SSAI

STATISTICAL SOCIETY OF AUSTRALIA INCORPORATED

NEWSLETTER

September 2009
Number 127

IN THIS ISSUE

Professor Denise Lievesley

Platinum sponsor
CSIRO

Australian Statistical Conference 2010

ASC 2010 is fortunate to have Professor Denise Lievesley, a leading social statistician in the international context as an invited plenary speaker. She is responsible for the use of statistical evidence as the basis for the development of sound public policies internationally and within the UK.

Before her current role as Head of the School of Social Science and Public Policy at King's College London, Denise enjoyed a distinguished career, which included the posts of founding Chief Executive of the English Information Centre for Health and Social Care; Director of Statistics at UNESCO (where she established its new Institute for Statistics), and Director of the UK Data Archive (and simultaneously Professor of Research Methods in the Mathematics Department, University of Essex). Most recently Denise was a special advisor at the African Centre for Statistics of the UN and was based in Addis Ababa.

Professor Lievesley's various roles have led her to work with ministers, ambassadors, senior civil servants and officials of international agencies, from which she has established a reputation for upholding the principles of statistical integrity, policy relevance and methodological transparency. Throughout her working life, Professor Lievesley has been committed to protecting the integrity of official statistics and ensuring that they remain free from political influence.

In the statistical community, her expertise and ability have been recognised with her election as President of the Royal Statistical Society (1999-2001) and of the International Statistical Institute (2007-9), the first ever woman to hold this office. Through these roles she has contributed to the formulation of both national and international policy on both statistics and evidence-based policy, and remains active in the development of social research methods and in research ethics. She has been the international representative on the Board of the American Statistical Association for the last three years and chairs the methodology committee of the European Social Survey.

At a time when in Australia we have governments talking about "evidence based decision making" and the health sector emphasising "evidence based medicine", it is particularly appropriate to have Denise Lievesley in Perth in December next year for ASC 2010.

John Henstridge
Deputy Chair Program Committee
ASC2010

Editorial

Alice Richardson.

This newsletter carries a bumper amount of Branch news, and we hope you enjoy reading about the activities that have taken place both in your Branch and others. Many people have taken on new responsibilities on Branch Councils in the last few months. I'd like to take this opportunity to thanks those people for volunteering their time, and wish them all the best in furthering the activities of the Branches.

I'd also like to welcome our new National President, Geoff Lee, and wish him all the best in his new position in the Society. He has set out am ambitious agenda in his column this issue, and I urge you all to read through it and think about how you can play your part in the plans put forward.

One of the program items for the next few years is a re-invigorated program of workshops around the country. This issue carries information about several up-coming workshops, as well as a report on a recently held workshop. The Editors of the newsletter are always keen to print reports on workshops you attend, so that future participants have a good idea of what to expect when they arrive. Please keep those newsletter contributions coming!

Alice Richardson *Michael Adena*

Alice Richardson
Editor

Michael Adena
Editor

CONFERENCES

New Zealand Statistical Association Conference
1–3 September 2009, Wellington, New Zealand.
http://msor.victoria.ac.nz/Events/NZSA2009

Young Statisticians' Conference 2009
25-26 September 2009, Sydney
http://www.statsoc.org.au/young-statisticians-conference.htm

20th National Conference of the Australian Society for Operations Research
27–30 September 2009, Gold Coast
http://www.asor.org.au/conf2009/index.php?page=1

The Australasian Conference on Statistical Methods for Genomic Data Analysis
5-6 October 2009, Queensland University of Technology, Brisbane QLD
http://genepi.qimr.edu.au/genomics/

Strictly Bayes (Formerly known as Spring Bayes)
26–28 October 2009, North Stradbroke Island, QLD
email bayes.admin@qut.edu.au

25th International Methodology Symposium, Statistics Canada "Longitudinal Surveys: from Design to Analysis".
27-30 October 2009, Gatineau, Qc, Canada
http://www.statcan.ca/english/conferences/symposium2009/.

SSAI and APBG Oncology Workshop
26 November 2009, Melbourne
https://www.statsoc.org.au/CPD4Regs

Australasian Region of the International Biometric Society Conference
29 November – 3 December 2009, Taupo, New Zealand
https://conference.fos.auckland.ac.nz/ibsar/index.html).

Tenth Biennial Islamic Countries Conference on Statistical Sciences (ICCS-X)
20–23 December 2009, The American University in Cairo (AUC), New Cairo, Egypt
http://www.isoss.com.pk/conference/info_conf.php

Statistical Modelling and Inference Conference to celebrate Murray Aitkin's 70th birthday
1–4 February 2010, Brisbane
http://www.aitkinconference.scitech.qut.edu.au/

ISBIS-2010, International Symposium for Business & Industrial Statistics
5–9 July 2010, Grand Hotel Bernardin, St. Bernardin Adriatic Resort & Convention Center, Portoroz (Portorose), Slovenia
www.action-m.com/isbis2010

International Biometrics Conference
5–10 December 2010, Florianopolis, Brazil
http://www.tibs.org/Interior.aspx

Australian Statistical Conference 2010
6 – 10 December 2010, Perth, WA
http://www.promaco.com.au/2010/asc/index.htm

58th Session of the International Statistical Institute
21-26 August 2011, Dublin, Ireland
http://www.isi2011.ie/

SSAI

OFFICE MOVE

The SSAI Office has moved!
Our new address is
PO Box 213, Belconnen ACT 2616
We are located on the ground floor of ABS House.
Phone 02 6251 3647
Fax 02 6251 0204
Email eo@statsoc.org.au
Website www.statsoc.org.au

EDITORS

Alice Richardson
School of ISE,
University of Canberra

Michael Adena
Covance

CORRESPONDENCE

Please direct all editorial correspondence to Alice Richardson
Email newsletter@statsoc.org.au

DISCLAIMER

The views of contributors to this Newsletter should not be attributed to the Statistical Society of Australia, Inc.

SUBSCRIPTIONS

The Newsletter of the Statistical Society of Australia is supplied free to all members of the society. Any others wishing to subscribe to the newsletter may do so at an annual cost of A$30.00 for an issue of four numbers.

ADVERTISING

Advertising will be carried in the Newsletter on any matters which the Editors feel are of interest to the members of the Society. For details of advertising rates, etc.

Contact the SSAI Executive Officer at newsletter@statsoc.org.au

DEADLINE FOR NEXT ISSUE:
10 November 2009

President's Message

Dear Members

This is my first newsletter column as incoming President of the SSAI, having formally taken over from William Dunsmuir at the conclusion of the Annual General Meeting of the SSAI Executive Council. I feel very honoured to have this role. It is traditional to thank one's predecessor for what they have achieved, but I won't do that for tradition's sake, I will do it because William has well and truly earned my thanks, indeed the thanks of the whole society, for the effort he has put in and the challenges that he has guided our Society through.

In my role as Vice President last year I saw the enormous amount of work and energy William put in. What I didn't fully appreciate until the close of the AGM, when it settled on my shoulders, was the sense of responsibility that goes with the position. So, William please accept these public thanks from me, on behalf of all of the Central Council, indeed all the society, for the leadership you have given us through what have been challenging times.

I have been tremendously heartened by the responses from branches, and individual members of the society, to the financial challenges we face. The Central Council has agreed on a plan extending over the next 4 years which involves:

- very careful management of our ongoing expenditure and income;
- arrangements with the individual branches which should enable us to resolve the situation with our creditor from the ASC 2008 within the next few months;
- a program of fund raising workshops in each state;
- a reinvigoration of our membership drive; and
- agreement to develop an institutional membership program.

These initiatives won't just happen on their own, we will all have to work at them if our Society is to survive and prosper in the future. However, I am convinced we can do this. Members were consulted in the last few months about the Central Council plans, and the responses were overwhelmingly in favour of accepting and implementing them. Many members have gone beyond simply approving the plans. A number have made quite generous personal donations to the society; others have shown their faith in the future of the society by renewing their membership for up to five years in advance; while others have volunteered their time and effort, either to present workshops or to help with other aspects of implementing the plans. It is because of this level of grass roots practical support that I am confident our Society will survive, and, dare I say it, even be strengthened by the challenges we face.

There is plenty more to do, and I encourage all members to consider how they might help and become involved in the Society either on these "national" initiatives, or in local branch activities. Everyone's circumstances are different, and that will guide what you can do, but I can personally attest to the satisfaction that comes from becoming actively involved in the activities of our Society. At the end of the day, our financial position is merely an enabler to our Society's main objective which is to further the study, application and good practice of statistical theory and methods, in all branches of learning and enterprise. The true measure of our success is how much enjoyment and success we derive from striving towards that lofty goal.

That's plenty of exhortation from the "pulpit", on to other issues. This is an appropriate time for me to ponder out loud about what the Society might achieve over the next few years. First, and obviously, a key focus will have to be our managing our financial situation. However, even while doing that, I would like at the same time to be progressing the Society's broader objectives. Taking as an example the program of fund raising workshops, this will be an excellent base to expand into a wider program of continuing professional development, with its links to our accreditation program. Likewise the institutional membership program, when it goes ahead, will be an opportunity to promote the accreditation program, this time to employers.

Second, I would like to see us maintain a strong program of professional conferences. Planning for the ASC in Fremantle in December 2010 is already underway, and as it draws closer, will become an increasingly important focus for our attention. The selection of keynote speakers, and the main program, is already shaping up well,

>>

SSAI EXECUTIVE COMMITTEE

Central Council:
President: Geoff Lee
Secretary: Dr Doug Shaw
secretary@statsoc.org.au

BRANCH PRESIDENTS AND BRANCH SECRETARIES

Canberra
President: Veronica Boero-Rodriguez
Secretary: Dr Ray Lindsay
secretary.actbranch@statsoc.org.au

New South Wales
President: Dr Eric Beh
Secretary: Dr Boris Choy
Boris.choy@uts.edu.au

Queensland
President: Dr Miranda Mortlock
Secretary: Helen Johnson
h.johnson@qut.edu.au

South Australia
President: Paul Sutcliffe
Secretary: Janine De Korte
janine.dekorte@abs.gov.au

Victoria
President: Dr Ian Gordon
Secretary: Dr Sue Finch
sfinch@unimelb.edu.au

Western Australia
President: Marty Firth
Secretary: Rebecca O'Leary
roleary@ichr.uwa.edu.au

SECTION CHAIRS

Bayesian Statistics
Scott Sisson
Scott.sisson@unsw.edu.au

Statistical Education (co-chairs)
Michael Martin
Michael.martin@anu.edu.au

Peter Howley
Peter.howley@newcastle.edu.au

Surveys and Management
Veronica Rodriguez
veronica.rodriguez@abare.gov.au

Young Statisticians (co-chairs)
Kevin Wang
wang_ko-kang@lilly.com

Further contact details for Society Secretaries and Section Chairs can be obtained by contacting the Society on (02) 6251 3647

Report
Liz Yuncken

Statisticians: meet your fans

President's Message cont.

and I would like to see us surround that with a rich program of satellite workshops and development courses, so that attending our conference can be a really enjoyable and attractive professional experience.

Our Young Statisticians group is active and vibrant, and have a conference planned in Sydney in September 2009. It's looking to be a fun and rewarding experience, open to both Young Statisticians and the young at heart, who may want to attend to share their experiences and wisdom. If you happen to know of a colleague or student who is a young statistician (or indeed you are one yourself), you might like to draw this opportunity to network to their attention and suggest they consider attending.

Third, I am personally interested in the long term sustainability of the statistical profession in Australia. I am thinking broadly here, with a definition of statistics closer to a "data science" which has many applications in scientific (and indeed commercial) endeavours, not just that important core "mathematical statistical" part of our profession. And I am thinking not just about the university education and research sector components, but also about the feeder groups in secondary school education, career paths after formal study, and even the societal forces which are driving the career decisions of children, young adults, and their parents. This is not a new topic, indeed over the last few years there have been several major reviews in this area. Nor is it a topic where we need necessarily share the same views about the problem (if there is one - I think there is), nor the solutions. But I am sure this is a topic which should be at the centre of our Statistical Society's list of important issues to address, and I look forward to being able to work with all the members of the Society to understand, lobby and act in this space.

Finally, to close this message I should say how humbled I am by the responsibility you have entrusted to me. I know what a wide range of interests, applications and experiences the members of our Society span. I've got a pretty good appreciation of how statistical methodology is used in Official Statistics, but that leaves many vitally important areas of our profession where I have only superficial knowledge of the real issues, challenges and opportunities. I look forward to serving as President of our Society, and using this opportunity to meet many of you, and to learn more about issues affecting statistics in academia, in research, and in commerce. Please feel free to contact me, and provide advice and suggestions.

Geoff Lee

It's time for mathematicians and statisticians to step forward and meet their fans while helping promote maths education in schools.

Mathematicians in Schools, a sub-program of the highly successful Scientists in Schools run by CSIRO Education, connects teachers with mathematicians and other professionals who use maths as a major component of their work.

The ongoing, professional partnerships engage students with real-life maths, support primary and high-school teachers and inspire mathematicians to see their work from a new perspective.

Over 1000 scientists from around Australia have already been linked with schools and treated like celebrities by the students they visit. Now it's the turn of the mathematicians!

Dr Louise Ryan, Chief of CSIRO Mathematical and Information Sciences, is the patron of Mathematicians in Schools and is excited by its potential.

"Mathematics underpins so many aspects of our lives and provides the foundation for science, engineering, economics, finance and many other professional fields. For all these reasons, it is critical to help students early on to appreciate the beauty and power of mathematics and to begin to understand how it plays out in so many areas."

Independent evaluation and anecdotal evidence from Scientists in Schools shows that students are inspired by their interaction with 'real scientists', resulting in a higher regard for science and a keener interest in science careers. And teachers' confidence in teaching science has markedly improved as a result of the support from a scientist, with flow-on effects to students.

Even the scientists have seen significant benefits from their involvement with frequent reports of the enjoyment of interacting with students and resulting re-engagement with their own work.

Mathematicians in Schools is new, but there's already a waiting list of maths teachers eager to interact with maths professionals. Why not link up with a teacher in your area to inspire the next generation?

Dr Ryan encourages mathematicians of all persuasions to engage with their local communities to share their knowledge of and passion for mathematics and its various subspecialties and related fields.

For details, and to register, visit www.mathematiciansinschools.edu.au

Mathematicians in Schools is a sub-program of Scientists in Schools and is an Australian Government initiative.

Liz Yuncken Senior Project Officer Scientists in Schools, CSIRO Education.

Delights of Fremantle

Australian Statistical Conference 2010

Auatralian Statistical Conference 2010 cont.

ASC2010 will be held in Fremantle which is located at the mouth of the Swan River on the West Coast of Australia and is part of the Perth metropolitan area. Perth city is only 20 kms away which is a 30 minute trip by train. Fremantle is a vibrant, eclectic city. It is home to some of Western Australia's most beautiful heritage buildings, great beaches, delicious seafood and talented artists and designers. Some of the most popular tourist sites are the WA Maritime Museum, Fremantle Prison, Fremantle Fishing Boat Harbour, the Cappuccino Strip and the historic Fremantle Markets. Alternatively you can catch a ferry to Rottnest Island or just relax and watch the sunset on South beach.

Boat Harbour WA.

Fishing Boat Harbour

Several hundred Italian fishermen came to Western Australia after the Eastern Goldfield discoveries in 1892 and 1893. In 1904, a Fremantle Fisheries Inspector reported that "The number of licenses issued has been 354 men and 124 boats, about 10 boats being engaged chiefly in the cray fishing. I regret to say that the industry is chiefly in the hands of the Greeks and Italians, both afloat and ashore. These prefer the small class of boats and work together in batches and send their fish to market by the same boat".

Today, the working harbour, which is located just across the road from the conference venue, is surrounded by fishing boats, yachts, cafes, restaurants and even a boutique brewery. The fishing boat harbour was the centre of activity during the America's Cup and is a popular place for locals and tourists. There is a fishing boat harbour walking trail which allows you to experience the working life and maritime history of Fremantle's waterfront. See where history was recreated with the building of the Endeavour and Duyfken replicas.

Cycling on Rottnest Island WA.

Rottnest Island

Rottnest Island is Western Australia's premier island getaway, located just off the coast of Perth. As a 'Class A' reserve, managed by the Rottnest Island Authority, the island boasts unique animal and plant life, as well as some of the most spectacular beaches you will ever see. The reefs provide divers with an abundance of beautiful corals and shipwreck to explore. There are no cars on the island so visitors have a chance to tour around the island by bicycle in a leisurely manner.

Fremantle Markets

These ever popular Markets, located within walking distance of the Esplanade Hotel, are open Friday to Sunday and Monday Public Holidays. Housed in a grand old Victorian building, over 150 stalls offer fresh food, fashion, crafts and gifts along with live entertainment from talented buskers.

Cappucino Strip – Fremantle WA.

Cappucino Strip

A visit to Fremantle would not be complete without having a coffee or meal on South Terrace, otherwise known as the Cappuccino Strip. This cosmopolitan area houses many coffee houses, restaurants and pubs where tourists and locals, including students from Fremantle's Notre Dame University, congregate to enjoy themselves, day or night, anytime of the week. The Sail and Anchor, Australia's first pub and microbrewery, is also on South Terrace.

Jane Speijers – Chair of the Organizing Committee ASC2010

Accredited Members

Meet our Accredited Members

Darssan Balasingam.

Dr Grgory Fant.

Manjula Schou.

Darssan Balasingam, St.George Clinical School, UNSW, Kogarah

Accredited since 2009 (GStat)

Statistical fields of Interest: Bayesian Statistics, Biostatistics, Clinical Trial Design

Darssan got his Bachelor of Science with Honours in Statistics in 2005 at the University of Jaffna, Sri Lanka. He has completed M.APP.Sc (Statistics and Operations Research) at the RMIT University in 2008.

Darssan is currently working as a Statistician at the University of NSW. He is involved with the analysis of the PAET (Patient-centered Adverse Event Tool) with an aim to develop a tool that is both reliable and responsive when used to measure adverse events associated with new or alternative therapeutic treatments. He is also working with the analysis of OSRA (Objective study in Rheumatoid Arthritis) project. This is looking at developing an 'evidenced-based' measure of inflammation in Rheumatoid Arthritis patients by correlating a measure of disease progression such as radiograph of the MCP joints to a biological surrogate. The OSRA is a randomized controlled blinded trial, designed to test the validity of surrogate markers rather than efficacy of drugs.

Dr Gregory Fant, US Department of Health and Human Services, Rockville, Maryland, USA

Accredited since 2009

Statistical fields of Interest: Public Health Statistics, applied statistics and research methods, statistical computing using SPSS, statistical education.

Dr. Fant is a Supervisory Health Scientist assigned to the Division of Science and Policy (DSP), HIV/AIDS Bureau (HAB), Health Resources and Services Administration (HRSA), U.S. Department of Health and Human Services (DHHS). He serves as the Associate Director, DSP, and assists the DSP Director in the day-to-day operations of the division including planning and assisting in the development of science and data policies supporting HRSA/HAB. Before his current assignment, Gregory was a health statistician in charge of the Scientific Registry of Transplant Recipients contract and the Executive Secretary of the DHHS Secretary's Advisory Committee on Organ Transplantation. Since joining the Federal Civil Service in 1997, Gregory has served as a Federal statistician in the Epidemiology and Data Analysis Branch, Office of Science and Epidemiology, HIV/AIDS Bureau; Department of Clinical Investigation, Walter Reed Army Medical Center, U.S. Department of Defense; and the Bureau of the Census, U.S. Department of Commerce. Dr. Fant earned his doctorate from the University of Nebraska; he is an elected member of the American College of Epidemiology.

Gregory completed an applied space science/ technology, leadership development, and team-building course at the U.S. Space and Rocket Center in Huntsville, Alabama, and earned his "Adult Advanced Space Academy Wings" in September 2008.

Favorite quote:

"An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question."

— *The first golden rule of mathematics, sometimes attributed to John Tukey.*

Manjula Schou

Accredited since 2009 (AStat)

Statistical fields of Interest: Biostatistics, Particular Focus on Clinical Trials/ Methodology

Manjula has worked as a statistician for over 15 years and has more than 12 years of experience working on pharmaceutical industry sponsored and investigator initiated clinical trials. Manjula was most recently with Pfizer Pharmaceuticals and is in the process of enrolling in a PhD at Macquarie University.

Hansoo Kim, GSK Australia

Accredited since 2008 (AStat)

Statistical Fields of Interest: Outcomes Research Statistics, Clinical and Biostatistics, Health Economic Modelling

Hansoo has degrees in mathematical and theoretical statistics from the University of Copenhagen, Denmark. He worked as a clinical biostatistician in Denmark and Singapore before coming to Australia. Here, Hansoo worked initially as an outcomes research statistician for Pfizer in Sydney from 2006 to 2009 and is currently working at GSK in Melbourne.

Show me the model! An exchange on ANZSTAT Feb 2009

$Y = X\beta + E$, where $E \sim N(0, \sigma^2 I)$

ANZSTAT Exchange

Alice Richardson

Edited by the participants and Alice Richardson for SSA newsletter Aug 2009

Patrick Cordue: I asked a question on GLMs a couple of days ago. In essence I was asking "what is the model — please write it down — you, know, like for a linear model: Y = a + bx + e, where e ~N(0,s^2) — can't we do that for a GLM?"

I come from a modelling background where the first step is to "write down the model"; the second step is to look for tools which will provide estimates of the unknown parameters; (I am assuming we already have a data set). If my model is a GLM, then I can just use glm() in R. So, I wanted to know the form of the GLM models for different families and link functions.

In particular, which implied simple additive errors ($Y = \mu + \varepsilon$) and which implied simple multiplicative errors ($Y = \mu * \varepsilon$)? (where $\mu = E(Y)$)

Bill Venables: This is a topic that has arisen elsewhere many times and not a few explanations have been faulty in the past (in particular in the first edition of the GLIM manual, no less!).

There has been a lot of work done on model checking, diagnostics and things like "goodness of link" tests, of course, (some would say too much). Good old Google should deliver a swathe of stuff, as usual (again, probably too much!).

Residuals, when properly defined, usually provide a good tool for this, too. However no single definition of residuals is universally optimal, and none really guarantees that you can think of them as "additive" or "multiplicative". They are simply quantities derived from the observations and the fitted model which, under the assumption that the model is correct, should behave as approximately iid Gaussian, and this in turn provides a check on the original assumptions - of sorts.

For example, there are "deviance residuals", which have the property that their sum of squares is the deviance. These should be approximately iid Gaussian (irrespective of the original distribution) if the modelling assumptions are correct. In turn this can be checked by normal scores plots and other devices. For some purposes "Pearson" residuals are to be preferred for use in diagnostic devices. These have the property that their sum of squares is the Pearson Chi-squared test of goodness of fit for the model. And so on. There are four commonly used definitions, in fact, which all become the standard definition for the case of normal linear models. Take your pick, really.

One thing to point out pretty strongly is that in the case of binary data, a pretty common kind of response these days, no definition of residuals has much use. Checking model assumptions there can be a tricky affair and lots of pretty unusual suggestions have been made at times for how to do it...

Modellers often like to think of the model as consisting of a fixed part, essentially deterministic, perturbed by errors. In this they are comforted and encouraged by the traditional way linear models are presented, i.e.

$Y = X\beta + E$, where $E \sim N(0, \sigma^2 I)$

This additive expression of how 'errors' come into the picture is not always possible, however. An alternative way of expressing this model is

$Y \sim N(X\beta, \sigma^2 I)$

where the stochasticity is embedded in the "N" part, without explicit reference to additivity. In this way the extension to, e.g. Poisson loglinear models, is straightforward:

$Y \sim Po(\exp(X\beta))$

where the stochasticity is now expressed by the "Po", but this time ther is no going backwards to an additive expression. The "errors" are not explicitly available. This is still a fully specified model, however, and you can define errors in any way you see fit, if that is your wish. e.g. $e = y - \exp(X\beta)$, $e = \log(y) - X\beta$ (bit embarrassing if y = 0, though!) etc. However don't ask for the distribution of these errors, because it is just not useful to do so. Rather deal with the model as expressed in the distributional form above. This gives you all you need.

The generalization to generalized linear models is then straightforward

$Y \sim f(\mathrm{invlink}(X\beta); \theta)$

where f describes the distributional family, X beta is the linear predictor, invlink is the inverse of the link function and theta is any additional parameter needed, like σ^2.

Why the "inverse link" is used here rather than just the link function is an accident of history, but an instructive one. Before GLMs the tradition was to transform the response to a scale in which something like an additive linear model did apply. The achievement of GLMs is really to switch the transformation to the other side and apply it to the mean of the y instead, rather than to y itself. This is then the inverse link function, with the link function itself still essentially the transformation that would have been historically applied to the response. Thus in the above Poisson example the link is the log function, but the inverse link is the exponential. This way of doing things has many advantages, but one obvious one is that it completely avoids any problem with zero observations, which were a great embarrassment in the old transformation days and there was a thriving cottage industry in how to deal with them. In fact the idea of a GLM grew out of a throw-away remark of Fisher when in 1935 Bliss pushed him for a solution to "the case of zero survivors" in probit analysis.

See http://digital.library.adelaide.edu.au/dspace/bitstream/2440/15223/1/126.pdf

This tiny appendix is the birthplace of generalized linear modelling. It all grew from that.

Murray Jorgensen: In GLMs there is no simple characterisation of how the systematic and random parts of the model combine to give you the data (other than the definition of the GLM, of course).

One of the achievements of modern statistical computing is the ability to express a wide variety of models with just a simple statement in the language of a statistical package. Patrick is right to be concerned to 'unpack' the statement and know just exactly what it implies about the distribution of the responses given the parameters and the values of the covariates.

Too many students seem content to just be able to manipulate model specification statements without any clear idea of what they mean. To me a model should be a prescription of how to generate new data conforming to the model. If you can't program that, you don't understand the model.

ANZSTAT Exchange

Alice Richardson

Bill Venables has pointed out the historical connection between links and transformations which is why links are defined as they are. I do think, though, that the use of the link function rather than the more natural inverse link function in a GLM plays a part in making the model more mysterious.

Apart from a number of regrettable typos I am a great fan of Gelman and Hill's book "Data Analysis Using Regression and Multilevel/ Hierarchical Models". They go to extreme lengths with some of their examples spelling out the gory details of the models in about three equivalent ways. But this sort of thing has to be done somewhere or users do not have a good enough understanding of just what they are fitting.

Sama Low Choy: This debate has been very illuminating. Thank you Patrick for posing the question.

Bill's seemingly obvious comments are quite powerful - on writing down the model so that the mean and variance are embedded within the key model statement about the assumed distribution, including the (inverse) "link" to covariates. Within a Bayesian context, it is perhaps even more important to be aware of this way of "writing" down the model, as implied by Murray's reference to Gelman & Hill's book.

Thought I would just add a few of comments/ queries that may be of interest from a "pedagogical" standpoint (blame John Marriott for explaining this word to us recently!)

ERRORS AND LEAST SQUARES

1. The way of writing down the Normal model with additive errors seems to have been motivated by early least squares (including partial LS, weighted LS) approaches to estimation. To do LS, you need the errors explictly $\varepsilon_i = Y_i - X_i*\beta$, so that you can minimize the LS criterion on the errors min $\sum \varepsilon_i^2$.

When undertaking estimation based on the likelihood (e.g. maximum likelihood), the distributional way of writing down the model becomes more useful. This includes Bayesian approaches where the likelihood is required - either for a computational approach (e.g. in WinBUGS where GLMs can be specified simply by specifying the likelihood, priors, and link to covariates) or for an analytic approach where the marginal posterior distributions of parameters are written down explicitly.

ERRORS AND PARAMETERIZATION

2. The parameterization of the model does not always lead to clear separation of errors (i.e. to provide an additive or multiplicative expression), let alone permit separation of the mean from the variance. I encourage you to investigate alternative parameterisations of GLMs or other non-Gaussian models where the covariates are related to the distribution of the response (e.g. Beta regression).

For example, a Beta regression may be parameterized so that the covariates are linked to the mean or even perhaps the mode; the remaining parameter (related to *both* the variance and the mean) can be expressed as the variance, the "effective sample size" or as correlation.

MODEL FORMULATION FOR ESTIMATION

3. More generally, the "way of writing down the model" can be very closely linked to the method of estimation; this seems to be true for a range of statistical models beyond GLMs. For example, consider the suite of clustering techniques (see nice exposition in Chambers, Tibshirani and Friedman's Statistical Learning text). In most cases the "model" is written down in a form that is suited to the usual method of estimation. See Christian Robert's recent efforts to re-express the Nearest Neighbour clustering technique to a particular form of Finite Mixture Models.

Another example is classification trees. To explain, the Recursive Partitioning Algorithm implemented in RPART (an R library by Atkinson & Therneau 1997) the model is written down in a way that is linked to the estimation method. However for a Bayesian estimation approach, the full likelihood model is required - see O'Leary et al (2008; Journal of Applied Statistics, 3(1)) with citation of seminal references Chipman, George & McCulloch (1998) and Denison, Mallick & Smith (1998).

ESTIMATION VS MODEL

4. When working with ecologists, it has often been confusing to me that sometimes the model is quoted via the estimation method. One example is the terminology used when referring to alternatives to logistic regression when modelling habitat. See Elith et al (2006; Ecography, 29(2): 129-151). Here the "maximum entropy" approach is mentioned which seems to refer to moment-matching on the mean and variance by applying the ubiquitous entropy model from statistical physics.

Geoff Jones: I've been waiting for someone to say something about quasilikelihood and generalized estimating equations. Let me try.

Isn't a GEE a weighted sum of the additive "residuals", Y - Model? So this way of looking at the model is actually quite useful.

Bill Venables: Y - Model is one of the four definitions of residual in a GLM. It's usually called the 'raw' residual, or something similar.

Of course this is useful. But if you are trying to understanding the genesis and theoretical basis of GLMs I think it is a bit of a red herring, at least at the start.

Murray H. Smith: In most GLMs the error is neither multiplicative nor additive. Parameterize the 1-parameter error family by the mean (fixing any dispersion or shape parameters, which is what pure GLM is with the added constraint that the error distribution belongs to a 1-parameter exponential family).

We can only write

$$y \sim \mu + \varepsilon \quad \text{or } y \sim \mu*\varepsilon$$

for e not depending on mu, if mu is a location or scale parameter for the error family. i.e.

$$y \sim f(\, y;\, \mu) \text{ where } f(y;\, \mu) = f(y - \mu;\, \mu{=}0) \text{ or}$$

$$y \sim f(\, y;\, \mu) \text{ where } f(y;\, \mu) = 1/\mu * f(y/\mu;\, \mu = 1)$$

The variance function $V(\mu)$, the variance expressed as a function of the mean, must be constant for an additive error and proportional to μ^2 for multiplicative.

Patrick Cordue: Clearly for discrete distributions, it makes no sense to look for a "building block" error e which can be added/ multiplied to/by the expectation to provide the response variable. My question was aimed at continuous distributions.

I deduced the following results from first principles:

For the Gaussian family, $Y = \mu + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$ (and E(Y) = mu =

m(eta) where eta is the linear combination of the explanatory/stimulus variables, and μ^{-1} is the link function) is a GLM. I take this to imply that when one fits a model using glm() with a Gaussian family and any link, that the implied error structure is additive.

For the Gamma family, Y = mu * e where e ~ Gamma(k, 1/k) is a GLM. I take this to imply that when one fits a model using glm() with a Gamma family and any link, that the implied error structure is multiplicative.

For the inverse Gaussian family the implied model does not have a simple additive or multiplicative error structure (someone might know how to write down the model in this case - but not me).

Thanks to everyone who provided comments and references

News on the GEE short course

Deakin University, Burwood Campus, Melbourne

Short course attendees.

A short course on "Longitudinal Data Analysis using GEE and QIC methods" was held at the Deakin University, Burwood campus, Melbourne, on the 5th June 2009. In this course, Dr Jisheng Cui introduced the advantages of longitudinal studies compared with cross-sectional studies in estimating the "cohort" and "age" effects separately. He also introduced the graphical display utilities of longitudinal data, statistical analysis methods using summary measures and Generalized Estimating Equation (GEE) methods in the analysis of normal, binomial and Poisson distribution data. Finally, Dr Cui described his recent research results on the selection of the best correlation structure and best GEE model using the quasi-likelihood under the independence model criterion (QIC) method. He gave several examples to demonstrate how to use the QIC program to calculate the QIC values to select the best correlation structure and best model for normal, binomial and negative binomial distribution data.

A total of 24 people attended this short course on a sunny Friday in Melbourne. Not only local people from the University of Melbourne, Monash University, La Trobe University, Royal Melbourne Institute of Technology (RMIT), Ballarat University, Deakin University, Cancer Council Victoria and Department of Primary Industries attended this short course, but also people from overseas and interstates, including New Zealand, Fiji, New South Wales, South Australia and Queensland. Each participant received a lecture note of 74 pages, which includes detailed examples and exercises to conduct the GEE and QIC analyses. Each exercise is accompanied with an answer and a corresponding illustrative data set. Most participants reflected that this short course is useful for analysing their own longitudinal data. Attached is a photo of the participants of the short course.

Dr Jisheng Cui

APBG
Australian Pharmaceutical
Biostatistics Group

Challenges of running oncology clinical trials

The APBG and the Statistical Society of Australia Inc are pleased to present the following workshop:

Thursday 26 November 2009

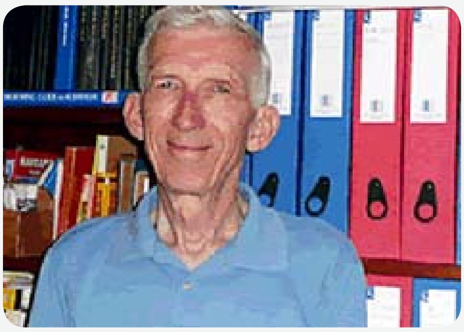
9.00am-5.00pm

Melbourne

If you are a Clinical Research Associate, a data managers, a statistician, a medical writer or researcher who undertakes clinical trials in oncology, this workshop is for you.

You will have the opportunity to share your experiences and challenges of working on clinical trials in oncology. The workshop aims to take you through some of the stages during the execution of a clinical trial in oncology. It will start with discussions around developing concept outlines and protocols for oncology studies, the designs of oncology clinical trials and the subsequent logistics from a statistical, monitoring and data management perspective. From there the focus will move to safety reporting and the use of data safety monitoring boards, the differences between MedDRA and CTC coding dictionaries and why you would use one rather than the other followed by how the TGA reviews oncology studies. After lunch the discussions will become more specific around the types of measures and the corresponding analyses and the benefits and pitfalls of using the RECIST criteria. The day will finish with some ideas on how to undertake a critical appraisal of the oncology literature and a look into the future promises and challenges of oncology studies with the exciting novel therapies. For more information and to see the detailed program, please go to:

https://www.statsoc.org.au/CPD4Regs.



Murray Aitkin.

Statistical Modelling and Inference Conference

to celebrate Murray Aitkin's 70th birthday

1 - 4 February 2010

QUT Gardens Point Campus, Brisbane, Queensland, Australia

This conference celebrates Murray Aitkin's 70th birthday and is also a celebration of statistical modelling and inference: the current and future state, as well as Murray's involvement and contributions to these fields. To find out more, please go to

http://www.aitkinconference.scitech.qut.edu.au/

Use of third and fourth order inclusion probabilities in survey sampling

Sarjinder Singh
Department of Mathematics
Texas A&M University-Kingsville
Kingsville, TX 78363, USA

E-mail: **sarjinder@yahoo.com**

ABSTRACT

In this paper, students doing sample survey courses can learn how to find the variance of the estimator of variance that makes use of third and fourth order inclusion probabilities. To my knowledge, in survey sampling, there is no book including Cochran (1977), Lohr (1999) and Singh (2003) which explains to the students to construct and use of the third and fourth order inclusion probabilities. Thus many students and teachers can get benefit from the exercise given below while dealing the problem of estimation of variance using unequal probability sampling.

Key words: Variance, Bias, Inclusion Probabilities.

PROBLEM AND SOLUTION

Problem: In a circus show the rating of performances of five animals: *Antelope, Bull, Camel, Deer and Elephant* have their respective scores of 12, 14, 10, 18 and 22. We wish to select a sample of $n = 4$ animals out of $N = 5$ animals with the following sampling scheme:

$s_1 = \{\text{Antelope, Bull, Camel, Deer}\}$, $s_2 = \{\text{Antelope, Bull, Camel, Elephant}\}$, $s_3 = \{\text{Antelope, Bull, Deer, Elephant}\}$, $s_4 = \{\text{Antelope, Camel, Deer, Elephant}\}$, and $s_5 = \{\text{Bull, Camel, Deer, Elephant}\}$,

such that

$P(s_1) = 0.2$, $P(s_2) = 0.1$, $P(s_3) = 0.3$, $P(s_4) = 0.2$, and $P(s_5) = 0.2$.

(a) Construct the first order inclusion probabilities.

(b) Construct the second order inclusion probabilities.

(c) Construct the third order inclusion probabilities.

(d) Construct the fourth order inclusion probabilities.

(e) Estimate the variance of the HT-estimator using the Sen-Yates-Grundy estimator of variance from each sample.

(f) Find the true variance from all the units in the population using Sen-Yates-Grundy form.

(g) Show that the Sen-Yates-Grundy's estimator remains unbiased.

(h) Find the variance of the estimator of the Sen-Yates-Grundy estimator by using the definition of variance.

(i) Use the third and fourth order inclusion probabilities to find the variance of the estimator of variance due to Sen-Yates-Grundy formula.

Solution. (a) First order inclusion probabilities

$$\pi_1 = P(s_1)+P(s_2)+P(s_3)+P(s_4) = 0.2+0.1+0.3+0.2 = 0.8$$
$$\pi_2 = P(s_1)+P(s_2)+P(s_3)+P(s_5) = 0.2+0.1+0.3+0.2 = 0.8$$
$$\pi_3 = P(s_1)+P(s_2)+P(s_4)+P(s_5) = 0.2+0.1+0.2+0.2 = 0.7$$
$$\pi_4 = P(s_1)+P(s_3)+P(s_4)+P(s_5) = 0.2+0.3+0.2+0.2 = 0.9$$
$$\pi_5 = P(s_2)+P(s_3)+P(s_4)+P(s_5) = 0.1+0.3+0.2+0.2 = 0.8$$

(b) Second order inclusion probabilities

$$\pi_{12} = P(s_1)+P(s_2)+P(s_3) = 0.2+0.1+0.3 = 0.6$$
$$\pi_{13} = P(s_1)+P(s_2)+P(s_4) = 0.2+0.1+0.2 = 0.5$$
$$\pi_{14} = P(s_1)+P(s_3)+P(s_4) = 0.2+0.3+0.2 = 0.7$$
$$\pi_{15} = P(s_2)+P(s_3)+P(s_4) = 0.1+0.3+0.2 = 0.6$$
$$\pi_{23} = P(s_1)+P(s_2)+P(s_5) = 0.2+0.1+0.2 = 0.5$$
$$\pi_{24} = P(s_1)+P(s_3)+P(s_5) = 0.2+0.3+0.2 = 0.7$$
$$\pi_{25} = P(s_2)+P(s_3)+P(s_5) = 0.1+0.3+0.2 = 0.6$$
$$\pi_{34} = P(s_1)+P(s_4)+P(s_5) = 0.2+0.2+0.2 = 0.6$$
$$\pi_{35} = P(s_2)+P(s_4)+P(s_5) = 0.1+0.2+0.2 = 0.5$$
$$\pi_{45} = P(s_3)+P(s_4)+P(s_5) = 0.3+0.2+0.2 = 0.7$$

(c) Third order inclusion probabilities

$$\pi_{123} = P(s_1)+P(s_2) = 0.2+0.1 = 0.3$$
$$\pi_{124} = P(s_1)+P(s_3) = 0.2+0.3 = 0.5$$
$$\pi_{125} = P(s_2)+P(s_3) = 0.1+0.3 = 0.4$$
$$\pi_{134} = P(s_1)+P(s_4) = 0.2+0.2 = 0.4$$
$$\pi_{135} = P(s_2)+P(s_4) = 0.1+0.2 = 0.3$$
$$\pi_{145} = P(s_3)+P(s_4) = 0.3+0.2 = 0.5$$
$$\pi_{234} = P(s_1)+P(s_5) = 0.2+0.2 = 0.4$$
$$\pi_{235} = P(s_2)+P(s_5) = 0.1+0.2 = 0.3$$
$$\pi_{245} = P(s_3)+P(s_5) = 0.3+0.2 = 0.5$$
$$\pi_{345} = P(s_4)+P(s_5) = 0.2+0.2 = 0.4$$

(d) Fourth order inclusion probabilities

$$\pi_{1234} = P(s_1) = 0.2$$
$$\pi_{1235} = P(s_2) = 0.1$$
$$\pi_{1245} = P(s_3) = 0.3$$
$$\pi_{1345} = P(s_4) = 0.2$$
$$\pi_{2345} = P(s_5) = 0.2$$

Let $\Theta_{ij} = (\pi_i \pi_j - \pi_{ij})$, $D_{ij} = \Theta_{ij}/\pi_{ij}$, $\gamma_{ij}^2 = \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$,

$\Theta_{12} = (\pi_1\pi_2 - \pi_{12}) = (0.8 \times 0.8 - 0.6) = 0.04$; $\Theta_{13} = (\pi_1\pi_3 - \pi_{13}) = (0.8 \times 0.7 - 0.5) = 0.06$

$\Theta_{14} = (\pi_1\pi_4 - \pi_{14}) = (0.8 \times 0.9 - 0.7) = 0.02$; $\Theta_{15} = (\pi_1\pi_5 - \pi_{15}) = (0.8 \times 0.8 - 0.6) = 0.04$

$\Theta_{23} = (\pi_2\pi_3 - \pi_{23}) = (0.8 \times 0.7 - 0.5) = 0.06$; $\Theta_{24} = (\pi_2\pi_4 - \pi_{24}) = (0.8 \times 0.9 - 0.7) = 0.02$

$\Theta_{25} = (\pi_2\pi_5 - \pi_{25}) = (0.8 \times 0.8 - 0.6) = 0.04$; $\Theta_{34} = (\pi_3\pi_4 - \pi_{34}) = (0.7 \times 0.9 - 0.6) = 0.03$

$\Theta_{35} = (\pi_3\pi_5 - \pi_{35}) = (0.7 \times 0.8 - 0.5) = 0.06$; $\Theta_{45} = (\pi_4\pi_5 - \pi_{45}) = (0.9 \times 0.8 - 0.7) = 0.02$

$D_{12} = \Theta_{12}/\pi_{12} = 0.04/0.6 = 0.06667$; $D_{13} = \Theta_{13}/\pi_{13} = 0.06/0.5 = 0.12$;

$D_{14} = \Theta_{14}/\pi_{14} = 0.02/0.7 = 0.02857$; $D_{15} = \Theta_{15}/\pi_{15} = 0.04/0.6 = 0.06667$;

$D_{23} = \Theta_{23}/\pi_{23} = 0.06/0.5 = 0.12$; $D_{24} = \Theta_{24}/\pi_{24} = 0.02/0.7 = 0.02857$;

$D_{25} = \Theta_{25}/\pi_{25} = 0.04/0.6 = 0.06667$; $D_{34} = \Theta_{34}/\pi_{34} = 0.03/0.6 = 0.05$;

$D_{35} = \Theta_{35}/\pi_{35} = 0.06/0.5 = 0.12$; $D_{45} = \Theta_{45}/\pi_{45} = 0.02/0.7 = 0.02857$;

$$\gamma_{12}^2 = \left(\frac{y_1}{\pi_1} - \frac{y_2}{\pi_2}\right)^2 = \left(\frac{12}{0.8} - \frac{14}{0.8}\right)^2 = 6.25;$$
$$\gamma_{13}^2 = \left(\frac{y_1}{\pi_1} - \frac{y_3}{\pi_3}\right)^2 = \left(\frac{12}{0.8} - \frac{10}{0.7}\right)^2 = 0.510204081;$$
$$\gamma_{14}^2 = \left(\frac{y_1}{\pi_1} - \frac{y_4}{\pi_4}\right)^2 = \left(\frac{12}{0.8} - \frac{18}{0.9}\right)^2 = 25;$$
$$\gamma_{15}^2 = \left(\frac{y_1}{\pi_1} - \frac{y_5}{\pi_5}\right)^2 = \left(\frac{12}{0.8} - \frac{22}{0.8}\right)^2 = 156.25;$$
$$\gamma_{23}^2 = \left(\frac{y_2}{\pi_2} - \frac{y_3}{\pi_3}\right)^2 = \left(\frac{14}{0.8} - \frac{10}{0.7}\right)^2 = 10.33163265;$$
$$\gamma_{24}^2 = \left(\frac{y_2}{\pi_2} - \frac{y_4}{\pi_4}\right)^2 = \left(\frac{14}{0.8} - \frac{18}{0.9}\right)^2 = 6.25;$$
$$\gamma_{25}^2 = \left(\frac{y_2}{\pi_2} - \frac{y_5}{\pi_5}\right)^2 = \left(\frac{14}{0.8} - \frac{22}{0.8}\right)^2 = 100;$$
$$\gamma_{34}^2 = \left(\frac{y_3}{\pi_3} - \frac{y_4}{\pi_4}\right)^2 = \left(\frac{10}{0.7} - \frac{18}{0.9}\right)^2 = 32.65306122$$
$$\gamma_{35}^2 = \left(\frac{y_3}{\pi_3} - \frac{y_5}{\pi_5}\right)^2 = \left(\frac{10}{0.7} - \frac{22}{0.8}\right)^2 = 174.6173469;$$
$$\gamma_{45}^2 = \left(\frac{y_4}{\pi_4} - \frac{y_5}{\pi_5}\right)^2 = \left(\frac{18}{0.9} - \frac{22}{0.8}\right)^2 = 56.25$$

(e) Estimators of variance using Sen-Yates-Grundy formula:

(i.) From the first sample, we have

$$\hat{v}(\hat{Y}_{HT})_{SYG(1)} = \sum_{i<j \in s_1} \sum D_{ij}\gamma_{ij}^2$$
$$= D_{12}\gamma_{12}^2 + D_{13}\gamma_{13}^2 + D_{14}\gamma_{14}^2 + D_{23}\gamma_{23}^2 + D_{24}\gamma_{24}^2 + D_{34}\gamma_{34}^2$$
$$= 0.06667 \times 6.25 + 0.12 \times 0.510204 + 0.02857 \times 25 + 0.12 \times 10.33163 + 0.02857 \times 6.25 + 0.05 \times 32.65306$$
$$= 0.416666666 + 0.061224489 + 0.714285714 + 1.239795918 + 0.178571428 + 1.632653061$$
$$= 4.243197276.$$

Probability Puzzle

Sarjinder Singh

(ii) From the second sample, we have

$$\hat{v}\left(\hat{Y}_{HT}\right)_{SYG(2)} = \sum\sum_{i<j\in s_2} D_{ij}\gamma_{ij}^2$$
$$= D_{12}\gamma_{12}^2 + D_{13}\gamma_{13}^2 + D_{15}\gamma_{15}^2 + D_{23}\gamma_{23}^2 + D_{25}\gamma_{25}^2 + D_{35}\gamma_{35}^2$$
$$= 0.06667\times 6.25 + 0.12\times 0.510204 + 0.06667\times 156.25 + 0.12\times 10.33163 + 0.06667\times 100 + 0.12\times 174.61735$$
$$= 0.416666666 + 0.061224489 + 10.41666667 + 1.239795918 + 6.666666667 + 20.95408163$$
$$= 39.75510204.$$

(iii) From the third sample, we have

$$\hat{v}\left(\hat{Y}_{HT}\right)_{SYG(3)} = \sum\sum_{i<j\in s_3} D_{ij}\gamma_{ij}^2$$
$$= D_{12}\gamma_{12}^2 + D_{14}\gamma_{14}^2 + D_{15}\gamma_{15}^2 + D_{24}\gamma_{24}^2 + D_{25}\gamma_{25}^2 + D_{45}\gamma_{45}^2$$
$$= 0.06667\times 6.25 + 0.02857\times 25 + 0.06667\times 156.25 + 0.02857\times 6.25 + 0.06667\times 100 + 0.02857\times 56.25$$
$$= 0.416666666 + 0.714285714 + 10.41666667 + 0.178571428 + 6.666666667 + 1.607142857$$
$$= 20.00000001$$

(iv) From the fourth sample, we have

$$\hat{v}\left(\hat{Y}_{HT}\right)_{SYG(4)} = \sum\sum_{i<j\in s_4} D_{ij}\gamma_{ij}^2$$
$$= D_{13}\gamma_{13}^2 + D_{14}\gamma_{14}^2 + D_{15}\gamma_{15}^2 + D_{34}\gamma_{34}^2 + D_{35}\gamma_{35}^2 + D_{45}\gamma_{45}^2$$
$$= 0.12\times 0.510204 + 0.02857\times 25 + 0.06667\times 156.25 + 0.05\times 32.65306 + 0.12\times 174.61734 + 0.02857\times 56.25$$
$$= 0.061224489 + 0.714285714 + 10.41666667 + 1.632653061 + 20.95408163 + 1.607142857$$
$$= 35.38605442.$$

(v) From the fifth sample, we have

$$\hat{v}\left(\hat{Y}_{HT}\right)_{SYG(5)} = \sum\sum_{i<j\in s_5} D_{ij}\gamma_{ij}^2$$
$$= D_{23}\gamma_{23}^2 + D_{24}\gamma_{24}^2 + D_{25}\gamma_{25}^2 + D_{34}\gamma_{34}^2 + D_{35}\gamma_{35}^2 + D_{45}\gamma_{45}^2$$
$$= 0.12\times 10.33163 + 0.02857\times 6.25 + 0.06667\times 100 + 0.05\times 32.65306 + 0.12\times 174.61734 + 0.02857\times 56.25$$
$$= 1.239795918 + 0.178571428 + 6.666666667 + 1.632653061 + 20.95408163 + 1.607142857$$
$$= 32.27891156.$$

(f) True variance is given by

$$V\left(\hat{Y}_{HT}\right)_{SYG} = \sum\sum_{i<j\in\Omega} \Theta_{ij}\gamma_{ij}^2$$
$$= \Theta_{12}\gamma_{12}^2 + \Theta_{13}\gamma_{13}^2 + \Theta_{14}\gamma_{14}^2 + \Theta_{15}\gamma_{15}^2 + \Theta_{23}\gamma_{23}^2 + \Theta_{24}\gamma_{24}^2 + \Theta_{25}\gamma_{25}^2 + \Theta_{34}\gamma_{34}^2 + \Theta_{35}\gamma_{35}^2 + \Theta_{45}\gamma_{45}^2$$
$$= 0.04\times 6.25 + 0.06\times 0.510204 + 0.02\times 25 + 0.04\times 156.25 + 0.06\times 10.33163 + 0.02\times 6.25 + 0.04\times 100 + 0.03\times 32.65306 + 0.06\times 174.61734 + 0.02\times 56.25$$
$$= 0.25 + 0.030612244 + 0.50 + 6.25 + 0.619897959 + 0.125 + 4.00 + 0.979591836 + 10.47704082 + 1.125$$
$$= 24.35714286.$$

(g) Unbiasedness: By the definition of expected value, we have

$$E\left[\hat{v}\left(\hat{Y}_{HT}\right)_{SYG}\right] = \sum_{t=1}^{5} p_t \hat{v}\left(\hat{Y}_{HT}\right)_{SYG(t)}$$
$$= p_1\hat{v}\left(\hat{Y}_{HT}\right)_{SYG(1)} + p_2\hat{v}\left(\hat{Y}_{HT}\right)_{SYG(2)} + p_3\hat{v}\left(\hat{Y}_{HT}\right)_{SYG(3)} + p_4\hat{v}\left(\hat{Y}_{HT}\right)_{SYG(4)} + p_5\hat{v}\left(\hat{Y}_{HT}\right)_{SYG(5)}$$
$$= 0.2(4.243197276) + 0.1(39.75510204) + 0.3(20.00000001) + 0.2(35.38605442) + 0.2(32.27891156)$$
$$= 24.35714286.$$

(h) Variance of the estimator of variance

$$V\left[\hat{v}\left(\hat{Y}_{HT}\right)_{SYG}\right] = \sum_{t=1}^{5} p_t\left\{\hat{v}\left(\hat{Y}_{HT}\right)_{SYG(t)} - E\left[\hat{v}\left(\hat{Y}_{HT}\right)_{SYG}\right]\right\}^2$$
$$= 0.2(4.243197276 - 24.35714286)^2 + 0.1(39.75510204 - 24.35714286)^2 + 0.3(20.00000001 - 24.35714286)^2 + 0.2(35.38605442 - 24.35714286)^2 + 0.2(32.27891156 - 24.35714286)^2$$
$$= 147.1975.$$

(i) Let $I_{ij} = \begin{cases} 1, & \text{if i and j} \in s \\ 0, & \text{otherwise} \end{cases}$ such that $E(I_{ij}) = \pi_{ij}$, $E(I_{ij}^2) = \pi_{ij}$, $E(I_{ij}I_{ik}) = \pi_{ijk}$ and $E(I_{ij}I_{kl}) = \pi_{ijkl}$, thus

$$V\left[\hat{v}\left(\hat{Y}_{HT}\right)_{SYG}\right] = E\left[\sum\sum_{i<j\in\Omega}\Theta_{ij}\left(\frac{I_{ij}}{\pi_{ij}} - 1\right)\gamma_{ij}^2\right]^2$$
$$= E\left[\Theta_{12}\left(\frac{I_{12}}{\pi_{12}}-1\right)\gamma_{12}^2 + \Theta_{13}\left(\frac{I_{13}}{\pi_{13}}-1\right)\gamma_{13}^2 + \Theta_{14}\left(\frac{I_{14}}{\pi_{14}}-1\right)\gamma_{14}^2 + \Theta_{15}\left(\frac{I_{15}}{\pi_{15}}-1\right)\gamma_{15}^2\right.$$
$$+ \Theta_{23}\left(\frac{I_{23}}{\pi_{23}}-1\right)\gamma_{23}^2 + \Theta_{24}\left(\frac{I_{24}}{\pi_{24}}-1\right)\gamma_{24}^2 + \Theta_{25}\left(\frac{I_{25}}{\pi_{25}}-1\right)\gamma_{25}^2 + \Theta_{34}\left(\frac{I_{34}}{\pi_{34}}-1\right)\gamma_{34}^2$$
$$\left. + \Theta_{35}\left(\frac{I_{35}}{\pi_{35}}-1\right)\gamma_{35}^2 + \Theta_{45}\left(\frac{I_{45}}{\pi_{45}}-1\right)\gamma_{45}^2 \right]^2$$
$$= \sum\sum_{i<j\in\Omega}\Theta_{ij}\left(\frac{1}{\pi_{ij}}-1\right)\gamma_{ij}^4 + \sum\sum_{i<\ j}\sum_{<k\in\Omega}\Theta_{ij}\Theta_{ik}\left(\frac{\pi_{ijk}}{\pi_{ij}\pi_{ik}}-1\right)\gamma_{ij}^2\gamma_{ik}^2 + \sum\sum\sum\sum_{i<\ j<k<l\in\Omega}\Theta_{ij}\Theta_{kl}\left(\frac{\pi_{ijkl}}{\pi_{ij}\pi_{kl}}-1\right)\gamma_{ij}^2\gamma_{kl}^2$$
$$= \Theta_{12}^2\left(\frac{1}{\pi_{12}}-1\right)\gamma_{12}^4 + \Theta_{12}\Theta_{13}\left(\frac{\pi_{123}}{\pi_{12}\pi_{13}}-1\right)\gamma_{12}^2\gamma_{13}^2 + \Theta_{12}\Theta_{14}\left(\frac{\pi_{124}}{\pi_{12}\pi_{14}}-1\right)\gamma_{12}^2\gamma_{14}^2$$
$$+ \Theta_{12}\Theta_{15}\left(\frac{\pi_{125}}{\pi_{12}\pi_{15}}-1\right)\gamma_{12}^2\gamma_{15}^2 + \Theta_{12}\Theta_{23}\left(\frac{\pi_{123}}{\pi_{12}\pi_{23}}-1\right)\gamma_{12}^2\gamma_{23}^2 + \Theta_{12}\Theta_{24}\left(\frac{\pi_{124}}{\pi_{12}\pi_{4}}-1\right)\gamma_{12}^2\gamma_{24}^2$$
$$+ \Theta_{12}\Theta_{25}\left(\frac{\pi_{125}}{\pi_{12}\pi_{25}}-1\right)\gamma_{12}^2\gamma_{25}^2 + \Theta_{12}\Theta_{34}\left(\frac{\pi_{1234}}{\pi_{12}\pi_{34}}-1\right)\gamma_{12}^2\gamma_{34}^2 + \Theta_{12}\Theta_{35}\left(\frac{\pi_{1235}}{\pi_{12}\pi_{35}}-1\right)\gamma_{12}^2\gamma_{35}^2$$
$$+ \Theta_{12}\Theta_{45}\left(\frac{\pi_{1245}}{\pi_{12}\pi_{45}}-1\right)\gamma_{12}^2\gamma_{45}^2 + \Theta_{13}\Theta_{12}\left(\frac{\pi_{123}}{\pi_{12}\pi_{13}}-1\right)\gamma_{12}^2\gamma_{13}^2 + \Theta_{13}^2\left(\frac{1}{\pi_{13}}-1\right)\gamma_{13}^4$$
$$+ \Theta_{13}\Theta_{14}\left(\frac{\pi_{134}}{\pi_{13}\pi_{14}}-1\right)\gamma_{13}^2\gamma_{14}^2 + \Theta_{13}\Theta_{15}\left(\frac{\pi_{135}}{\pi_{13}\pi_{15}}-1\right)\gamma_{13}^2\gamma_{15}^2 + \Theta_{13}\Theta_{23}\left(\frac{\pi_{123}}{\pi_{13}\pi_{23}}-1\right)\gamma_{13}^2\gamma_{23}^2$$
$$+ \Theta_{13}\Theta_{24}\left(\frac{\pi_{1234}}{\pi_{13}\pi_{24}}-1\right)\gamma_{13}^2\gamma_{24}^2 + \Theta_{13}\Theta_{25}\left(\frac{\pi_{1235}}{\pi_{13}\pi_{25}}-1\right)\gamma_{13}^2\gamma_{25}^2 + \Theta_{13}\Theta_{34}\left(\frac{\pi_{134}}{\pi_{13}\pi_{34}}-1\right)\gamma_{13}^2\gamma_{34}^2$$
$$+ \Theta_{13}\Theta_{35}\left(\frac{\pi_{135}}{\pi_{13}\pi_{35}}-1\right)\gamma_{13}^2\gamma_{35}^2 + \Theta_{13}\Theta_{45}\left(\frac{\pi_{1345}}{\pi_{13}\pi_{45}}-1\right)\gamma_{13}^2\gamma_{45}^2 + \Theta_{14}\Theta_{12}\left(\frac{\pi_{124}}{\pi_{14}\pi_{12}}-1\right)\gamma_{14}^2\gamma_{12}^2$$

>>

Use of third and fourth order inclusion probabilities in survey sampling cont.

$$+\Theta_{14}\Theta_{13}\left(\frac{\pi_{134}}{\pi_{14}\pi_{13}}-1\right)\gamma_{14}^2\gamma_{13}^2+\Theta_{14}^2\left(\frac{1}{\pi_{14}}-1\right)\gamma_{14}^4+\Theta_{14}\Theta_{15}\left(\frac{\pi_{145}}{\pi_{14}\pi_{15}}-1\right)\gamma_{14}^2\gamma_{15}^2$$

$$+\Theta_{14}\Theta_{23}\left(\frac{\pi_{1234}}{\pi_{14}\pi_{23}}-1\right)\gamma_{14}^2\gamma_{23}^2+\Theta_{14}\Theta_{24}\left(\frac{\pi_{124}}{\pi_{14}\pi_{24}}-1\right)\gamma_{14}^2\gamma_{24}^2+\Theta_{14}\Theta_{25}\left(\frac{\pi_{1245}}{\pi_{14}\pi_{25}}-1\right)\gamma_{14}^2\gamma_{25}^2$$

$$+\Theta_{14}\Theta_{34}\left(\frac{\pi_{134}}{\pi_{14}\pi_{34}}-1\right)\gamma_{14}^2\gamma_{34}^2+\Theta_{14}\Theta_{35}\left(\frac{\pi_{1345}}{\pi_{14}\pi_{35}}-1\right)\gamma_{14}^2\gamma_{35}^2+\Theta_{14}\Theta_{45}\left(\frac{\pi_{145}}{\pi_{14}\pi_{45}}-1\right)\gamma_{14}^2\gamma_{45}^2$$

$$+\Theta_{15}\Theta_{12}\left(\frac{\pi_{125}}{\pi_{15}\pi_{12}}-1\right)\gamma_{15}^2\gamma_{12}^2+\Theta_{15}\Theta_{13}\left(\frac{\pi_{135}}{\pi_{15}\pi_{13}}-1\right)\gamma_{15}^2\gamma_{13}^2+\Theta_{15}\Theta_{14}\left(\frac{\pi_{145}}{\pi_{14}\pi_{15}}-1\right)\gamma_{14}^2\gamma_{15}^2$$

$$+\Theta_{15}^2\left(\frac{1}{\pi_{15}}-1\right)\gamma_{15}^4+\Theta_{15}\Theta_{23}\left(\frac{\pi_{1235}}{\pi_{15}\pi_{23}}-1\right)\gamma_{15}^2\gamma_{23}^2+\Theta_{15}\Theta_{24}\left(\frac{\pi_{1245}}{\pi_{15}\pi_{24}}-1\right)\gamma_{15}^2\gamma_{24}^2$$

$$+\Theta_{15}\Theta_{25}\left(\frac{\pi_{125}}{\pi_{15}\pi_{25}}-1\right)\gamma_{15}^2\gamma_{25}^2+\Theta_{15}\Theta_{34}\left(\frac{\pi_{1345}}{\pi_{15}\pi_{34}}-1\right)\gamma_{15}^2\gamma_{34}^2+\Theta_{15}\Theta_{35}\left(\frac{\pi_{135}}{\pi_{15}\pi_{35}}-1\right)\gamma_{15}^2\gamma_{35}^2$$

$$+\Theta_{15}\Theta_{45}\left(\frac{\pi_{145}}{\pi_{15}\pi_{45}}-1\right)\gamma_{15}^2\gamma_{45}^2+\Theta_{23}\Theta_{12}\left(\frac{\pi_{123}}{\pi_{23}\pi_{12}}-1\right)\gamma_{23}^2\gamma_{12}^2+\Theta_{23}\Theta_{13}\left(\frac{\pi_{123}}{\pi_{23}\pi_{13}}-1\right)\gamma_{23}^2\gamma_{13}^2$$

$$+\Theta_{23}\Theta_{14}\left(\frac{\pi_{1234}}{\pi_{23}\pi_{14}}-1\right)\gamma_{23}^2\gamma_{14}^2+\Theta_{23}\Theta_{15}\left(\frac{\pi_{1235}}{\pi_{23}\pi_{15}}-1\right)\gamma_{23}^2\gamma_{15}^2+\Theta_{23}^2\left(\frac{1}{\pi_{23}}-1\right)\gamma_{23}^4$$

$$+\Theta_{23}\Theta_{24}\left(\frac{\pi_{234}}{\pi_{23}\pi_{24}}-1\right)\gamma_{23}^2\gamma_{24}^2+\Theta_{23}\Theta_{25}\left(\frac{\pi_{235}}{\pi_{23}\pi_{25}}-1\right)\gamma_{23}^2\gamma_{25}^2+\Theta_{23}\Theta_{34}\left(\frac{\pi_{234}}{\pi_{23}\pi_{34}}-1\right)\gamma_{23}^2\gamma_{34}^2$$

$$+\Theta_{23}\Theta_{35}\left(\frac{\pi_{235}}{\pi_{23}\pi_{35}}-1\right)\gamma_{23}^2\gamma_{35}^2+\Theta_{23}\Theta_{45}\left(\frac{\pi_{2345}}{\pi_{23}\pi_{45}}-1\right)\gamma_{23}^2\gamma_{45}^2+\Theta_{24}\Theta_{12}\left(\frac{\pi_{124}}{\pi_{24}\pi_{12}}-1\right)\gamma_{24}^2\gamma_{12}^2$$

$$+\Theta_{24}\Theta_{13}\left(\frac{\pi_{1234}}{\pi_{24}\pi_{13}}-1\right)\gamma_{24}^2\gamma_{13}^2+\Theta_{24}\Theta_{14}\left(\frac{\pi_{124}}{\pi_{24}\pi_{14}}-1\right)\gamma_{24}^2\gamma_{14}^2+\Theta_{24}\Theta_{15}\left(\frac{\pi_{1245}}{\pi_{24}\pi_{15}}-1\right)\gamma_{24}^2\gamma_{15}^2$$

$$+\Theta_{24}\Theta_{23}\left(\frac{\pi_{234}}{\pi_{24}\pi_{23}}-1\right)\gamma_{24}^2\gamma_{23}^2+\Theta_{24}^2\left(\frac{1}{\pi_{24}}-1\right)\gamma_{24}^4+\Theta_{24}\Theta_{25}\left(\frac{\pi_{245}}{\pi_{24}\pi_{25}}-1\right)\gamma_{24}^2\gamma_{25}^2$$

$$+\Theta_{24}\Theta_{34}\left(\frac{\pi_{234}}{\pi_{24}\pi_{34}}-1\right)\gamma_{24}^2\gamma_{34}^2+\Theta_{24}\Theta_{35}\left(\frac{\pi_{2345}}{\pi_{24}\pi_{35}}-1\right)\gamma_{24}^2\gamma_{35}^2+\Theta_{24}\Theta_{45}\left(\frac{\pi_{245}}{\pi_{24}\pi_{45}}-1\right)\gamma_{24}^2\gamma_{45}^2$$

$$+\Theta_{25}\Theta_{12}\left(\frac{\pi_{125}}{\pi_{25}\pi_{12}}-1\right)\gamma_{25}^2\gamma_{12}^2+\Theta_{25}\Theta_{13}\left(\frac{\pi_{1235}}{\pi_{25}\pi_{13}}-1\right)\gamma_{25}^2\gamma_{13}^2+\Theta_{25}\Theta_{14}\left(\frac{\pi_{1245}}{\pi_{25}\pi_{14}}-1\right)\gamma_{25}^2\gamma_{14}^2$$

$$+\Theta_{25}\Theta_{15}\left(\frac{\pi_{125}}{\pi_{25}\pi_{15}}-1\right)\gamma_{25}^2\gamma_{15}^2+\Theta_{25}\Theta_{23}\left(\frac{\pi_{235}}{\pi_{25}\pi_{23}}-1\right)\gamma_{25}^2\gamma_{23}^2+\Theta_{25}\Theta_{24}\left(\frac{\pi_{245}}{\pi_{25}\pi_{24}}-1\right)\gamma_{25}^2\gamma_{24}^2$$

$$+\Theta_{25}^2\left(\frac{1}{\pi_{25}}-1\right)\gamma_{25}^4+\Theta_{25}\Theta_{34}\left(\frac{\pi_{2345}}{\pi_{25}\pi_{34}}-1\right)\gamma_{25}^2\gamma_{34}^2+\Theta_{25}\Theta_{35}\left(\frac{\pi_{235}}{\pi_{25}\pi_{35}}-1\right)\gamma_{25}^2\gamma_{35}^2$$

$$+\Theta_{25}\Theta_{45}\left(\frac{\pi_{245}}{\pi_{25}\pi_{45}}-1\right)\gamma_{25}^2\gamma_{45}^2+\Theta_{34}\Theta_{12}\left(\frac{\pi_{1234}}{\pi_{34}\pi_{12}}-1\right)\gamma_{34}^2\gamma_{12}^2+\Theta_{34}\Theta_{13}\left(\frac{\pi_{134}}{\pi_{34}\pi_{13}}-1\right)\gamma_{34}^2\gamma_{13}^2$$

$$+\Theta_{34}\Theta_{14}\left(\frac{\pi_{134}}{\pi_{34}\pi_{14}}-1\right)\gamma_{34}^2\gamma_{14}^2+\Theta_{34}\Theta_{15}\left(\frac{\pi_{1345}}{\pi_{34}\pi_{15}}-1\right)\gamma_{34}^2\gamma_{15}^2+\Theta_{34}\Theta_{23}\left(\frac{\pi_{234}}{\pi_{34}\pi_{23}}-1\right)\gamma_{34}^2\gamma_{23}^2$$

$$+\Theta_{34}\Theta_{24}\left(\frac{\pi_{234}}{\pi_{34}\pi_{24}}-1\right)\gamma_{34}^2\gamma_{24}^2+\Theta_{34}\Theta_{25}\left(\frac{\pi_{2345}}{\pi_{34}\pi_{25}}-1\right)\gamma_{34}^2\gamma_{25}^2+\Theta_{34}^2\left(\frac{1}{\pi_{34}}-1\right)\gamma_{34}^4$$

$$+\Theta_{34}\Theta_{35}\left(\frac{\pi_{345}}{\pi_{34}\pi_{35}}-1\right)\gamma_{34}^2\gamma_{35}^2+\Theta_{34}\Theta_{45}\left(\frac{\pi_{345}}{\pi_{34}\pi_{45}}-1\right)\gamma_{34}^2\gamma_{45}^2+\Theta_{35}\Theta_{12}\left(\frac{\pi_{1235}}{\pi_{35}\pi_{12}}-1\right)\gamma_{35}^2\gamma_{12}^2$$

$$+\Theta_{35}\Theta_{13}\left(\frac{\pi_{135}}{\pi_{35}\pi_{13}}-1\right)\gamma_{35}^2\gamma_{13}^2+\Theta_{35}\Theta_{14}\left(\frac{\pi_{1345}}{\pi_{35}\pi_{14}}-1\right)\gamma_{35}^2\gamma_{14}^2+\Theta_{35}\Theta_{15}\left(\frac{\pi_{135}}{\pi_{35}\pi_{15}}-1\right)\gamma_{35}^2\gamma_{15}^2$$

$$+\Theta_{35}\Theta_{23}\left(\frac{\pi_{235}}{\pi_{35}\pi_{23}}-1\right)\gamma_{35}^2\gamma_{23}^2+\Theta_{35}\Theta_{24}\left(\frac{\pi_{2345}}{\pi_{35}\pi_{24}}-1\right)\gamma_{35}^2\gamma_{24}^2+\Theta_{35}\Theta_{25}\left(\frac{\pi_{235}}{\pi_{35}\pi_{25}}-1\right)\gamma_{35}^2\gamma_{25}^2$$

$$+\Theta_{35}\Theta_{34}\left(\frac{\pi_{345}}{\pi_{35}\pi_{34}}-1\right)\gamma_{35}^2\gamma_{34}^2+\Theta_{35}^2\left(\frac{1}{\pi_{35}}-1\right)\gamma_{35}^4+\Theta_{35}\Theta_{45}\left(\frac{\pi_{345}}{\pi_{35}\pi_{45}}-1\right)\gamma_{35}^2\gamma_{45}^2$$

$$+\Theta_{45}\Theta_{12}\left(\frac{\pi_{1245}}{\pi_{45}\pi_{12}}-1\right)\gamma_{45}^2\gamma_{12}^2+\Theta_{45}\Theta_{13}\left(\frac{\pi_{1345}}{\pi_{45}\pi_{13}}-1\right)\gamma_{45}^2\gamma_{13}^2+\Theta_{45}\Theta_{14}\left(\frac{\pi_{145}}{\pi_{45}\pi_{14}}-1\right)\gamma_{45}^2\gamma_{14}^2$$

$$+\Theta_{45}\Theta_{15}\left(\frac{\pi_{145}}{\pi_{45}\pi_{15}}-1\right)\gamma_{45}^2\gamma_{15}^2+\Theta_{45}\Theta_{23}\left(\frac{\pi_{2345}}{\pi_{45}\pi_{23}}-1\right)\gamma_{45}^2\gamma_{23}^2+\Theta_{45}\Theta_{24}\left(\frac{\pi_{245}}{\pi_{45}\pi_{24}}-1\right)\gamma_{45}^2\gamma_{24}^2$$

$$+\Theta_{45}\Theta_{25}\left(\frac{\pi_{245}}{\pi_{45}\pi_{25}}-1\right)\gamma_{45}^2\gamma_{25}^2+\Theta_{45}\Theta_{34}\left(\frac{\pi_{345}}{\pi_{45}\pi_{34}}-1\right)\gamma_{45}^2\gamma_{34}^2+\Theta_{45}\Theta_{35}\left(\frac{\pi_{345}}{\pi_{45}\pi_{35}}-1\right)\gamma_{45}^2\gamma_{35}^2$$

$$+\Theta_{45}^2\left(\frac{1}{\pi_{45}}-1\right)\gamma_{45}^4$$

=0.04167+0.00000+0.02381+0.17361+0.00000+0.00595+0.11111-0.10884-1.74617-0.08036
+0.00000+0.00094+0.00219+0.00000+0.00380-0.00164-0.08163+0.01000+0.06415-0.01476
+0.02381+0.00219+0.10714+0.59524-0.13284+0.00128-0.57143-0.02332-2.24508+0.01148
+0.17361+0.00000+0.59524+26.04166-2.58291-0.22321+2.77778-2.72108+0.00000+1.33929
+0.00000+0.00380-0.13284-2.58291+0.38427+0.01107+0.00000+0.20242+1.29894-0.29888
+0.00595-0.00164+0.00128-0.22321+0.01107+0.00670+0.09524-0.00583-0.56127+0.00287
+0.11111-0.08163-0.57143+2.77778+0.00000+0.09524+10.66666-1.74149+0.00000+0.85714
-0.10884+0.01000-0.02332-2.72108+0.20242-0.00583-1.74149+0.63973+3.42107-0.05248
-1.74617+0.06415-2.24508+0.00000+1.29894-0.56127+0.00000+3.42107+109.76835+1.68381
-0.08036-0.01476+0.01148+1.33929-0.29888+0.00287+0.85714-0.05248+1.68381+0.54241

=147.19750.

which is same as given in (h).

REFERENCES

Cochran, W.G. (1997). *Sampling Techniques*. 3rd Ed., Wiley.

Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury.

Singh (2003), "*Advanced Sampling Theory with Applications: How Michael Selected Amy*, pp. 1-1247, Kluwer Academic Publishers, The Netherlands"

NSW Branch
Stephen Bush

New South Wales Branch Events

For the NSW Branch, 2009 continues to be a busy year. Our branch meetings have been well attended, and several important events are on their way.

Our June meeting guest was Gareth Peters, who has recently been appointed as a lecturer at UNSW. He presented an interesting talk on the modelling of operational risk, incorporating dependencies between the many different types of operational risk. This talk drew upon his experience working for the Commonwealth Bank, and is based on research undertaken jointly between the CMIS section of CSIRO and RiskLab in ETH, Zurich.

Our July meeting was held at the University of Sydney, and our guest was Pieter Kroonenberg for Leiden University in The Netherlands. In his talk, Professor Kroonenberg discussed the methodological challenges present when a variable cannot satisfactorily enter a principal components analysis linearly. He discussed this problem, and some useful strategies for overcoming this, in the context of a study of environmental characteristics of the Great Barrier Reef.

In our August meeting, our guest will be Professor Ian Marchner from Macquarie University who will be speaking about subgroup analysis for randomised clinical trials, and in October we have confirmed that Dr Pavel Shevchenko from CSIRO will be our guest speaker.

In addition to our regular meetings, NSW Branch has two landmark events in the second half of 2009. The first of these is the 2009 Young Statisticians Conference, to be held at the University of Technology Sydney on September 25 and 26. Organisation for this event is well underway, with an interesting set of Keynote Speakers, a careers session, and many varied submitted talks. While the speakers may be young statisticians, we strongly encourage everyone to register and hear what the future leaders of the society are working on.

The second landmark event is the 10th annual JB Douglas postgraduate awards, to be held at the University of New South Wales on November 25. For 10 years now, this afternoon has given many emerging researchers the opportunity to present their research to the society, and compete for prizes. The judges and audience have time and time again commented on the high calibre of the speakers at this event. For those who are PhD students and their supervisors, we encourage you to participate in this rewarding experience. For the statistical community in general, we encourage you to come along and celebrate this landmark event, and support our emerging researchers.

Stephen Bush

Queensland Branch July Events

QLD Branch
Miranda mortlock

The new Committee at ABS at the July meeting. Dr Helen Thompson, Dr Rob Reeves, Professor Chris Lloyd (invited guest speaker), Dr Helen Johnson, Dr Nancy Spencer, Dr Peter Baker and Dr Miranda Mortlock.

Dr Rob Reeves, Prof Chris Lloyd and Peter Burke (State and Territory Statistical Services, ABS).

Our meeting was hosted by the ABS.

We were pleased to have the first meeting with our new Queensland SSAI banner. This was made possible by the generous sponsorship of SAS.

The Queensland Branch of the Stats Society held a general meeting and talk at the ABS offices in on Tuesday 14th of July 2009. Chris Lloyd, Associate Professor at the Melbourne Business School, gave a fascinating exposition on the topic of Blogs. Chris would be familiar to many statisticians as the instigator and chief contributor to "Fishing in the Bay", a statistics blog which seeks to engender discussion and conversation among the professional statistics community, especially in Australia.

But first, what the heck is a blog? Chris obliged our diverse audience by explaining for the uninitiated the concept of a blog, short for Web-log, an internet based tool for publishing and archiving dated essays. The original blogs were on line, open, diary entries, but now fulfill a range of functions, ranging from family travelogues to corporate propaganda and everything in between. Chris enthusiastically championed their potential as mechanisms for scholarly societies to disseminate information, and to encourage discussion and dialogue. Of course Chris's opinion was backed up by experience with his own statistics blog, and he took the audience through the ups and downs, and the remarkably simple process, of setting up and running a Blog.

The simplest method, Chris explained, was just to navigate to the site blogger.com, and sign up. In a few minutes, you can be writing your own blog! For those a little more internet savvy, one can download and install Wordpress, a variety of blog software, on their departmental world wide web server – and have many more options for customising the look and operation of your blog.

Chris has used his blog to disseminate opinions and information concerning all varieties of statistics. Astute readers will already have noted the references in the name of Chris's Blog to two of the greats from the opposing camps of the profession, of course I refer to Fisher, and Bayes, in tune perhaps with the pragmatism of the modern day which seeks to reconcile rather than divide. Chris then discussed a number of the issues that he had published on the blog. One of these was the issue of academic publishing, in which Chris presented a reasoned and cogent argument for the abandonment of paper based journal publishing, in favour of electronic publishing. The central features of his proposal were to have a nominal charge for uploading papers, after which they are accessible for free to all comers. Reviewers can log in and review the papers, giving a score, much as buyers and sellers are given a score on e-Bay. Reviewers must give their full name and affiliation, and reviewers are also rated for their quality. Quality of papers is then judged by their score given by reviewers, weighted by the reviewers' own quality!

Of course Chris acknowledged the vested interests involved in journal publication, and realistically thought the scheme was some way off! However this should not prevent us from imagining how such a scheme could work now.

Another of his Blogs that Chris discussed was the issue of a "bogus" journal, set up by an international publisher at the behest of a large pharmaceutical company, raising issues of independence and bias in the research published. Indeed this was one of the motivations for re-thinking scientific publication, which is currently resource intensive, costly, and depending on the good-will of academic reviewers who perform reviews for free.

Chris finished his talk by exhorting the audience to become members of "Fishing in the Bay", and to contribute to posting their views and raising issues for discussion, and hoped that the blog might at some future time be taken on by the Statistics Society as one of its activities.

Queensland Branch council member Dr Rob Reeves gave a vote of thanks after the talk, complementing Professor Lloyd on his contribution to the "Blogosphere", and seconding Chris's call for members to become involved with the discussion on Fishing in the Bay.

Miranda Mortlock

SA Branch
David Hirst

South Australian Branch News

May Meeting: Field Trials, Pedigrees and Statistics

The May meeting of the South Australian branch was held in the CSIRO building at the Waite Precinct, where the branch was treated to an entertaining presentation by Dr Helena Oakey from the Australian Research Centre for Health of Women and Babies (ARCH). The subject of the talk was Helena's PhD thesis on field trials and pedigrees, which she conducted at the University of Adelaide under the supervision of Ari Verbyla, Wayne Pitchford and Brian Cullis.

Helena started her talk with a summary of the design of field trials for crop breeding programs. In these trials, genetic varieties of field crops such as wheat or barley are tested to see which varieties have the best properties (e.g. yield, pest resistance) for commercial release. The crop breeders are also interested in finding out which combinations of varieties can be used as parents to produce the best hybrid crops. Initially, a large number of varieties are tested at a small number of experimental sites. The best performing varieties are then selected for further testing at a greater number of sites.

Helena then described a quantitative genetics model that can be used for analysing field trial data. The model can be thought of as having three different levels of partitioning. In the first level, the response of interest (e.g. yield) is partitioned into fixed environmental effects, random genetic effects and residual effects. The environmental effects are due to differences between the sites in the trial, the genetic effects are due to differences between the varieties and the residual effects are due to non-genetic factors such as spatial trends within the sites. The model allows for the genetic effects of the varieties to be different at each site, in which case the variance matrix of the genetic effects can be split into separate variance matrices for the sites and the varieties.

In order to more accurately identify the best hybrid varieties for the plant breeders, in the second level the genetic effects in the quantitative model are partitioned into additive, dominance and residual non-additive effects. The additive effects represent the breeding performance of individual varieties, while the dominance effects represent the joint performance of the parents of each variety. These effects are unable to be quantified directly, but Helena showed that if the pedigree of the varieties is known, then the relationships that correspond to the additive and dominance effects can be determined. However, this can be very computationally intensive if there are a large number of varieties.

In the final level, the dominance effects are further partitioned into between family and within family dominance effects. As varieties with the same parents have the same dominance relationships, this can reduce the number of dominance effects that need to be calculated. This is especially true if there are many varieties from the same family, so that the number of families is a lot less than the number of varieties.

Finally, Helena illustrated the benefits of partitioning the dominance effects by using an example from a sugarcane breeding program. The field trials in this program tested 2267 varieties of sugarcane, and there was pedigree information available for all of these varieties. As the varieties in the trial were from only 187 families, the number of parameters that had to be estimated was reduced from over 2 million to around 20 thousand by partitioning the model.

After the talk, a number of the attendees went to dinner at a nearby Thai restaurant.

David Hirst

Coverage Adjustments in the UK Census

Canberra Branch
James Brown

James Brown discussing the UK Population Census results.

In the July meeting of the Canberra Branch James Brown from the Institute of Education, University of London, gave a fascinating presentation on the lessons learnt from coverage adjustments that were carried out by the Office for National Statistics (ONS) in the 2001 UK Population Census, and how these were being adapted to the 2011 census. Undercoverage occurs in censuses when individuals are not given a census form or fail to return a form. This may occur because households may fail to be listed by a census collector, or certain individuals (e.g. young men) are difficult to contact or are less likely to respond.

James began his presentation by giving a brief explanation of why coverage adjustments had become so important in the UK. Despite the 1991 Census having a 96.5 % coverage rate, which is reasonably high, the census validation survey identified a failure of the level of coverage in many 'local authorities' (small regions) for which estimates were required. This lead to a complete overhaul in the way the 2001 census was conducted. In particular, for the first time in UK a large-scale independent coverage survey (CS) was conducted. It was run a very short time after the census and information in this survey was used to weight-up counts from the census, and prior to the release of final results it was also used to impute records for preliminary results. James noted that the use of the coverage survey for imputation was something that few other countries did, but it made good practical sense because of the high quality information collected in the survey.

For weight adjustments an approach called Dual-System Estimation (DSE) was used. Persons included in the CS were matched against census records using a combination of automatic and clerical matching. DSE works by using a ratio adjustment of the raw census counts based on whether individuals in the CS responded to the census or not. The CS was stratified into estimation areas by grouping local authorities into populations of around 0.5 million using geography (there are approximately 100 of these in England and Wales with Scotland and Northern Ireland similarly stratified) and by a variable measuring the likelihood of response.
By necessity, the CS was a large survey (over 350,000 households in the UK). To ensure one of the key assumptions of DSE estimation is met, that is the assumption of independence between the CS and Census, different collectors had to be used in each local area, and the listing of households had to be redone without any reference to the listings done in the census. For estimates at the individual local authority level, small area estimation techniques were used, and estimates were robustified against extreme weights by a variety of procedures.

Planning for the 2011 census and coverage survey is well underway. Some of the changes in methodology planned include: a pure post-out, post-back system for the census as this is expected to give higher response rates in problem areas, a longer fieldwork procedure in the census aimed at further improving response, but which may lead to an increase in non-sampling errors in the CS due to movement of individuals between the time of the census and survey, and the use of post-stratification in the CS rather than geographic stratification of local authorities. The impact of these effects is currently being investigated by James and ONS using computer simulation techniques.

James finished his presentation by highlighting some topics that needed further investigation. These include: better use of administrative data, better methods for robustness against extreme situations, more efficient techniques for likely scenarios, better approaches to deal with overcounts, which are likely to be more of an issue in 2011, and a general ambition to build on the success of the 2001 census.

James Brown

Canberra Branch
Phil Kokic

Business analytics and model design

In the May meeting of the Canberra Branch Rohan Baxter from the Australia Taxation Office gave a very interesting presentation on business analytics and risk probability model design. The objective of business analytics in the taxation office is to assess the risk of taxpayers not meeting their taxation obligations. The four key taxation obligation areas with risk models are registration, lodgment, reporting and payment. For the payment obligation, where a taxpayer has an overdue tax payment, the payment risk model is used to determine the best strategy to collect the tax debt while minimising the overall costs of doing so. Payment collection actions can range from no immediate action (the least expensive) to expensive legal options.

Three stages are required for business analytics: modelling and algorithm development, systems integration, and business benefit focus. In his presentation Rohan highlighted the importance of all three stages for a successful implementation in the ATO and it required experts from different areas working together effectively. The Analytics Unit, where Rohan works, is mainly responsible for the first stage.

The analytic risk models are defined in business terms at multiple levels. For example, the risk of incorrect reporting can be assessed at the level of a particular field in a form, or the form level, or the taxpayer level for all their forms, or at a group of taxpayer level (e.g. for taxpayers with the same tax agent). Risk scores can be calculated at each level and a combined risk score is also computed.

An important lesson for risk model design has been to take into account the application context. The available treatments and their cost/benefits decisions need to be taken into account for effective model design. For example, the choice between a simple model covering a large taxpayer segment (e.g. businesses) may be appropriate compared to a stratified model (e.g. for businesses in different industries).

To perform the cost-benefit analysis firstly rules or actions are defined for different risk score quantiles. These indicated what action needs to be taken against non-compliance (e.g. letter, phone call or firmer action). These define the costs. Next, measures of success are constructed, e.g. total collectible debt, recovery rate and turnover rate. This process is typically optimized to obtain the best net benefit, and different statistical models for scoring will produce different results. Rohan presented an application of the approach to retail banking. Finally he summarised how network models could be used in this context to score risk.

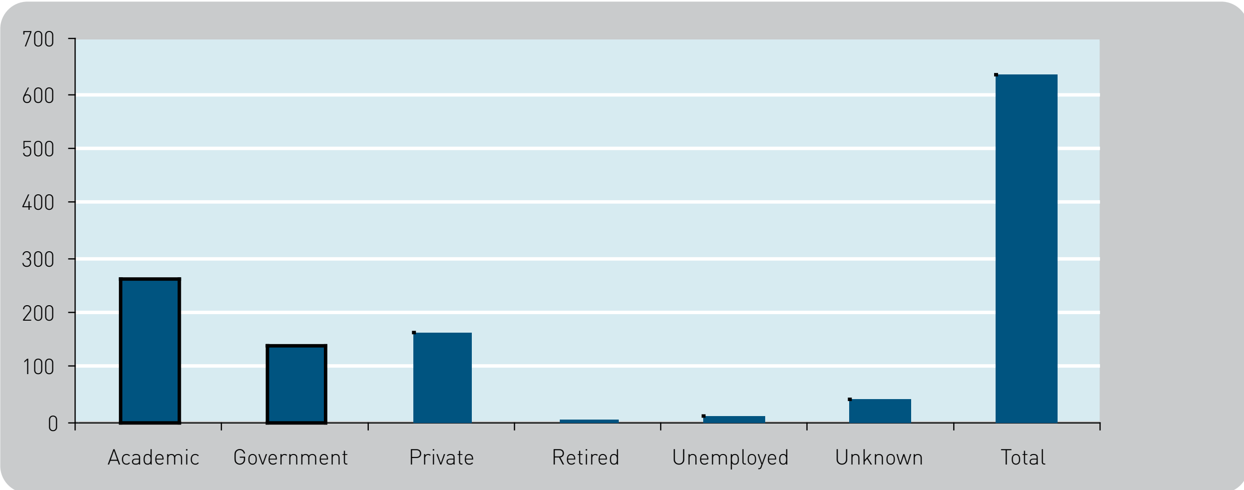
In answer to a question from the audience Rohan indicated that approximately 80 people are currently working in the Analytics teams across Australia, so it appears that its contributions are highly valued within ATO.

Phil Kokic

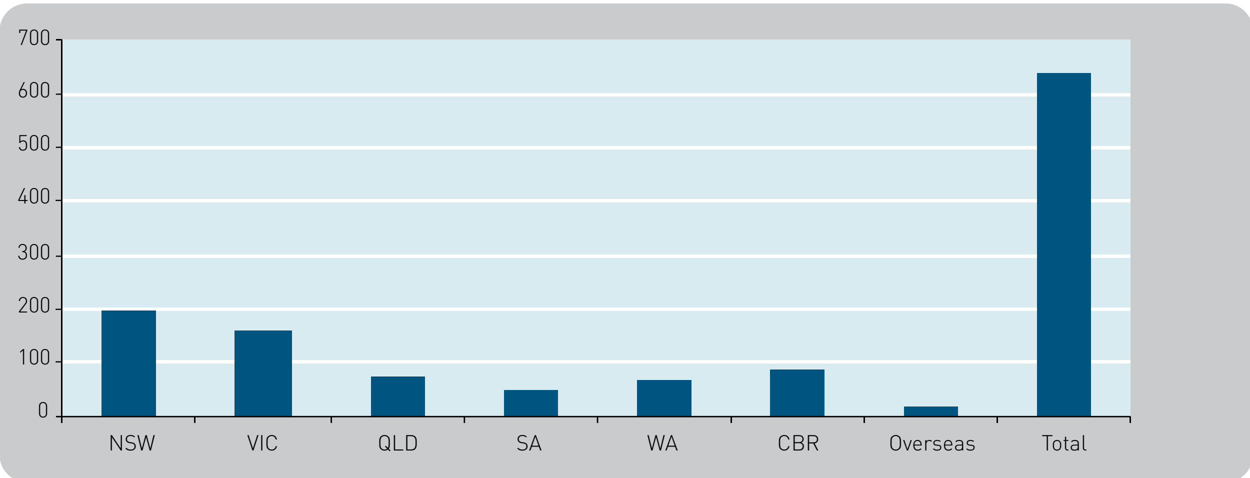
Member Numbers

SSAI Membership

SSAI Membership Numbers by Employment Affiliation (@19 August 2009)



Membership Numbers by SSAI Branch (@19 August 2009)



Please do not forget to advise the SSAI Office of any address or email address changes, so that we can stay in touch with you!

You can also update your details by logging in to the website as a member. If you have misplaced your login details, please contact the SSAI Office at eo@statsoc.org.au for assistance.

Back page

Keep in touch with SSAI through our website

WWW.STATSOC.ORG.AU

4436 DESIGN DIRECTION