HCOs. Combining with EB shrinkage estimators, it facilitates practicable reports for HCOs.

The annual trend report introduced in the past decade identifies the individual HCO's performance compared with both the entire system of 'peer' organisations, via means, 20^{th} and 80^{th} centile rates, and themselves, based on trend analyses of their 6-monthly rates, after considering both within-HCO and between-HCO variation. It includes both graphical and numerical comparisons, including cumulative excess counts of observations above that expected for the HCO and indicates the HCO's relative performance to itself and the other HCOs contributing information to the CI over time.

Systems theory-based approach to improvement

Chuang and Inder (2009) [2] identified an holistic healthcare systems relationship model and four distinct control and communication relationships, see Figure 1. In summary, P1 is a control relationship that represents hierarchically determined practitioner standards, generated from the accreditation system and given to the HCO-level system. P2 is a communication relationship which represents the communication of outcomes of the QMR system to the HCO-level system for its own internal control response. P3 communicates associations in the outputs of the accreditation and QMR systems. P4 is a control relationship which provides feedback from the QMR's output (CI reports) to the accreditation system's input.

Figure 1. Healthcare systems relationship*



*I – Input, P – Process, O - Output

The three major systems identified in Figure 1 form the health administration system. The safety and

quality characteristics are an emergent property of this system as a whole, not simply its individual system components. The question arises as to the impact of each of these relationships.

Considerable international research into the P1 relationship has revealed varying findings regarding the effectiveness of the accreditation system for providing quality of care [4-7]. Research into the P2 relationship, assessing the effectiveness of the QMR system for improving the quality of care also revealed varying degrees of success. Analysis of the P3 relationship identifies gaps that permit associations between the two outputs to be due to chance. In cases where HCOs achieve accreditation and their QMR approaches are deemed acceptable, partial, inconsistent and conflicting success in improving quality has resulted [8, 9]. For brevity the reader is invited to view Chuang and Inder (2009) for the summary of the results for the P1-P3 relationships.

3. Current

The QMR System

The CI reports involve retrospective analysis and reporting. Such reports, however, could be complemented by tools, such as control charts, that enable HCOs to monitor their performance during the six-month periods of data collection.

A new control chart for monitoring clinical indicator (CI) data based upon the beta-binomial posterior predictive (BBPP) distribution was compared with the more commonly used Bernoulli cumulative sum (Bernoulli CUSUM) chart. Control chart limits were generated and run lengths were simulated for 3894 parameter combinations (π, D_i, θ_i, σ and percentage change in the underlying proportion (required for Bernoulli CUSUM chart)) reflecting the types of CI data collated.

For the case where the underlying proportion of cases with an event of interest had to be estimated, the BBPP chart was shown to have the desired smaller out-of-control average run length in 71% of the simulations.

Whilst these results are promising the charts are yet to be included in the reports and provided to the HCOs.

The Control Relationship (P4)

The P4 control relationship, not previously explored in detail, is crucial towards creating a positive correlation, or *communication*, between the QMR and accreditation systems and achieving continuous quality improvement in the system's outcomes. Accreditation surveyors were identified as a key system component in activating the relationship; if the QMR's CI reports had been both utilised by HCOs to guide quality improvement as well as referenced by surveyors as a tool to assess quality improvement in HCOs, then surveyors could produce valuable feedback to HCOs via the accreditation process.

The authors' study of ACHS accreditation surveyors has revealed half used the CI reports most or all of the time and half also found solely positive responses from the HCOs when discussing the reports, with 20% solely negative (reflecting ignorance of relevance and use). 75% to 89% of surveyors perceived the reports to be useful for the quality and safety objectives for each of senior executive officials, clinicians, safety and quality managers, and accreditation surveyors. Changes to processes to ensure CI reports are not omitted from pre-survey packages along with improved education of surveyors and HCOs on how to better utilise the reports for the purposes of improvement in the safety and quality of healthcare were revealed as significant factors that would increase their usage.

4. Future research

In order to create a well-designed closed feedback loop among the accreditation, QMR systems and HCO-level systems, a series of adaptive studies related to the P1-P4 relationships need to be pursued. The P4 relationship has begun to be explored with the surveyor-based adaptive study reported upon in Section 3. These results identify the relative use of the CI reports by accreditation surveyors and HCOs and the perceived appropriateness of the reports. Implementation of the conclusions augurs well for surveyors ultimately being able to produce valuable feedback to HCOs via the accreditation process. However, that result is still far from a *fait accompli*. A study investigating the current decision process of surveyors undertaking accreditation investigations and how CI reports may need to be modified to support the decision process is a next required step towards achieving the desired results.

Association studies between the accreditation and QMR systems are required to support the P3 relationship. These studies will involve assessing the amount of association between the accreditation results and the CI reports and the potential for mapping the CIs to the accreditation standards.

The P2 relationship relates to the communication of outcomes of the QMR system to the HCO-level system. Whilst the study results reported in Section 3 identify more positive than negative responses from HCOs when surveyors discuss the CI reports with them, it also identifies scope for improvement. Studies to assess the needs of the HCO regarding the interpretation, understanding, value, timeliness and

perceived usefulness of the existing CI reports for the HCO's own internal control response will help establish the desired P2 relationship.

Studies to assess the value, and potential development, of risk mapping style CI reports that reduce the amount of interpretation and better describe the information at hand will be required.

Finally, the required P2 relationship can only be achieved by the ongoing development of statistical tools to complement existing methods and reports. To this end, continued testing and ultimate introduction of control charts, development of the models underpinning the analyses as well as the introduction of pro-forma for HCOs on how to collect their CI data and appropriate sampling methods for the HCOs to reduce their efforts in collecting their CI data, is continuing.

5. Conclusion

The ACHS CI data is the largest source of data that attempts to measure the quality of care in Australia and New Zealand. The QMR system generating the CI reports has continued to be developed and improved over the past decade.

Given the Bayesian paradigm within which the CI data have been analysed and reported, it is encouraging that there appears to be a parameter space in which the Bayesian-based BBPP control chart detects changes in the underlying proportion more quickly than the CUSUM alternative. It is feasible to consider using a particular chart for a given CI; continued investigation is warranted.

The results of the accreditation surveyor study identified factors affecting the use of the CI reports and their perceived usefulness. The combination of the recommendations and the relatively positive reported use and perceived usefulness of the reports indicates that implementing the control relationship between the QMR and accreditation systems is a promising expectation. There are, however, other key system components, such as the survey method and accreditation standards, which play critical roles in the feedback loops and building the control relationship, warranting further studies.

The future studies into the P1-P4 relationships in the healthcare system model will occur within the setting of the ACHS and provide guidance on policy and improve the health care system and its outcomes. However, the findings in this setting extend to both the *international* health care setting and the international non-health care setting. Industry, government, education, business each has performance measurement and reporting systems and accreditation (both internal and external) processes. The systems theory-based relationships and the conclusions reached in the health care setting can apply and provide guidance, and a platform for future research, in these non-health care settings.

Acknowledgements

Academy of the Social Sciences of Australia and Department of Innovation, Industry, Science and Research for their financial support.
Australian Council on Healthcare Standards for allowing the survey to be conducted and their ongoing support of these improvement projects.

References

[1] Australasian Clinical Indicator Report 2001-2009: Determining the Potential to improve quality of care: 11th Edition.

[2] S.W. Chuang**,** K. Inder, "An Effectiveness Analysis of Healthcare Systems Using Systems Theoretic Approach", *Health Services Research* 2009, 9:195.

[3] P.P. Howley, R.W. Gibberd, "Using hierarchical models to analyse clinical indicators: a comparison of the gamma-Poisson and beta-binomial models", *Int J Qual Health Care,*. 15 (4) 319-329, 2003.

[4] M.P. Pomey, A.P. Contandriopoulos, P. Francois, D. Bertrand, "Accreditation: a Tool for Organizational Change in Hospitals?" *Int J Qual Health Care*, 17:113-24, 2004.

[5] M. Sheahan, "Customer focus: patient, organisation and EQuIP in collaboration", *J Qual Clin Pract*, 19:139-144, 1999

[6] V. Daucourt, P. Michel, "Results of the first 100 accreditation procedures in France", *Int J Qual Health Care*, 15(6):463-471, 2003.

[7] P. Mazmanian, J. Kreutzer, C. Devany, K. Martin, "A survey of accredited and other rehabilitation facilities: education, training and cognitive rehabilitation in brain-injury programmes", *Brain Inj*, 7(4):319-331,1993.

[8] C.H. Fung, Y.W. Lim, S. Mattke, C. Damberg, P.G. Shekelle: "Systematic review: the evidence that publishing patient care performance data improves quality of care", *Ann Intern Med*,148(2):111-123, 2008.

[9] D. Greenfield, G. Braithwaite: "Health sector accreditation research: a systematic review". *Int J Qual Health Care* 2008,20(3):172-183.

Constrained Ordination Analysis in Metagenomics Microbial Diversity Studies

Olivier Thas

Ghent University, Belgium
Olivier.thas@Ugent.be

Yingjie Zhang

Ghent University, Belgium
liv.zhangcn@gmail.com

Abstract

Canonical or constrained correspondence analysis (CCA) is a very popular method for the analysis of species abundance distributions (SAD), particularly when the study objective is to explain differences between the SADs at different sampling sites in terms of local environmental characteristics. These methods have been used successfully for moderately sized studies with at most a few hundred sites and species. Current molecular genomics high throughput sequencing techniques allow estimation of SADs of several tens of thousands of microbial species at each sampling site. A consequence of these deep sequencing results is that the SADs are sparse, in the sense that many microbial species have very small or zero abundances at many sampling sites. Because it is well known that CCA is sensitive to these phenomena, and because CCA depends on restrictive assumptions, there is need for a more appropriate statistical method for this type of metagenomics data. We have developed a constrained ordination technique that can cope with sparse high throughput abundance data. The method is related to the statistical models of Yee (2004, Ecological Monographs, 74(4), pp. 685-701), Zhu et al. (2005, Ecological Modelling, 187, pp. 524-536) and Yee (2006, Ecology, 87(1), pp. 203-213). However, instead of assuming a Poisson model for the abundances, we consider a hurdle model with a truncated Poisson component. The new method is applied to a study on microbial communities in Antarctic lakes. The Roche 454 sequencing technique is used to give SADs of several of thousand microbial species in samples from 50 lakes. The study objective is to estimate the relative importance of environmental lake characteristics and of the geographic coordinates of the lakes in explaining differences between the SADs.

Key words: biodiversity, correspondence analysis, hurdle model, zero-inflation

Computational and Statistical Drug Design: Self-Organising Maps (SOMs) versus Mixtures in Drug Discovery

Irene Hudson

The University of Newcastle, Australia
Irene.Hudson@newcastle.edu.au

Andrew Abell

The University of Adelaide, Australia
Andrew.abell@adelaide.edu.au

Shalem Lee

The University of Adelaide, Australia
Shalem.lee@adelaide.edu.au

Abstract

The principle that chemical structure determines molecular activity has meant that preliminary High Throughput Screening of molecules in search of new drugs has focused on identifying molecules with a similar structure to a known active molecule. High throughput docking of small molecule ligands into high resolution protein structures is now standard in computational approaches to drug discovery, where the receptor structure is kept fixed, while the optimal location and conformation of the ligand is sought via sampling algorithms. A software package called GLIDE is now widely used to screen libraries of ligands (of millions of compounds) and to estimate how well the ligand docks. The aim of this study is to identify superior parameters for the prediction of binding affinity of ligands, and to distinguish between high affinity binders and essential non-binders of calpain in the treatment of cataracts. Using a two-level clustering approach, SOM followed by K-means (KM) clustering, we prove that the structural parameters, namely the number of hydrogen bonds and warhead distance, combined with the 8 subcomponents of GLIDE (or with GLIDE alone), significantly better predict true binding affinity than using GLIDE alone, as a one off score to assess binding affinity in drug design investigations. SOM results are compared to mixture approaches. The mathematical tools developed may be applicable also to the more general and complex challenge of performing docking when scoring and analytically determined activities do not correlate, as is the case for covalent inhibitors.

Key words: Drug discovery, Drug design, hybrid SOMs, mixtures, Cataracts

Index of Authors