BIOMETRICS BY THE BLOWHOLES
Abstract Booklet

Australasian Region of The International Biometric Society
Kiama, New South Wales, Australia

December 4–8, 2011

Contents

Invited Speakers

Kinship, heritability and genetic effect sizes

David Balding
University College London Genetics Institute
d.balding@ucl.ac.uk
Coauthors: Doug Speed, Vincent Plagnol (UCL Genetics Institute)

Relatedness is a fundamental concept in genetics, and there are several kinship coefficients to measure it. Relatedness is part of the cause of associations between genetic types and disease, the study of which is currently revolutionising our understanding of common complex diseases, as well as plant and animal breeding programmes. But relatedness it is also a principle confounder in such studies because of its role in population structure and related effects. Kinship coefficients provide the basis for estimates of heritability - the proportion of variance in a phenotype of interest that can be attributed to genetic factors. Although familiar and important, relatedness and kinship are surprisingly difficult to define and measure in useful and principled ways, and this has caused some confusion as traditional approaches based on pedigrees are giving way to methods based on genome-wide markers. Currently there is renewed interest in these concepts, as the emergence of high throughput sequencing technologies has made it feasible to identify genomic segments shared by apparently unrelated people from remote common ancestors. This raises new possibilities to understand relatedness as both a friend and a foe in the analysis of genetic mechanisms of disease, and to exploit it in population linkage methods that combine some aspects of both linkage and association methods of gene mapping. Connected with these ideas, new ways have recently arisen to measure and exploit heritability from genome-wide markers. I will review the above inter-related topics, and discuss the role of various underlying assumptions, for example about the distribution of genetic effect sizes.

Link functions in multi-locus models: implications for testing, prediction, and interpretation

David Clayton
Cambridge Institute for Medical Research, University of Cambridge
dc208@cam.ac.uk

"Complex" diseases are, by definition, influenced by multiple causes, both genetic and environmental and statistical work on the joint action of multiple risk factors has, for more than 40 years, been dominated by the generalized linear model. In genetics, models for dichotomous traits have traditionally been approached via the model of an underlying, normally distributed, liability. This corresponds to the generalized linear model with binomial errors and a probit link function. Elsewhere in epidemiology, however, the logistic regression model, a GLM with logit link function, has been the tool of choice, largely because of its convenient properties in case–control studies.

The choice of link function has usually been dictated by mathematical convenience, but it has some important implications in (a) the choice of association test statistic in the presence of existing strong risk factors, (b) the ability to predict disease from genotype given its heritability, and (c) the definition, and interpretation of epistasis (or epistacy). I will review these issues and propose a new association test.

Towards an improved multi-environment trial analysis for the National Variety Trials system

Brian Cullis
School of Mathematics and Applied Statistics, Faculty of Informatics, University of Wollongong and Environmental Informatics, CSIRO
bcullis@uow.edu.au
Coauthors: Alison Smith (University of Wollongong) and Robin Thompson (Rothamstead Research, UK)

The National Variety Trials (NVT) system was established in 2005 and is supported by grain growers through the Grains Research and Development Corporation (GRDC). It has the remit to provide sound information to growers, advisors and researchers on the performance and characteristics of grain crop varieties.

The key trait for consideration is grain yield and an important outcome is to provide an annual analysis of yield data for all crops. This involves both so-called single site analysis and an overall multi-environment trial (MET) analysis. The single site analysis follows the methods of Stefanova *et. al.,* (2009, JABES). The MET analysis currently in use has been described by Smith *et. al.,* (2001a, ANZJS) and comprises a two-stage analysis. In the first stage individual trials are analysed to obtain variety means (and weights) to be used in the second stage. The second stage involves a mixed model with variance components for the variety main effects and a partitioning of the variety by trial (V×T) interaction into (a minimum of) variety by region, variety by year and variety by region by year effects. Use of this approach is a legacy of the approaches used for the state-based testing programs that were in place prior to the commencement of NVT.

It is well known that this partitioning of V×T interaction is not competitive with the use of the factor analytic (FA) modelling approach advocated by Smith *et. al.,* (2001b, Biometrics). Implementation of an FA modelling approach for the NVT-MET analysis is challenging for several reasons. Firstly, connectivity across years can be poor. Secondly, there is a large number of trials so that FA models fitted to V×T effects suffer from problems of numerical instability as well as computational constraints for most hardware platforms. Furthermore, although FA models provide a more realistic model for explaining V×T interaction, it may be preferable to search for a more parsimonious approach for MET data-sets with relatively large numbers of trials.

In this paper we present a new approach to the NVT-MET analysis which involves a so-called informed reduced factor model. This model aims to preserve the attractive properties of the FA model in explaining V×T interaction, but reduces the computational burden and numerical instability via a dimension reduction in the environmental space of the MET. The approach is illustrated using the MET data-set for main season wheat from 2005 to 2010 with over 1100 trials.

Nonparametric spatial models for extremes: Application to extreme temperature data

Montserrat Fuentes
North Carolina State University, U.S.A
fuentes@ncsu.edu
Coauthors: John Henry and Brian Reich (North Carolina State University, U.S.A)

Estimating the probability of extreme temperature events is difficult because of limited records across time and the need to extrapolate the distributions of these events, as opposed to just the mean, to locations where observations are not available. Another related issue is the need to characterize the uncertainty in the estimated probability of extreme events at different locations. Although the tools for statistical modelling of uni-variate extremes are well-developed, extending these tools to model spatial extreme data is an active area of research.

In this work, in order to make inference about spatial extreme events, we introduce a new nonparametric model for extremes. We present a Dirichlet-based copula model that is a flexible alternative to parametric copula models such as the normal and t-copula. This presents the most flexible multivariate copula approach in the literature, and allows for non-stationarity in the spatial dependence of the extremes. The proposed modelling approach is fitted using a Bayesian framework that allows us to take into account different sources of uncertainty in the data and models. We apply our methods to annual maximum temperature values in the east-south-central United States.

Some issues in the design and analysis of multi-environment trial data in plant breeding and cultivar testing

Hans-Peter Piepho
Universitat Hohenheim, Germany
Hans-Peter.Piepho@uni-hohenheim.de

Series of field experiments conducted at multiple environments constitute an integral part of any plant breeding program and cultivar testing system. One of the main challenges in the analysis of multi-environment trial (MET) data is how to model between-trial (genotype-environment interaction) and within-trial variation. The problems are in many ways related to those in meta-analysis of medical multi-centre trials. With the advent of molecular marker data at unprecedented density, a further challenge is posed by the integration of phenotypic with genotypic data in plant breeding programs, for example in genomic selection, where the bottleneck nowadays is shifting from the genotyping to the phenotyping side. In my talk, I will review several modelling options for MET data using a number of examples, mainly from plant breeding and cultivar testing. In addition, I will consider the experimental design of MET, focussing on augmented p-rep designs.

Biometry - a lost art?

Louise Ryan
CSIRO Mathematics, Informatics and Statistics
Louise.Ryan@csiro.au

As articulated succinctly by our societys flagship journal, biometry refers to the development of statistical and mathematical methods applicable to data analysis problems in the biological sciences. Although the roots of our discipline are founded in the work of people like Fisher and others who were deeply involved in genetics, there are some who argue that biometry has become a lost art and that newer disciplines such as bioinformatics and computational biology have taken over. In this talk, I will discuss some of the commonalities and distinctions between these various fields and put forward a vision for how biometricians can thrive in todays genomic era. The talk will draw on many real world examples ranging from crop science, through to biodiversity and human studies of ageing.

Variational Bayes and genome-wide association studies

Matt Wand
University of Technology, Sydney
Matt.Wand@uts.edu.au
Coauthors: David Balding (University College London), Shen Wang and Sarah Neville (University of Wollongong)

Variational Bayes is a fast alternative to Markov chain Monte Carlo for approximate inference in hierarchical Bayesian models. We describe variational Bayes and its use in genome-wide association studies, in which hundreds of thousands of single-nucleotide polymorphism genotypes are simulataneously screened. New variational Bayes methodology, involving the negative-exponential-gamma penalisation, is explained and illustrated.

Biometrics Showcase

Variance estimation for systematic designs in spatial surveys

Rachel Fewster
Department of Statistics, University of Auckland
r.fewster@auckland.ac.nz

In spatial surveys for estimating the density of animals or plants in a survey region, systematic designs will usually yield lower variance than random designs. However, estimating the systematic variance is well-known to be problematic. Existing methods tend to overestimate the variance, so although the variance is genuinely reduced, it is over-reported, and the gain from the more efficient design is lost. I will describe a new estimator for systematic variance, based on modeling the encounter process over space. The new 'striplet' estimator has negligible bias and excellent precision in a wide range of scenarios, including strip-sampling, distance-sampling, and quadrat-sampling surveys, and including populations that are highly trended or have strong aggregation of objects. The estimator can make a dramatic impact on reported variance. I will show the results of applying different estimators to survey data for the spotted hyena in the Serengeti National Park, Tanzania, where the reported coefficient of variation is nearly halved by application of the new estimator. Simulations verify the correctness of the reduced estimate.

RAD Biodiversity: Relating different aspects of biodiversity to the environment

Scott Foster
CSIRO's Division of Mathematics, Informatics and Statistics and CSIRO's Wealth from Oceans Flagship
scott.foster@csiro.au
Coauthors: Piers Dunstan (CSIRO's Wealth from Oceans Flagship)

Biodiversity is an important topic of ecological research as it is central to natural resource management. A common form of data collected to investigate patterns of biodiversity is the number of individuals of each species at a series of locations. These data contain information on the number of individuals (abundance), the number of species (richness), and the relative proportion of each species within the sampled assemblage (evenness). If there are enough sampled locations across an environmental gradient then the data should contain information on how these three attributes of biodiversity change over gradients. We show that the rank abundance distribution (RAD) representation of the data provides a convenient method for quantifying these three attributes constituting biodiversity. We present a statistical framework for modeling RADs and allow their multivariate distribution to vary according to environmental gradients. The method is motivated by, and applied to, a large-scale marine survey off the coast of Western Australia, Australia. It provides a rich description of biodiversity and how it changes with environmental conditions.

Bootstrap tests of risk difference

Chris Lloyd
University of Melbourne
c.lloyd@mbs.edu

Parametric bootstrap tests have extremely good frequentist properties for discrete data. We give some examples and demonstrate how to compute the bootstrap P-value using importance sampling.

Open capture-recapture models with heterogeneity

Shirley Pledger
Victoria University of Wellington, New Zealand
shirley.pledger@vuw.ac.nz
Coauthors: Kenneth H. Pollock (North Carolina State University, USA) James L. Norris (Wake Forest University, North Carolina, USA)

Estimation of abundance is important in both open and closed population capture-recapture analysis. However unmodelled heterogeneity of capture probability leads to negative bias in abundance estimates. Here we discuss a suite of open population capture-recapture models which use finite mixtures to model heterogeneity of capture and survival probabilities. Model selection and parameter estimation use likelihood-based methods. Our example using Australian brushtail possums in New Zealand exhibits realistic abundance estimates. We also appraise the amount of overestimation of survival arising in previous methods which condition on the first capture of each animal. Simulations are used to evaluate the main features of the new models.

Contributed Talks

Expectile and Quantile Regression Using the Idea of Bayesian Semiparametric Regression

Arash Ardalan
University of Auckland
arash@stat.auckland.ac.nz
Coauthors: Matt Wand (University of Technology, Sydney), Thomas Yee (University of Auckland)

Quantile regression is gradually developing as a comprehensive approach to the regression analysis. Semiparametric quantile regression is an enhancement of parametric quantile regression that uses penalised spline basis functions to achieve greater flexibility. Several semiparametric regression models have useful formulations as hierarchical Bayesian models, with variance component parameters used to control the degrees of freedom of smooth functions. Markov chain Monte Carlo (MCMC) software can be used for fitting and inference for hierarchical Bayesian models. In this article we focus on quantile and expectile regression using the idea of Bayesian semiparametric situations. In addition, we describe variational Bayes in quantile regression which is a fast alternative to Markov chain Monte Carlo for approximate inference in hierarchical Bayesian models.

The Christchurch Earthquakes. What cost?

David Baird
VSN NZ Ltd
david@vsn.co.nz
Coauthors: George Hooper , NZ Earthquake Commission, Christchurch.

This talk will give an overview of the Christchurch earthquakes, and work I have been involved in with the NZ Earthquake Commission (EQC) trying to quantify this. This has involved two surveys, and and modelling work to quantify the cost of the residential rebuild and the displacement of people, among other aspects of interest.

Comparison of Infant Growth Models

Ken Beath
Macquarie University
ken.beath@mq.edu.au

Assessment of factors influencing infant growth are best performed using a modelling approach, however this is difficult due to the high initial rate of growth and wide variability. The aim is to obtain a model which produces a good fit to the data with a minimum number of parameters. A number of parametric models have been used, motivated mainly by ability to fit data, rather than biological considerations. Biologically it is unlikely that growth can be modeled by a simple function, so a semi-parametric model appears more appropriate and may produce more easily interpretable parameters. A semi-parametric model is based around a flexible shape which is common to all subjects, combined with parameters that transform the curve for individual subjects, with only few models of this type available. The models are compared using data from the CAPS study.

Comparisons within randomised groups can be very misleading

Martin Bland
University of York, UK
mb55@york.ac.uk
Coauthors: Douglas G Altman Centre for Statistics in Medicine University of Oxford, UK

Rather than comparing the randomised groups in a clinical trial directly, researchers sometimes look at the change in the measurement between baseline and the end of the trial; they test whether there was a significant change from baseline, separately in each randomised group. They report that this difference is significant in one group but not in the other, and conclude that this is evidence that the treatments are different. Several examples will be given, including a recent trial which received wide publicity, in which participants were randomised to receive either an anti-ageing cream or a placebo. We will show by simulation and theoretically that this approach is fundamentally flawed and capable of giving alpha errors as high as 50 per cent.

Diagnosing Outbreaks: Detecting Multivariate Anomalies in Presentations to Hospital Emergency Departments

Sarah Bolt
CSIRO Mathematics, Informatics and Statistics
sarah.bolt@csiro.au
Coauthors: Ross Sparks (CSIRO), James Lind (Gold Coast Hospital, Queensland)

While predictive tools are already being implemented to assist in forecasting the total volume of patients to Emergency Departments (Jessup et al, 2010, Journal of Health Organization and Management, 24:306-318), these tools are unable to detect and diagnose when these estimates fall short. Yet early detection of the types of patients presenting in unusually high numbers would help authorities to manage limited health resources and communicate effectively about risk, both in a timely fashion.

So we will examine an anomaly detection tool to do just that: detect when and in what way Emergency Departments in Queensland are exceeding forecasted patient volumes. The tool in question is the EWMA Surveillance Tree methodology initially proposed by Sparks and Okugami (2010, Interstat) for the monitoring of vehicle crashes. The approach incorporates three major aspects:

1. The modelling of existing behaviour.

2. The use of EWMA (exponentially weighted moving averages) smoothing to both observed and expected counts in order to build in temporal memory.

3. Lastly, the growing and pruning of decision trees in order to find areas of high deviation from expected counts in the multivariate space. The pruning procedure is chosen to control the false alarm rate.

We will provide a description of the application and results of this approach in the surveillance of the volume of patients to 18 Emergency Departments across Queensland.

Estimating the dominance relationship matrix using a simulation approach

David Butler
Agri-Science Queensland, Dept of Employment, Economic Development & Innovation
david.butler@deedi.qld.gov.au
Coauthors: Ari Verbyla (CSIRO & University of Adelaide)

For a given trait, dominance genetic effects result from the joint action of pairs of alleles at a locus. Related individuals may have these pair of alleles in common through shared ancestry, and a dominance relationship matrix (D) that represents the probability that individuals share the same pair of alleles by descent can be constructed. This D matrix can be used in the analysis, together with the additive relationship matrix (A), to further partition total genetic effects. The dominance matrix D can be very difficult to construct using direct methods, and a monte-carlo simulation approach has been implemented to approximate D. This method repeatedly traverses the pedigree, sampling two genes for each individual (one from each parent). The sampled genes of all individuals are examined pair wise and the counts of events contributing to each of 15 identity states accumulated. For inbred individuals, the genes are sampled with replacement f times, where f is the filial generation. A number of relationship matrices, including A and D, can be calculated from these identity states.

Biostatistics at the coal-face: are we discharging our scientific responsibilities?

John Carlin
Murdoch Children's Research Institute & University of Melbourne
john.carlin@mcri.edu.au

Using a recent medical collaboration that led to a "high-impact" publication as an example, I will discuss how statisticians can provide crucial input to the scientific process. Empirical research involves inductive reasoning but most applications of statistical inference in practice fall back on false applications of deductive logic. In particular, the widespread designation of associations for which P<0.05 on a hypothesis test as "significant", with this identified with the concept of an empirical "finding", and more dangerously the identification of "P>0.05" as indicating lack of association, is essentially unscientific and leads to much confusion in the scientific literature. This is of course a very old and unoriginal point but it remains remarkably pertinent and I maintain that statisticians have a responsibility to actively inject appropriate representations of uncertainty into scientific reporting. A related evergreen problem is the widespread misunderstanding of power calculations, which collaborating statisticians could do more to dispel, especially in relation to the remarkably popular retrospective power calculation. The project that I will discuss also illustrates how zealous journal editors can corrupt the reporting of research by bringing inadequate statistical understanding to their role; with powerful high-impact journals a good deal of statistical and scientific sophistication is needed to deal with these problems.

Building blocks for the modelling of quantitative information in the interpretation of forensic DNA evidence

James Curran
Department of Statistics, University of Auckland
j.curran@auckland.ac.nz
Coauthors: Hannah Kelly, Jo Bright, John Buckleton

The equipment used for genotyping forensic DNA evidence provides information about the amount of DNA present as well as information regarding the genotypes of the contributors. Most forensic laboratories seek to take advantage of this information in their interpretation because it can aid in selecting ”feasible” genotypes for the true, but unknown, contributors.

In this talk I will describe some work we have done to model the distribution of a parameter known as heterozygous balance and discuss how this might be used in future interpretation models.

This is joint work with my PhD student Hannah Kelly, and Jo Bright and John Buckleton from ESR, Auckland.

Bayesian ensemble methods for survival prediction in gene expression data

Kim-Anh Do
University of Texas M.D. Anderson Cancer Center
kim@mdanderson.org
Coauthors: Vinicius Bonato, Veera Baladandayuthapani, Bradley Broom, Ken Aldape, Eric Sulman

We propose a Bayesian ensemble method for survival prediction in high-dimensional gene expression data. We specify a fully Bayesian hierarchical approach based on an ensemble sum-of-trees model and illustrate our method using three popular survival models. Our nonparametric method incorporates both additive and interaction effects between genes, which results in high predictive accuracy compared to other methods. In addition, our method provides model-free variable selection of important prognostic markers based on controlling the false discovery rates; thus providing a unified procedure to select relevant genes and predict survivor functions. We assess the performance of our method on several simulated and real microarray data sets. We show that our method selects genes potentially related to the development of the disease as well as yields predictive performance that is very competitive to many other existing methods.

A trap-centric model for capture/recapture

Chris Field
Dalhousie University, Canada
field@mathstat.dal.ca
Coauthors: Alan Welsh (ANU), Cornelis Potgeiter and Marc Genton (Texas A&M)

The approach to modeling capture/recapture considered here departs from the classic model in that a trap-centric approach is favoured above an animal-centric approach. It is believed that both the location of traps, as well as spatial features measurable at the trap locations, have a noticeable affect on the probability of capture. The model considered here allows for the incorporation of spatial and temporal effects through the use of covariates and is based on the log-odds ratio of success probabilities for the traps. We estimate the population size using likelihood methods and assess the variability via the bootstrap.

Why are generalised linear mixed models (GLMMs) overlooked when response variables are ordered categories?

Susan Fletcher

Agri-Science Queensland, Department of Employment, Economic Development and Innovation

susan.fletcher@deedi.qld.gov.au

Coauthors: Damian Collins (NSW Department of Primary Industries) and Alison Kelly (Agri-Science Queensland, Department of Employment, Economic Development and Innovation)

Pathology experiments on agricultural crops regularly use rating scale categories to quantify disease expression in plants. The recent extension of generalised linear models to a mixed model framework, allowing random terms to be included in the model, is attractive for analysing structured experiments. While these models provide facilities to capture the discrete properties of categorical variables, this analysis method is generally overlooked in favour of linear models that imply the response variable is of equidistant scale. Averaging over categorical scales is not a sensible approach for analysis because the interval between categories is often unknown or unequal.

Data sets consisting of disease ratings on a scale of one to nine were used to compare the analysis of ordered categorical variables using linear mixed models and generalised linear mixed models (GLMMs). Results highlight the sensitivity of the analysis method when some categories have limited response. Issues regarding sparseness of data are explored by comparing predictions from an analysis where the scale was reduced from nine down to seven categories. The concept of proportional odds is considered and different link functions are compared using deviance tests. Through the analysis of these data sets, we provide a link between the theory and the practicalities of fitting a GLMM to ordinal data.

Analysis of short interrupted time series: A restricted maximum likelihood approach

Andrew Forbes
Monash University
Andrew.Forbes@monash.edu
Coauthors: Muhammad Akram (Monash University) Catherine Forbes (Monash University)

Interrupted time series designs arise often in the evaluation of population health intervention programs, such as mass media campaigns. The data consists of the repeated observation of a variable in the population before and after a population level intervention, such as in a mass media campaign to promote HIV testing. These time series are often very short in length, and as such they pose challenges to the use of routine statistical methods for time series analysis, largely due to the poor estimation of the required autocorrelation parameters with these methods.

In this talk we consider an AR(1) regression model for short interrupted time series. Prior work in the literature has proposed a variety of methods, the most complex of which involves a double application of the bootstrap in which bias correction of a standard residual- based estimator of the autocorrelation parameter is performed in the first application and variance estimation for regression model parameter estimators in the second. In this talk we propose and evaluate an alternate approach using restricted maximum likelihood (REML) for estimation of the autocorrelation parameter not previously applied in the interrupted time series literature. Using monte carlo simulations we compare the performance of regression model parameter estimators using REML with that of a variety of existing estimators, both with and without the double application of the bootstrap. We further evaluate a Satterthwaite degrees of freedom estimation approach both with and without an expected information matrix modification. Our results indicate that the performance (bias, size, power, confidence interval coverage) of the REML approach with corrected degrees of freedom and without any bias correction matches or exceeds that of double-bootstrapped approaches. This finding has the potential to enable simpler and more efficient analyses of short interrupted time series as well as providing an opportunity for a detailed study of design aspects of such series which is currently lacking in the literature.

Accelerated longitudinal designs

Sally Galbraith
The University of New South Wales
sally.galbraith@unsw.edu.au
Coauthors: Adrian Mander (MRC Biostatistics Unit, Cambridge) Jack Bowden (MRC Biostatistics Unit, Cambridge)

Longitudinal studies are ideal for investigating how characteristics of individuals change with age. Conventional longitudinal studies of age-related development take a single cohort of individuals at the same initial age and follow this cohort over time, collecting measurements on each individual at a number of different ages. By contrast, an accelerated longitudinal study follows multiple cohorts, each one starting at a different age. For example, an age range of 12 to 19 years could be covered with a single cohort measured annually from ages 12 to 19, or alternatively with one cohort measured from ages 12 to 15, and a second cohort measured from ages 16 to 19. The obvious advantage of an accelerated longitudinal design is its ability to span the age range of interest in a shorter period of time than would be possible with a conventional longitudinal design. This reduced duration may also be beneficial for lessening the impact of dropout. One potential disadvantage of an accelerated longitudinal design is the possibility of the existence of a cohort effect. Design of an accelerated longitudinal study requires consideration of a number of parameters. Specific to this type of study are the number of cohorts and the extent of overlap between cohorts, whereas common to any longitudinal study, the frequency and timing of measurements also needs to be set. Varying these parameters may produce a large collection of candidate designs, so the question of how to choose the best design arises. In addition, the study may be constrained to a maximum duration, number of participants, or number of measurements. This talk will consider criteria for choosing amongst designs, and evaluate the effect of varying design parameters against these criteria. I will also discuss the impact of dropout as well as methods for detecting cohort effects.

Modeling interrupted time series to evaluate prevention and control of infection in health care

Val Gebski
NHMRC Clinical Trials Centre, University of Sydney
`val@ctc.usyd.edu.au`
Coauthors: Kate Ellingson Jonathan Edwards, John Jernigan (Centres for Disease Control & Prevention Atlanta)and David Kleinbaum (Emory University, Atlanta).

In the US, methicillin-resistant Staphylococcus aureus (MRSA) is the most common cause of skin and soft tissue infections in patients presenting to emergency departments and is endemic in many hospitals (1). Interventions to reduce transmission include emphasising hand hygiene, active surveillance culturing, and educating healthcare workers in the culture of infection control. Common methods for evaluating interventions to reduce the rate of new Staphylococcus aureus (MRSA) infections in hospitals use segmented regression or interrupted time-series analysis. We describe approaches to evaluating interventions introduced in different healthcare units at different times. We compare fitting a segmented Poisson regression in each hospital unit with pooling the individual estimates by inverse variance. An extension of this approach to accommodate potential heterogeneity allows estimates to be calculated from a single statistical model: a stacked model. It can be used to ascertain whether transmission rates before the intervention have the same slope in all units, whether the immediate impact of the intervention is the same in all units, and whether transmission rates have the same slope after the intervention. The methods are illustrated by analyses of data from a study at a Veterans Affairs hospital. Both approaches yielded consistent results. Where feasible, a model adjusting for the unit effect should be fitted, or if there is heterogeneity, an analysis incorporating a random effect for units may be appropriate

Supervised visualisation methods for exploring genome-wide association studies: An application to the WTCCC Type 1 Diabetes data.

Alexandra Gillett
University of New South Wales, Sydney, Australia
a.gillett@student.unsw.edu.au
Coauthors: Susan Wilson (Australian National Univeristy, Australia), Sally Galbraith (Univeristy of New South Wales, Australia)

Genome-wide association studies are an important tool for identifying genetic variation associated with disease. In a genome-wide association study (GWAS) a dense set of single nucleotide polymorphisms (SNPs) is genotyped, across the genome, to investigate the role of the most common form of genetic variation in disease or, to identify the genetic loci that are risk factors for disease. The high dimensional nature of GWAS datasets, with the number of SNPs far exceeding the number of samples, makes browsing the data for underlying patterns a challenge. Visualisation uses dimension reduction methods which aim to represent high dimensional data in 2 or 3 dimensions whilst preserving similarities between data points. Metrics used to judge similarity can incorporate class information, directing dimension reduction to focus on retaining differences between classes. Applying such methods to case-control GWAS data gives insight into the underlying patterns which distinguish case from control and can be used in complement with modelling techniques to communicate important features of the dataset when classifying. In this talk I shall review several supervised visualisation methods including supervised multidimensional scaling (SMDS), model-based visualisation and the supervised neighbour retrieval visualiser (SNeRV). Methods will be demonstrating using the WTCCC Type 1 Diabetes case-control dataset. The quality of the resulting visualisations will be compared by class prediction accuracy.

A Sampling strategy for fitting large linear mixed models

Arthur Gilmour
University of Wollongong
arthur.gilmour@cargovale.com.au
Coauthors: Sue Brotherstone (University of Edinburgh, Scotland), Robin Thompson (Rothamsted Research, England)

While mixed model software can fit quite large complex mixed models, one can always envisage analyses which are too big to fit directly, even with sparse AI technology. This paper presents a sampling strategy for fitting these large models. It uses ideas from Gibbs sampling. The motivating example is the analysis of 345,000 records on 3 disciplines on 19,829 horses representing 3017 sires and ridden by 11,841 riders. The records represent 6,875 competition classes when horse/rider combinations are assessed at 4 grades. There are 12 traits based on scores on combinations of discipline and grade. The basic model was a sire model with mixed linear effects, fitted within each discipline-grade:
y = mean + Gender + age + age2 + Class + Sire + Horse + Rider + e
with Sire, Rider, Horse and e random effects. The immediate interest was to generalise this univariate model to a multivariate model and estimate residual variances for the 12 traits and the 12 x 12 variance-covariance matrices of the 12 traits at the sire, horse and rider strata. Any attempt to fit directly the full model with the large number of sire, horse, rider effects is extremely unlikely to work. Typically, one would attempt the 66 bivariate models and then synthesize the joint matrices. A method called data augmentation was used to perform a 12-trait multivariate analysis. This data augmentation is based on previous work where computational requirements are reduced, by repeatedly fitting sub-models in an overlapping series, with each sub-model being fitted in turn to data adjusted for effects not in the current sub-model. This greatly reduces the computations. A simplified form of Gibbs sampling is used to add noise to the updated estimates at each step, thus preventing bias in the estimated variance parameters. The calculations were carried out in a development version of ASReml 3. For example in the univariate case we might successively fit 3 sub-models of the form
y - [Sire + Horse + Rider] = mean + Gender + age + age2 + Class + e,
y - [Class + Rider] = mean + Gender + age + age2 + Sire + Horse + e,
y - [Class + Sire + Horse] = mean + Gender + age + age2 + Rider + e,
where the terms in square brackets are augmentations to the data y using values for the effects imputed from the previous fits. By using these three sub-models, the computational burden is reduced because Horse is nested within Sire but Rider and Class are cross-classified with Horse. Each model is fitted in turn to the augmented data and the variance components estimated from the 3 sub-models. However, to overcome biases, the fixed (random) effects used to augment the data have noise added according to the estimation (prediction error) variance of the effects. After a burnin period, the variance parameter estimates and the solutions from the successive runs are averaged. The results from this procedure are compared with estimates from other approaches.

A composite sampling strategy for the design and analysis of cereal resistance trials

Beverley Gogel
The University of Adelaide
beverley.gogel@adelaide.edu.au
Coauthors: Brian Cullis, Alison Smith

Cereal cyst nematodes (CCN) are microscopic worms that invade the developing roots of seedlings which can result in reduced plant biomass and significant yield loss. They are common throughout the cropping regions of Australia. Designed field trials are conducted annually to assess the relative resistance of cereal cultivars to these nematodes. This requires a measure of the nematode population per plot both at seeding and immediately post harvest. The process of collecting soil samples and then processing them in the laboratory to obtain these measures is time consuming and costly. Typically measures are obtained for all plots, that is, for all replicates of all varieties. A new sampling strategy has recently been proposed in which the individual replicates for a subset of the varieties are measured while a composite sample across some or all of the replicates for the remaining varieties is processed (Smith et al. 2011). This approach allows an efficient mixed model analysis and is widely applicable particularly in the context of variety evaluation trials where traits of interest are often expensive to measure. Another powerful application of this strategy is in mapping a trial area for a trait of interest prior to judicious allocation of the experimental material in a well designed experiment. Both applications can result in significant cost savings. This talk will describe the application of composite sampling in the context of a series of cereal resistance trials currently being conducted by the Molecular Diagnostics Group, South Australian Research and Development Institute (SARDI).

Smith, A.B., Thompson, R., Butler, D.B. and Cullis, B.C. (2011). The design and analysis of variety trials using mixtures of composite and individual plot samples. *Applied Statistics*, 60(3):437-455.

Propensty score and hierarchical Bayes methods for longitudinal profiling of hospital performance

Patrick Graham
University of Otago, Christchurch
patrick.graham@otago.ac.nz

Tracking hospital outcomes over time permits hospitals exhibiting unusual trends in performance to be identified. If there is evidence that some hospitals are improving more rapidly than others additional investigations may reveal innovations in organisation or practice which could lead to improvements in outcomes across the hospital system. Hierarchical Bayesian modelling provides a framework for longitudinal modelling of hospital performance. In principle, hierarchical Bayesian methods permit adjustment for case-mix variations, modelling of the effect of hospital level attributes and time trends, as well as improved precision of hospital effect estimates due to partial pooling of information across hospitals. However, because of the large number of patient attributes typically required to adequately control case-mix variations and the large size of hospital outcomes datasets, fitting a full hierarchical Bayesian model can be challenging in standard computing environments. In this paper I outline an alternative, two-stage, strategy in which multiple category propensity score methods are used to adjust for case-mix variations and hierarchical Bayesian methods are used to model propensity score stratified, hospital-specific outcomes. An unusual feature of this application of propensity score methods is that it seems necessary to regard time as an exposure variable along with hospital of treatment. The underlying potential outcomes model, from which the propensity score methods are derived, will be briefly outlined and the methodology will be illustrated via application to an analysis of 30 day post-admission mortality risks for acute myocardial infarction patients admitted to New Zealand public hospitals between 2001 and 2007.

Ordinal regression models for continuous scales

Gillian Heller
Department of Statistics, Macquarie University
gillian.heller@mq.edu.au
Coauthors: Maurizio Manuguerra (Macquarie University)

Ordinal regression analysis is a convenient tool for analyzing ordinal response variables in the presence of covariates. We extend this methodology to the case of continuous self-rating scales such as the Visual Analog Scale (VAS) used in pain assessment, or the Linear Analog Self-Assessment (LASA) scales in quality of life studies. These scales measure subjects perception of an intangible quantity, and cannot be handled as ratio variables because of their inherent nonlinearity. We express the likelihood in terms of a function connecting the scale with an underlying continuous latent variable and approximate this function either parametrically or non-parametrically. Then a general semi-parametric regression framework for continuous scales is developed. We analyse two data sets to compare our method to the standard discrete ordinal regression model, and the parametric to the non-parametric versions of the model. The first data set uses VAS data from a study on the efficacy of low-level laser therapy in the treatment of chronic neck pain; the second comes from a study on chemotherapy treatments in advanced breast cancer and looks at the impact of different drugs on patients quality of life. The continuous formulation of the ordinal regression model has the advantage of no loss of precision due to categorization of the scores and no arbitrary choice of the number and boundaries of categories. The semi-parametric form of the model makes it a flexible method for analysis of continuous ordinal scales.

Regression to the mean

Harold Henderson
AgResearch Ruakura, New Zealand
harold.henderson@agresearch.co.nz

Regression to the mean is a classic statistical paradox. Some of its history is discussed and some examples are given.

Assessing similarity of DNA profiles

Graham Hepworth
Statistical Consulting Centre, The University of Melbourne
hepworth@unimelb.edu.au
Coauthors: Ian Gordon

The genetic similarity of strains of a pathogen can be assessed using a matrix of dissimilarities derived from bands in their DNA profile which are present or absent. It is often of interest to compare groups of strains which are differentiated according to the possession of an attribute, such as the presence of HIV. We show the limitations of a previously proposed statistic for determining if a group of strains is more similar than expected, and propose a new statistic based on similarity between strains within the group of interest and with those outside.

Such a statistic needs to account for the dependence in the raw data, and we use the correlation between elements of the dissimilarity matrix to investigate how this dependence results in underestimation of the variance if unaccounted for. Our work is applied to examples involving the pathogenic yeast Candida, which causes thrush in humans.

Extrapolating cumulative incidence for survival estimation in multi-state models of randomized trial outcomes.

Malcolm Hudson
Department of Statistics, Macquarie University and Faculty of Medicine, University of Sydney
malcolm.hudson@mq.edu.au
Coauthors: Serigne Lo (The George Institute, Australia) and Stephane Heritier (The George Institute and University of Sydney, Australia)

Semi-Markov processes have gained popularity as multistage disease models describing a patients history of events over time. In this talk we extend the indirect approach of Lo, Heritier and Hudson (2009, Computational Statistics and Data Analysis **59**, 531-541) used to model major cardiovascular endpoints of the LIPID trial by relaxing the fully parametric formulation. The technique does not require the proportional hazard assumption to hold at each step, allows semi-parametric estimation of clinically relevant quantities such as the hazard ratio, the overall survival or absolute benefit from treatment. Application in a large randomized clinical trial will be presented.

The presence of high proportions of censored data in survival analysis of randomized trials introduces significant difficulties in analyses employing semi-Markov models for multiple events or states encountered prior to an endpoint of interest.

In the paper cited above we have successfully applied a parametric model in a large trial (LIPID) to study progression from randomization to death with the presence of intervening strokes, with a high proportion of type 1 censoring. This censoring occurs in each transient model state because of the prescribed period of follow up in the trial design. We employed saddle-point inversion involving cumulants to estimate survival and hazard parameters in these models, having specified parametric distributions of time to each event.

Because of the difficulty of correctly specifying each parametrization, there is great appeal in non-parametric estimation of the holding times in component states of the model. But censoring introduces difficulties since censored cases must be assigned a subsequent pathway to the endpoint, and progression through states after the period of follow up is unknown.

We provide methods of extrapolating non-parametric survival and cumulative incidence functions beyond the period of follow up. These enable semi-parametric estimation of survival to a pre-specified endpoint of a randomized trial, and hazard function. Saddle-point methods are demonstrated to be effective in this estimation, and in comparisons of overall survival in different arms of the trial. The result is a method that does not depend heavily on parametric assumptions.

Finally, a sensitivity analysis is undertaken to examine the extent of dependence of the findings on assumptions involved in extrapolation.

Keywords: cumulative incidence, multi-state model, survival analysis.

Combining pooled and individual test data to estimate herd-level prevalence

Geoff Jones
Massey University, NZ
g.jones@massey.ac.nz
Coauthors: Wes Johnson (UC Irvine, USA) Cristobal Verdugo (Massey University, NZ) Cord Heuer (Massey University, NZ)

The use of an expensive but sensitive diagnostic test with pooled samples can be a cost-effective way of monitoring herds for the presence of disease. For example the faecal culture test for *Mycobacterium avium* subsp. *paratuberculosis* (MAP) can be applied to pooled faecal samples derived from random samples of animals in a deer herd, rather than to individual animals, in order to test the infection status of the herd.

If a less accurate but cheaper test is available, it may be advantageous to also apply this to individual animals in a defined testing regime. To examine the performance of one such regime in pastoral farmed livestock in New Zealand, random samples from 99 deer herds were subjected to pooled faecal sampling, with a follow-up individual blood serum test if the pooled faecal test was negative.

We discuss the difficulties in analyzing such data, in particular the modelling of prevalence at herd and individual level and the absence of a gold-standard test to measure true disease status.

The effect of a preliminary test of homogeneity of stratum-specific odds ratios on confidence intervals for these ratios

Paul Kabaila
Department of Mathematics and Statistics, La Trobe University, Melbourne
P.Kabaila@latrobe.edu.au
Coauthors: Dilshani Tissera, Department of Mathematics and Statistics, La Trobe University, Melbourne

Consider a case-control study in which the aim is to assess the effect of a factor on disease occurrence. We suppose that this factor is dichotomous. Also suppose that the data consists of k strata, with a two-by-two table for each stratum. A commonly-proposed procedure for the analysis of this type of data is the following (see e.g. Section 13.5 Rosner, 2006). We carry out a preliminary test of homogeneity of the stratum-specific odds ratios. If the null hypothesis of homogeneity is accepted then inference about the stratum-specific odds ratios proceeds on the assumption that these odds ratios are equal. If, on the other hand, this hypothesis is rejected then inference about the stratum-specific odds ratios is carried out without assuming that these odds ratios are necessarily equal. We examine the effect of this procedure on confidence intervals constructed for the stratum-specific odds ratios. The literature on the effect of preliminary model selection on confidence regions begins with the very important work of Freeman (1989) on the effect of a preliminary test of no differential carryover in a two-treatment two-period crossover trial on the confidence interval for the difference of treatment effects. This literature is reviewed by Kabaila (2009). We find that the preliminary test of homogeneity of the stratum-specific odds ratios has a harmful effect on the coverage probabilities of the confidence intervals for these odds ratios.

References

Freeman, P.R. (1989). The performance of the tow-stage analysis of two-treatment, two-period crossover trials. Statistics in Medicine, 8, 1421-1432.

Kabaila, P. (2009). The coverage properties of confidence regions after model selection. International Statistical Review, 77, 405-414.

Rosner, B. (2006). Fundamentals of Biostatistics, 6th edition. Pacific Grove, CA: Thomson.

Estimation in a linear mixed model with a non-positive definite variance matrix

Alison Kelly
Agri-Science Queensland, Department of Employment Economic Development and Innovation, Leslie Research Centre, Toowoomba, QLD
alison.kelly@deedi.qld.gov.au
Coauthors: Brian Cullis, Arthur Gilmour and Robin Thompson

Residual maximum likelihood estimation is routinely used to estimate variance parameters in mixed models. When the variance matrix, G, of a set of random effects is non-positive definite (npdG), difficulties arise because estimation algorithms involve computing the inverse of this matrix. In general, G could be non-positive definite by design, due to the terms in the mixed model equations. Examples include use of mixed models for fitting cubic smoothing splines or the estimation of dominance effects in the analysis of hybrid plant breeding data using pedigree information. Alternatively, a positive definite G could become non-positive definite due to the response data when updates are obtained from the iterative solution of the score equations. In existing software, npdG by design can be accommodated by applying constraints but handling npdG due to the data is more challenging. Using an iterative scheme such as the expectation-maximisation algorithm ensures updates of G remain positive definite but these schemes are known to be slow. Matrix bending may also be used to ensure the updates of G remain positive definite. We present novel approaches that accommodate npdG and are easily implemented within the framework of the average information algorithm.

Functional longitudinal response modelling

Steve Lane
The University of Melbourne
s.lane@pgrad.unimelb.edu.au

Functional response data appear in many biological studies. Motivated by the prediction of tree diameter density functions, we present a nonparametric method that allows prediction of longitudinal functional responses, accounting for (possible) covariate information.

Recent developments in exploratory and integrative multivariate approaches for 'omics' data: application to a kidney transplant study

Kim-Anh Le Cao
Queensland Facility for Advanced Bioinformatics, University of QLD
k.lecao@uq.edu.au

With the availability of many 'omics' data, such as transcriptomics, proteomics or metabolomics, the integrative or joint analysis of multiple datasets from different technology platforms is becoming crucial to unravel the relationships between different biological functional levels. However, the development of such analyses is a major computational and technical challenge as most approaches suffer from high data dimensionality, as the number of measured biological entities (the variables) is much greater than the number of samples or patients. Promising exploratory and integrative approaches have been recently developed for that purpose, such as sparse variants of Principal Component Analysis, and Partial Least Squares regression, in order to select the relevant variables related to the system under study. We will illustrate a whole range of these methodologies to a kidney transplant study from the PROOF Centre (Centre of Excellence for the Prevention of Organ Failure, Vancouver) that includes transcriptomics, proteomics and clinical data. We will show how we can get a deeper understanding of the data and select potential biomarkers to classify acute rejection or non rejection of kidney transplant. All these methodologies are implemented in the R package mixOmics as well as in its associated web-interface http://mixomics.qfab.org/

Fitting hierarchical GLMs

Youngjo Lee
Department of Statistics, Seoul National University, Seoul, South Korea
`youngjo@snu.ac.kr`

Hierarchical GLMs –GLMs with random effects– provide a very rich class of models for data analyses. Now general HGLMs codes are available in CRAN R packages. I want to show the current status of codes and demonstrate the models which can be fitted using the current R-codes.

Choice of prognostic estimators in joint models by estimating differences of expected conditional Kullback-Leibler risks

Benoit Liquet
INSERM, University Victor Segalen, Bordeaux, France
Benoit.Liquet@isped.u-bordeaux2.fr

Prognostic estimators for a clinical event may use repeated measurements of markers in addition to fixed covariates. These measurements can be linked to the clinical event by joint models that involve latent features. When the objective is to choose between different prognosis estimators based on joint models, the conventional Akaike information criterion (AIC) is not well adapted and decision should be based on predictive accuracy. We define an adapted risk function called expected prognostic cross-entropy (EPCE). We define another risk function for the case of right-censored observations, the expected prognostic observed cross entropy (EPOCE). These risks can be estimated by leave-one-out crossvalidation, for which we give approximate formulas and asymptotic distributions. The approximated crossvalidated estimator CVPOLa of EPOCE is studied in simulation and applied to the comparison of several joint latent-class models for prognosis of recurrence of prostate cancer using prostate specific antigen (PSA) measurements.

Rank tests for data from complex surveys

Thomas Lumley
Department of Statistics, University of Washington
`t.lumley@auckland.ac.nz`
Coauthors: Alastair J. Scott

Data from complex multistage survey samples such as NHANES are increasingly important in the health sciences, and researchers expect to be able to use all the statistical techniques familiar from independent data. Design-based versions of rank tests such as the Wilcoxon test have not been developed, except for a few special cases, so researchers are using rank tests for independent data instead. We show how to construct general rank tests under complex sampling, both for comparing groups within a survey and for using a national survey as a reference distribution.

Penalized likelihood approaches for Cox model fitting with interval censored data

Jun Ma
Department of Statistics, Macquarie University
jun.ma@mq.edu.au
Coauthors: Jinqing Li, Stephane Heritier, Ian Marschner

We consider the problem of Cox model fitting when the time to event observations are interval censored. We propose a novel estimation procedure developed from the constrained penalized likelihood maximization, with both baseline hazard and regression coefficients estimated simultaneously. The penalty function is used to smooth the baseline hazard and, moreover, the baseline hazard is subjected to the non-negativity constraint. An efficient alternating optimization procedure using the Newton algorithm and the multiplicative iterative (MI) algorithm is developed to maximize the constrained penalized likelihood function. We demonstrate the successfulness of this method by a simulation study and an application to a real data.

Key words: Cox regression model, interval censored observations, maximum constrained penalized likelihood, multiplicative iterative algorithm.

Generalised linear models in R: Problems and fixes

Ian Marschner
Department of Statistics, Macquarie University
ian.marschner@mq.edu.au

The standard function for fitting a generalised linear model (GLM) in R is glm. This function implements iteratively reweighted least squares with some modifications to prevent divergence of the iterative sequence. These modifications often allow glm to successfully fit a GLM in numerically unstable situations, such as when the estimate lies on the boundary of the parameter space. However, there are many other cases where glm should be able to converge but does not. These problems are most common with non-standard link functions such as log link binomial or identity link Poisson models, but they can even occur with canonical models such as logistic regression. Various examples of aberrant behaviour in glm will be discussed, and a simple proposal to address these problems will be presented. This involves an additional modification to the IRLS algorithm, in which a line search is used to force the deviance to decrease at each iteration. The proposed change has been implemented in an R add-on package called **glm2**, which contains a function called glm2. This function operates identically to glm aside from the improved computational algorithm. Results will be presented illustrating superior performance of glm2 compared to glm.

A multivariate omnibus test: Swiss Army Knife or plastic spork?

Brian McArdle
Department of Statistics, University of Auckland
b.mcardle@auckland.ac.nz

I introduce a way of combining p-values (based on Fishers Omnibus test) from separate univariate tests to perform a test of a multivariate hypothesis. I investigate its potential to handle common situations using ecological examples that are currently difficult or intractable: eg. multivariate linear mixed models, multivariate a posterior comparisons. I hope to show its potential for a wide variety of other situations.

Health Effects of the September 2009 Dust Storm in Sydney, Australia: Did Emergency Department Visits and Hospital Admissions Increase?

Alistair Merrifield
NSW Health
amerr@doh.health.nsw.gov.au
Coauthors: Suzanne Schindeler, Bin Jalaludin, Wayne Smith.

A large growing body of literature supports the association between exposure to particulate air pollution and adverse health outcomes. During September 2009, a rare large dust storm event was experienced in Sydney, NSW, Australia. Extremely high levels of respirable particles were recorded. We conducted an analysis to determine whether the dust storm was associated with increases in all-cause, cardiovascular, respiratory and asthma-related ED presentations and hospital admissions. We used Poisson generalized additive models to model the ED presentations and hospital admissions and adjust for pollutants, humidity, temperature and day of week effects to obtain estimates of relative risks (RR), 95% confidence intervals and p-values associated with the dust storm.

The dust storm period was associated with large significant increases in asthma ED visits (RR approximately 1.26-1.27, $p < 0.01$) and asthma hospital admissions (RR 1.15, p=0.01), and to a lesser extent, all ED visits (RR 1.09-1.10, $p < 0.01$), all-cause hospital admissions (RR 1.03-1.04, $p < 0.01$) and respiratory ED visits (RR 1.10-1.11, $p < 0.01$). There was no significant increase in cardiovascular ED visits (p=0.39) or cardiovascular hospital admissions (p=0.43-0.96). Age-specific analyses showed the dust storm wasnt associated with increases in respiratory or asthma ED visits in the 65+ year age group; the < 65 year group had higher risks of respiratory and asthma-related ED presentations. We recommend public health measures, especially targeting asthmatics, should be implemented during future dust storm events.

Estimating the number of salmon returning to spawn

Russell Millar
Dept of Statistics, University of Auckland
r.millar@auckland.ac.nz
Coauthors: Sam McKechnie and Chris Jordan

If observed numbers of spawning salmon are plotted against sampling date then the area under the curve (AUC) gives an estimate of spawner-days. Dividing AUC by spawner lifetime, and adjusting for observer efficiency, gives an estimate of spawner escapement. In particular, the trapezoidal form of AUC estimator has been widely used over the last several decades, despite the absence of a direct method for calculating its variance. For this reason, an alternative estimator (Hilborn et al, 1999) of escapement was developed using a maximum likelihood (ML) model of spawner arrivals. However, this alternative has not been widely used, perhaps due to its complexity and concerns over validity of assumptions. Here, a simpler ML approach is used to estimate AUC by fitting a model directly to spawner numbers. It can be fitted using existing generalized linear modeling software, and provides an explicit variance estimator for AUC. Simulations show that it is robust to violations of model assumptions, and has better performance than the more complex estimator.

EEG amplitude as an indicator of brain maturation in premature infants

Michael Navakatikyan
Australian Health Services Research Institute, University of Wollongong, Wollongong, NSW, Australia & Department of Statistics, University of Auckland, Auckland, New Zealand
mnavakat@uow.edu.au
Coauthors: Deirdre OReilly (Harvard Medical School, Boston, MA, USA), Marcia Filip (Brigham and Womens Hospital, Boston, MA, USA), Deirdre Greene (Brigham and Womens Hospital, Boston, MA, USA), Linda Van Marter (Harvard Medical School, Boston, MA, USA)

The aim of the work was an assessment of the strength of the association between amplitude of electroencephalographic activity (EEG) and postmenstrual age (PMA). Two-channel EEG recordings 3 to 6 hours long were collected 4 to 10 times from 26 premature infants (at $<$ 28 weeks of gestational age) totalling 177. The raw EEG was converted into range-EEG (rEEG) measure of EEG amplitude. Mean, median (Me), lower (5th) and upper (95th) percentiles (LP, UP), indices of spread (width = UP-LP, standard deviation and coefficient of variation), asymmetry = ((UP-Me)-(Me-LP))/(UP-LP) were calculated for each 1-minute epoch; and their medians over the whole recording were taken for analysis. Association with PMA was studied using linear mixed models, and measured as fixed-effects R-squared. Simulation was performed to predict 2.5% and 97.5% boundaries for normal values of the rEEG indices. As post-menstrual age advances the general tendency can be described as significant increase in the value of LP and decrease in the values of UP, spread and asymmetry. The most prominent change was observed for the indices of spread calculated on log-transformed values of rEEG (fixed-effects R-squared = 0.84-0.89). Thus, indices of rEEG spread can be regarded as indicators of neonatal brain maturation and used for neonatal EEG monitoring.

Key words: Electroencephalography (EEG); Premature infant; Monitoring; EEG amplitude; Range-EEG; Postmenstrual age (PMA)

Not all black-boxes have the answers an NIR calibration story

Sharon Nielsen
Charles Sturt University School of Computing and Mathematics, Australia
snielsen@csu.edu.au
Coauthors: Ken Russell (Charles Sturt University School of Computing and Mathematics), Alison Kelly (Agri-Science Queensland, DEEDI) & Glen Fox (The Queensland University, QAAFI).

Near infrared (NIR) spectral analysis is a rapid, non-destructive assessment tool widely used to predict or describe quality parameters of materials under investigation. Reflectance values from the NIR region of the electromagnetic spectrum (700 nm 2400 nm) are measured on samples of the material, while quality measurements are made on the same samples using usual laboratory techniques. The calibration equation is developed to explain the relationship between the scanned reflectance data and the laboratory data.

There are a range of statistical techniques currently used in the calibration-prediction process. These do not always account for the complex correlation structure within NIR spectral data. Some of the statistical methods currently used include multiple linear regression, principal component regression, partial least squares, factor analysis and artificial neural networks. Linear mixed model developments have been extended to include the capacity to model correlated data, such as NIR spectral data, but the application of linear mixed models in the field of NIR spectral analysis has been limited.

In my talk, I will investigate the statistics underlying the present procedures in NIR analysis. In addition I will seek to outline the potential role of the linear mixed model in the field of NIR analysis.

A genomic application of mutual information between discrete and continuous variables to identify gene modules

Chris Pardy
UNSW, Prince of Wales Clinical School
cpardy@unsw.edu.au
Coauthors: Susan Wilson

Genomic experiments give large quantities of high-dimensional data. Systems biology attempts to integrate the multiple sources of these data into a coherent description of the connections between processes within an organism. The data include gene expression levels, single nucleotide polymorphisms (SNPs) and clinically measured outcomes. Information theoretic and machine learning techniques such as mutual information (MI) and clustering are particularly useful in this context.

We extend a previous approach by Zhang and Horvath (2005, *Statistical Applications in Genetics and Molecular Biology* 4(1), 1128) by using MI as a measure of association that is valid for both continuous and discrete variables. This allows us to create a single information matrix including both types of data to use as a distance measure for clustering and network inference. The network can be grouped according to association with important clinical measurements with the aim of identifying modules containing biological pathways and highly connected "hub" genes.

We use kernel density estimation to develop nonparametric estimators for the MI between gene expression levels and SNPs. The continuous variables are modeled as a mixture with conditional distributions for each level of the discrete variable, leading to a joint distribution with both continuous and discrete parts. We derive expressions for the information between the two using Shannon and Renyi entropies, showing that the model has a reasonable interpretation. Invariance properties of MI are also used to fit flexible parametric models. Our results show good agreement with previous analyses while incorporating additional information from the discrete SNPs.

Statistical analysis for firing neurons

Tony Pettitt
Queensland University of Technology
a.pettitt@qut.edu.au
Coauthors: Christopher Drovandi (Queensland University of Technology), James McKeone (Queensland University of Technology), Gareth Ridall (Lancaster University, UK)

This talk concerns statistical models for data derived from measuring neurons when they fire, usually a voltage in the form of an action potential. It gives a short introduction to the types of data available when considering sensory neurons and contrasts these with the types of data when considering motor neurons. For the latter case an important indicator of disease progress or, conversely , successful treatment of injured nerves, is the number of functional motor neurons serving a muscle. The talk will describe Bayesian statistical models to count this number, illustrating this with data from animal experiments and patients.

Risk-based trace priorities during disease outbreak

Joanne Potts
University of Melbourne
pottsj@unimelb.edu.au
Coauthors: Mark Burgman, Martin Cox

During an incursion of a pest or disease, the BioSIRT (Biosecurity Surveillance Incident Response and Tracing) software application can be used to prioritise trace events (i.e. movement of potentially infected/infested material between an infected area and another area). As currently implemented, the BioSIRT user specifies trace direction (forward, or backward), category of moved material (e.g. farm machinery), contact type (direct, or indirect) and date of movement relative to day zero (where day zero is the estimate of the earliest date of contact with the disease or pest). The trace is then automatically assigned a priority within BioSIRT, by matching any combination of these input variables with a look-up table, predefined by domain experts (e.g. epidemiologists). This project aims to develop transparent, structured models for assessing trace priorities with an emphasis on accountability. We present a spatially-explicit, stochastic, state-transition model, based on graph theory, where pest or disease spread occurs across a network of nodes (i.e. susceptible populations). Dispersal can occur via numerous mechanisms (e.g. infested propagation material). Various rules used to prioritise traces, and thus contain the disease, can be investigated via a simulation study. We parameterise the model for a citrus canker (Xanthomonas citri ssp citri) case study.

Maximizing MAXENT: Improvements to MAXENT through Poisson point process models

Ian Renner
University of New South Wales
Ian.Renner@unsw.edu.au
Coauthors: David Warton

MAXENT is a method of species distribution modelling (SDM) that has taken off in the ecology literature. A comprehensive study suggests that MAXENT outperforms nearly all other univariate SDM methods. But can we do even better? In this talk, I will demonstrate the equivalence of MAXENT and a Poisson point process model. I will then make use of this equivalence to explore improvements to MAXENT subsequently available through application of data-driven penalization and diagnostic tools to assess goodness-of-fit.

Designing multinomial experiments using the Integrated Mean Square Error criterion

Ken Russell
Charles Sturt University, Wagga Wagga NSW, Australia
kerussell@csu.edu.au
Coauthors: Gwenda Thompson, Australian Bureau of Statistics

Consider a multinomial experiment where the value of a response variable falls in one of k classes. Let π_{ij} represent the probability that the ith experimental unit yields a response that falls in the jth class. By modelling $\ln(\pi_{ij}/\pi_{i1})$ as a linear function of the values of m predictor variables, the results of the experiment may be analysed using a Generalized Linear Model.

It is common to construct such designs using the D-optimality criterion, which considers the covariance matrix of the unknown parameters in the linear function (e.g., Zocchi & Atkinson, *Biometrics*, 1999). Instead, we use the Integrated Mean Square Error criterion, which considers the properties of the predictors of the probabilities of falling in the k classes. This approach will be outlined and some examples presented.

Metaheuristic approach to the design of gene expression studies

Penny Sanchez
School of Mathematics and Statistics, University of South Australia, Mawson Lakes Campus, Mawson Lakes, SA, 5095
penny.sanchez@unisa.edu.au
Coauthors: Gary Glonek, University of Adelaide; Andrew Metcalfe, University of Adelaide.

Gene expression studies are aimed at investigating the behaviour of genes under a variety of conditions. Microarrays are a powerful technology that enables the investigation of many thousands of genes simultaneously. Utilisation of this technology has created the potential to make substantial advances in areas of bioinformatics. Rigorous experimental design is essential to make the most effective use of available resources in experimental situations. This presentation focuses on the search for optimal or near-optimal designs for large and complex comparative microarray experiments in cases where it is infeasible to carry out an exhaustive search of the design space. To do so, the metaheuristic approach of Pareto simulated annealing in the framework of response surface methodology is developed and applied in the microarray context. This employs a sample of generating designs that search the design space in an intellegent way based on the setting of appropriate tuning parameters that affect the performance of the approach. The approach will be demonstrated and discussed with relevance to gene expression studies aimed at making advances in the areas of medical and plant research.

Building Online Biostatistical Reporting Solutions with Business Intelligence Software: Tensions and Triumphs

James Scandol
NSW Department of Health, North Sydney 2060, Australia
james.scandol@doh.health.nsw.gov.au
Coauthors: Helen Moore, Lina Persson, Mark Cerny and Hanna Noworytko

The demand for online reporting solutions for both public and corporate needs has seen the development of sophisticated business intelligence software by large software companies. When using these tools for biostatistical applications, then the requirements for robust statistical reporting do not always marry easily to the architecture and functionality of business intelligence software. This presentation outlines these issues and presents the solutions that were developed for Health Statistics NSW (www.healthstats.doh.health.nsw.gov.au). The issues fell into three general groups: statistical analytics; security and privacy; and presentational flexibility (text, tables and charts). Statistical analytics were managed by continuing to do all analyses on internal SAS-based systems. Security and privacy issues were resolved by maintaining unit-level processing on highly secure internal systems. Presentational flexibility was resolved by developing a series of templates that enabled that rapid generation of statistically acceptable charts and tables. Acknowledging that these templates will not suit all purposes, various options for downloading data are also provided. High-performance and secure biostatistical reporting systems that can deliver text, charts and tables to hundreds of simultaneous users are complex to build and require significant resources to design, develop and test. Health Statistics NSW is an example of how such systems can be built using commercial business intelligence software whilst retaining statistical integrity. The final product is an application that is easy to use, enables users to find new information, encourages users to explore related data sets and allows users to create highly specialised reports.

The Kent regression model for compositional data

Janice Scealy
Centre for Mathematics and its Applications, Australian National University
janice.scealy@anu.edu.au
Coauthors: Alan Welsh (Australian National University, Australia)

Square root transformed compositional data can be modelled using the Kent distribution for directional data. The advantage of this approach is that it handles zero components directly and the covariance structure is not restrictive. In this talk we summarise a regression model based on the Kent distribution for modelling compositional data responses and we describe some properties of the model. To demonstrate this new modelling technique in practise, we analyse data containing compositions of foraminifer, a marine micro-organism measured at different depths. We show that the traditional modelling approach based on logratio transforms does not work well for this dataset. We also discuss some current and future research directions and briefly describe how to extend the Kent regression model to incorporate heteroscedasticity.

Tests for the Cox Model with data from a complex survey

Alastair Scott
Department of Statistics, University of Auckland
a.scott@auckland.ac.nz
Coauthors: Thomas Lumley

Time-to-event data are important in many areas, particularly the medical and social sciences, and complex probability samples are an increasingly common source of such data. These include multistage national surveys such as the National Health and Nutrition Examination Surveys' (NHANES) Linked Mortality Files, as well as stratified two-phase epidemiologic studies using case-cohort and complex case-control designs. The wide range of applications of survival analysis to survey data is illustrated by a Google Scholar search which produces more than 17,000 items containing both "NHANES" and "survival analysis".

Software to fit the Cox model to survey-sampled data is now widely available, using the weighted partial likelihood approach of Binder (1992) (see also Lin, 2000), and generalizations that make use of whole-cohort or whole-population information. One major gap in this software is the equivalent of the likelihood-ratio test and related model-selection quantities such as AIC and BIC. The main purpose of this talk is to fill this gap by developing an analogue of the partial likelihood ratio test for survey data.

We also look at methods for computing the asymptotic distribution and at ways of improving the small sample performance. We illustrate with examples using data from NHANES and from a stratified case-cohort study.

References

Binder, D.A.(1992). Fitting Cox's proportional hazards model from survey data. Biometrika, 79, 139-147. Lin, D.Y.(2000). On fitting Cox's proportional hazards model to survey data. Biometrika, 87, 37-47.

Quantifying the effect of sampling for biodiversity modelling

Hideyasu Shimadzu
Geoscience Australia
hideyasu.shimadzu@ga.gov.au
Coauthors: Scott D. Foster (CSIRO Mathematics, Informatics and Statistics)
Ross Darnell (CSIRO Mathematics, Informatics and Statistics)

Quantifying biodiversity is a challenge for answering scientific as well as conservation management questions. As biodiversity can only be assessed with biological samples collected by surveys, the modelling process therefore needs to incorporate how the adopted sampling scheme affects the biological samples. This talk focuses on quantifying the effect of a sampling technique widely used in marine surveys. We model the sampling process as random sampling from a multi-species composition using a multivariate hypergeometric distribution and quantify the effect using attenuation of species abundance distributions (SADs). This attenuation allows an appropriate statistical modelling of biodiversity on a conditional modelling framework that regards the observation as a deduction of the true biological quantity that we never observe. Our modelling approach is illustrated with data collected by a marine survey.

Controlling multiplicity in healthcare performance monitoring

Nokuthaba Sibanda
School of Mathematics, Statistics & Operations Research, Victoria University of Wellington
nokuthaba.sibanda@vuw.ac.nz

Statistical process control charts are widely accepted as a tool for monitoring quality indicators for hospital units and individual practitioners to ensure high quality patient care is achieved and maintained. In a given setting, quality of care may encompass multiple indicators that are simultaneously monitored, with each indicator on a separate chart.

Despite on-going debates, it is clear from a statisticians viewpoint that control charts form a sequence of hypothesis tests. Therefore, when multiple indicators are monitored simultaneously, control limits should ideally be adjusted for multiple testing. This problem is usually ignored, leading to drastically increased false alarms and unnecessary, sometimes costly, initiatives to resolve problems that do not really exist.

I will describe various approaches for controlling the overall error rate for a family of p charts used to monitor the quality of care in a maternity unit. Indicators used include key maternal and infant outcomes, some of which are correlated and together can be used to test the hypothesis All is well in the maternity unit. I will explore approaches for controlling family-wise error and false-detection rates. Results of simulations used to explore Average Run Lengths using the two approaches will be presented.

What types of climate measurements best predict the distribution of biodiversity?

Eve Slavich
University of New South Wales
eve.slavich@student.unsw.edu.au
Coauthors: David Warton, Daniel Ramp, Michael Ashcroft, John Gollan

What is the most ecologically relevant way to measure climate? Community-level modelling is increasingly used to inform conservation management and investigate the effects of climate change, by predicting the distribution of an assemblage of species from a climate surface. There are a number of competing methodologies for inferring the climate surfaces. For example, popular methods often use weather station data to produce climate surfaces from elevation and location information (e.g. BIOCLIM) while recently developed methods use field based measurements of climate and a wider range of climate predictors (e.g. cold air drainage) to produce a climate surface. It is currently unclear how these different surfaces perform at predicting the distribution of communities of species. To address this, in this talk I compare the performance of these two types of climate surfaces at modelling a community of ferns at fine scales in the Hunter Catchment area of NSW, using multivariate extensions of generalised linear models.

Change-point detection in DNA copy number variants

Georgy Sofronov
Macquarie University
georgy.sofronov@mq.edu.au

Recent biological studies show the close relationship between chromosomal regions aberrant in copy number and diseases like cancer, mental retardation and diabetes. Therefore, identifying genomic regions associated with systematic aberrations provides insights into the initiation and progression of a disease, and improves the diagnosis, prognosis and therapy strategies. With more and more large datasets emerging, there is a need for efficient algorithms that automatically detect change points and the same time provide some estimate of error for this detection process. In this talk, we consider various approaches to change-point detection in DNA copy number variants, using Monte Carlo simulation to find estimates of change-points as well as parameters of the process on each segment.

Pedigrees in the Analysis of Yield Protein Relationship for Multi-Environment Lupin Breeding Trials

Katia Stefanova
The UWA Institute of Agriculture, UWA, 35 Stirling Highway, Crawley WA 6009
katia.stefanova@uwa.edu.au

Multiplicative mixed models are routinely used for the analysis of multi-environment trial (MET) data. Recent papers by Beeck et al (2010, Genome, 53, 992-1001), Cullis et al (2010, Genome, 53, 1002-1016) and Piepho et al (2008, Euphytica, 161, 209-228) discuss the inclusion of pedigree information when analyzing traits as yield and oil content. Oakey et al (2007, Theor Appl Genet, 114,13191332) suggested decomposition of the total genetic effects into additive, dominance and residual non-additive components in the context of factor analytic model applied to a MET data set (Smith et al, 2005, J Agr Sci 143, 449-462). In this paper, we utilize the above approaches while investigating the yield protein relationship for a set of 6 lupin Angustifolius (narrowleaf lupin) breeding trials at 3 environments for the period 2008-2009. The analysis adjusts for the spatial variation in the field, models the variance structure of the genotype by environment (GxE) interaction effects, incorporating the pedigree information for both traits, yield and protein.

Behaviour of higher criticism and competing tests for sparse normal mixtures near the detection boundary

Michael Stewart
University of Sydney
michael.stewart@sydney.edu.au

Consider testing that a sample is standard normal against the alternative that it is from the (mixture) distribution of $Z + I\mu$, where Z is standard normal independently of a Bernoulli(p) random variable I. This apparently simple parametric test of $p = 0$ or $\mu = 0$ has deceptively complicated asymptotic properties and work on elucidating them goes back at least to the seminal paper of Hartigan (1986). In recent times "sparse" versions of this scenario have been used as models for large-scale multiple testing problems where only a small proportion $p \approx n^{-\beta}$, $0 < \beta < 1/2$ of n independent hypotheses being tested are false, for which μ then represents a common effect size.

John Tukey proposed a method of assessing the significance of a body of tests by firstly standardising the empirical CDF of the P-values (as if they were all uniform on (0,1)) and then looking at the maximum of this process over a prespecified interval in the neighbourhood of zero. He called this procedure "higher criticism" and its power properties were analysed for the mixture testing problem above by Donoho and Jin (2004). They showed that if $\mu = \sqrt{2r\log(n)}$ then a "detection boundary" $\rho(\beta)$ exists; if $r > \rho(\beta)$ then the limiting power is 1 but if $r < \rho(\beta)$ the limiting power is simply the level of the test. They also pointed out that the likehood ratio test for this problem has the same detection boundary.

We provide a higher-order analysis which explains what happens at the detection boundary itself, both under the basic model above and a more realistic variant where the false hypotheses can have different effect sizes. In particular we show that the tests have different higher-order power properties.

Confidence interval construction for disease prevalence based on partial validation series

Man-Lai Tang
Department of Mathematics, Hong Kong Baptist University
mltang@math.hkbu.edu.hk
Coauthors: Qiu, Shi-Fang and Poon, Wai-Yin

It is desirable to estimate disease prevalence based on data collected by a gold standard test, but such tests are often limited due to cost and ethical considerations. Data with partial validation series thus become an alternative. The construction of confidence intervals for disease prevalence with such data is considered. A total of 12 methods, which are based on two Wald-type test statistics, score test statistic, and likelihood ratio test statistic, are developed. Both asymptotic and approximate unconditional confidence intervals are constructed. Two methods are employed to construct the unconditional confidence intervals: one involves inverting two one-sided tests and the other involves inverting one two-sided test. Moreover, the bootstrapping method is used. Two real data sets are used to illustrate the proposed methods. Empirical results suggest that the 12 methods largely produce satisfactory results, and the confidence intervals derived from the score test statistic and the Wald test statistic with nuisance parameters appropriately evaluated generally outperform the others in terms of coverage. If the interval location or the non-coverage at the two ends of the interval is also of concern, then the aforementioned interval based on the Wald test becomes the best choice.

High dimensional and random whole genome average interval mapping

Julian Taylor
School of Agriculture, Food and Wine, The University of Adelaide
julian.taylor2@gmail.com
Coauthors: Ari Verbyla

Whole genome average interval mapping (WGAIM) is an approach for QTL analysis proposed by Verbyla, Cullis and Thompson (2007). All intervals on a linkage map are included simultaneously in the analysis as a simple random effect. A forward selection approach is used with selected QTL placed in the fixed effects part of the model. WGAIM has been implemented in an R (R Development Core Team, 2011) package wgaim (Taylor et al, 2011); see also Taylor and Verbyla (2011). Two issues arise in this method. The first issue is selection bias whereby estimated QTL effects are inflated. A random effects version of WGAIM is proposed and is shown to reduce the bias. The second issue involves situations where the number of intervals exceeds the number of observations and computation becomes expensive. An approach is outlined that reduces the dimension to the number of genetic lines in the data being analysed. Both the random WGAIM formulation and the dimension reduction technique have been implemented in the wgaim package.

References

R Development Core Team (2009) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org, ISBN 3-900051-07-0

Taylor JD, Diffey S, Verbyla AP, Cullis BR (2011) wgaim: Whole Genome Average Interval Mapping for QTL detection using mixed models. R package version 1.00-1

Taylor, JD and Verbyla, AP (2011) R package wgaim: QTL analysis using complex linear mixed models. Journal of Statistical Software 40(7), 1-18. http://www.jstatsoft.org/v40/i07/

Integrating Graphics and Analyses for Microarray Data

Antony Unwin
University of Augsburg, Germany
unwin@math.uni-augsburg.de
Coauthors: Alex Gribov (University of Augsburg)

Microarray data are often graphically presented by heatmaps. This is only effective if the rows and columns are ordered, using clustering or seriation methods. Deciding which of many methods to use is dependent on computational power (many of the methods only work on small datasets) and on the value of the results obtained. Experimenting with alternatives, comparing the results, and drawing conclusions, all benefit from a close integration of graphics and analyses. For instance, comparing clusterings in the graphical display of a confusion matrix requires a sorting algorithm to reveal the structure, and linking to other displays is very useful for explaining that structure. The ability to work with multiple linked graphics in parallel with direct access to analytic methods is important for flexible, exploratory analyses. How these demands may be met in software and what sort of tools can be combined, are the subject of this paper. The talk will be illustrated using Alex Gribovs interactive graphics software SEURAT (seurat.r-forge.r-project.org), which uses analytic methods available in R via Simon Urbaneks Rserve.

Key words: Microarrays, clustering, SEURAT, Rserve

The Swings and Roundabouts of Methods and Models for Genomic Selection

Klara Verbyla
Mathematics, Informatics and Statistics, CSIRO, Australia
klara.verbyla@csiro.au
Coauthors: Ari Verbyla (Mathematics, Informatics and Statistics, CSIRO, Australia)

Genomic Selection (GS) is a selection technique that utilises genome-wide DNA markers to improve the accuracy of selection for quantitative traits. GS is currently revolutionising animal breeding with a saturation of studies demonstrating the advantages of GS over traditional selection methods. This paired with the dramatic reduction in genotyping costs has lead to the commercial implementation of GS in dairy cattle.

Fundamental to the success of GS has, and continues to be, the development of statistical methods to allow accurate prediction of breeding values. The major challenge of genomic prediction is to accurately model the true quantitative trait loci (QTL) effects. This challenge is made difficult by disparity between the large number of markers (p) and the number of records (n) that are available to estimate the marker effects i.e. $p > n$.

Due to the interest of plant breeders in implementing GS, I review the methods and models used for GS and examine the challenges of successfully implementing GS for plant species.

Performance of model selection criteria in longitudinal data analysis

You-Gan Wang
The University of Queensland
you-gan.wang@uq.edu.au
Coauthors: Lin-Yee Hin and Vincent Carey

Selecting an appropriate working correlation structure is pertinent to longitudinal data analysis because an inappropriate choice will lead to inefficient parameter estimation. We investigate a selection of criteria including QIC, CIC and the Gaussian likelihood. Extensive simulation studies indicate that the CIC has remarkable improvement to QIC and the simple Gaussian likelihood also works well for nonnormal data in selecting the correct correlation structures. We illustrate our findings using some real data sets.

Advances in species distribution modelling in ecology

David Warton
University of New South Wales
David.Warton@unsw.edu.au
Coauthors: Alex Brown, Ian Renner, and Luke Wilson

Species distribution modelling (SDM), where one models the likelihood of occurrence of a species as a function of a suite of environmental variables, has received a lot of recent attention in ecology. In fact, ISI Essential Science Indicators lists SDM as the fastest moving research front in the environmental sciences. This talk will review some recent contributions to the SDM literature from the UNSW Eco-Stats group: using point process models for presence-only data; a model-based approach to account for observer bias; using the LASSO to "borrow strength" across species; incorporating traits into models to explain differing environmental response across species. Future directions for research will also be outlined.

Analysing Occupancy Surveys

Alan Welsh
The Australian National University
Alan.Welsh@anu.edu.au
Coauthors: C.F. Donnelly, D.B. Lindenmayer

Occupancy surveys are surveys of sites which are designed to collect data for the purpose of either estimating the proportion of sites that are occupied or modelling the probability that a site is occupied by a particular animal species. The standard approach (MacKenzie et al, 2006, Occupancy Estimation and Modelling: Inferring Patterns and Dynamics of Species Occurrence) recognises that the binary occupancy variable may be measured with error, assuming particularly that a site may be occupied without the species being detected on the site. This situation is handled by assuming that the occupancy status of a site does not change over a period of time and then making multiple visits to at least some sites over that period. Data collected in this way can then be used to model the probability of detection and hence to adjust the estimates of occupancy for potential nondetection. In this talk, we will discuss some aspects of the analysis of occupancy surveys. In particular, we will discuss the effect of different designs and clarify some issues of model identifiability. We will present the analysis of some bird data and discuss the interpretation of the results.

The importance of carry-over effects in experimental design and analysis

Emlyn Williams
Statistical Consulting Unit, Australian National University
emlyn.williams@anu.edu.au
Coauthors: Jun Imaki, Angeline Tjhin, Antonia Vincent

In many situations multiple measurements are made on experimental material. Often it is simply repeat measurements in time. But in other cases different treatments are successively applied, for example in a lactation experiment different diets may be given to cows over several time periods. In the latter situation there is the possibility of a carry-over effect from the previous treatment. Experimental designs that accommodate such effects are called crossover designs and they are used extensively in practice.

In this talk I will discuss a range of crossover design types, such as first and second order additive, self-adjacency and placebo models. I will give two examples of crossover designs that have been used to spectacular effect. The first case is an experiment run in the Linguistics Department at ANU involving the evaluation of essays by assessors. The second case comes from the Psychology Department at ANU and investigates the response time of primary school children to different arithmetic tasks.

On Detection of Differential Expression using RNA-Seq Data

Susan Wilson
ANU & UNSW
sue.wilson@anu.edu.au

Ultra high-throughput sequencing of RNA (RNA-Seq) has emerged as a powerful technology for profiling transcript abundances. Its main advantage is its ability to profile the transcriptome directly, and so no prior knowledge of the queried transcriptome is needed. The data produced by RNA-Seq are abundance counts. Two widely used R packages for determining differential expression for such data are edgeR and DESeq. Using recently published data we have explored the similarities and differences between the effects of the different assumptions made by these packages. Currently this technology is relatively expensive. So biologists need to find the balance between having (i) greater sequencing depth for each sample and (ii) more replicates at reduced sequencing depth. Such choices affect detection of differential expression for those transcripts that are less strongly expressed. Using real and simulated data, these experimental design issues are explored.

Mapping multiple quantitative traits using Structural Equation Models

Lisa Woods
Victoria University of Wellington
woodslisa1@myvuw.ac.nz
Coauthors: Nokuthaba Sibanda

Quantitative traits are continuous physical properties displayed by an organism, such as yield, which are influenced by regions of the genome known as quantitative trait loci (QTL). The identification and mapping of QTL is of interest to geneticists and breeders looking to select for particular traits.

Multiple trait mapping is useful as it examines the correlations between traits such that, unlike single trait analysis, it allows the user to fit more complex and accurate biological models. Many multiple trait mapping methods have been developed, for example: Seemingly Unrelated Regression, Structural Equation Modelling (SEM), composite interval mapping for multiple traits.

SEM has an advantage over other multiple trait methods in that it allows incorporation of the causal structure. Other methods only work to map QTL and test for pleiotropy, while incorporating correlations between traits. By incorporating the causal structure, direct and indirect QTL effects on each trait can be estimated, thus allowing more accurate inferences of the genetic architecture to be made.

In most analyses that use SEM, the causal structure is assumed to be known a priori. We will investigate the use of a Bayesian mixture SEM to infer the causal structure, effects and location of QTL that influence multiple traits. We use simulated data to explore the accuracy of our method under various scenarios, including the effect of potentially confounding environmental factors.

Comparison of ANOVA, Tobit model and Two-part model for analysing sensory data

Hwan-Jin Yoon
Statistical Consulting Unit, The Australian National University, Canberra, ACT 0200, Australia
hwan-jin.yoon@anu.edu.au
Coauthors: Alan Welsh Centre for Mathematics and its Applications, Mathematical Sciences Institute, The Australian National University, Canberra, ACT 0200, Australia

Emlyn Williams Statistical Consulting Unit, The Australian National University, Canberra, ACT 0200, Australia

In designed experiments, we face the zero-inflated data from time to time. One area in which data commonly occur is Food Science. Often, sensory data in food science contain a large number of zeroes, due to the limits scale used. Therefore, in this case, ANOVA may not be a good choice. If the observed zeroes are due to censoring rather than true zeroes, the Tobit model (Tobin, 1958) might be used. Alternatively, a two-part model can be applied for analysing zero-inflated data. Using Pangborn sensory dataset (first replication only), Marin-Galiano and Kunert (2006) compared the ANOVA and the Tobit model using permutation tests and concluded that ANOVA is better suited than the Tobit model to analyse sensory data: (1) ANOVA keep the test level whereas the Tobit model fails to keep the test level most of the time except for the case of high amount of zeroes and (2) ANOVA is much easier to implement and better known. Does the permutation test really work for the zero-inflated data? Guillet et al. (2001) state that the Tobit model is a generalization of ANOVA. They conclude that the Tobit model is better for analysing sensory data. We redo the analyses using complete Pangborn sensory dataset and compare the results between ANOVA, Tobit model and the two-part model.

Two optimization strategies of multi-stage design in clinical proteomic study

Irene Zeng
Department of statistics, Univerisity of Auckland
zeng@stat.auckland.ac.nz
Coauthors: Thomas Lumley, Kathy Ruggiero, Ralph Stewart

In the majority of the current reported proteomic studies, laboratory selections and clinical validation of the protein markers were two separated processes in study design. It has been suggested[1] that the lack of success in translation is due to a lack of connection between laboratory proteomics and clinical proteomics. The National Cancer Institute (NCI) suggested a three-stage workflow of clinical proteomic study in order to drive the laboratory discovery to clinical utility. Following the NCI suggestion, we propose an optimized multi-stage systematic design for clinical proteomic study. We compute the operating characteristics of the multistage study as a function of sample sizes and nominal Type-I error rates at each stage, and then optimize over these parameters to find the study with greatest expected number of true discoveries under constraints on cost and false discoveries. A simulated annealing algorithm is used to find the optimal solution in the defined region. We show that this approach is feasible and leads to efficient designs. We also investigated the use of biological grouping information in the first stage of selection, and found improved performance when the grouping is informative, with little loss in performance when the grouping is uninformative.

[1] Department of Statistics, Department of Medicine, University of Auckland. Report from the Wellcome Trust/EBI "Perspectives in Clinical Proteomics" retreat - A strategy to implement Next-Generation Proteomic Analyses to the clinic for patient benefit: Pathway to translation. Proteomics Clin. Appl. 2010, 4, 883-887.

Posters

Distributional properties of Harvest Index

Delma Greenway
School of Agriculture and Food Sciences, The University of Queensland, Brisbane 4067
del.greenway@uq.edu.au
Coauthors: Olena Kravchuk (School of Agriculture, Food and Wine, The University of Adelaide, Adelaide 5000; School of Agriculture and Food Sciences, The University of Queensland, Brisbane 4067)

Improvement in harvest index (HI), calculated as the ratio of grain yield (Y) to above ground biomass including grain (X+Y), has been a major focus in breeding experiments of small grain crops over the past 50 years. Harvest index can be interpreted as the proportion of the biomass that is converted into grain, HI = (1+X/Y)-1. At maturity (X, Y) is a binormal variable with parameters (μX, μY, σX, σY, ρ) . The distribution of HI depends on the means (μ) and standard deviations (σ) of the primary variables and ρ, their correlation coefficient. Such distribution is not normal, exhibiting skewness and heavy tails. Various normalising data transformations have been applied in grain research with log transformations being the most common. However, there has been little focus on a distributional analysis of the index. At the same time, harvest index has been shown to be crop specific, resulting in the need to examine its parameters on a crop by crop basis. In this work, we draw the attention of the reader to the sensitivity of the parameters of the distribution of HI to the correlation between the grain yield and the rest of the above ground biomass, and the need to take this into account when designing studies concerned with the estimation of the index.

Characterisation of neural ensemble activity with a stochastic point process framework

Maurizio Manuguerra
Statistics Department, Macquarie University
maurizio.manuguerra@mq.edu.au

In this work the functional relations among a set of neurons recorded in-vivo have been studied. Neural recordings, regarded as stochastic point processes, can be characterised by their conditional intensity function, a generalisation of the rate function of a Poisson process, and analysed in the GLM framework. The aim of this study is to estimate the intensity of the relations and the temporal distances between the recorded neurons. This information is crucial to a few theories, and in particular to the theory of polychronous groups by Izhikevich, that is the base for further research in development. As the point process likelihood function is based on a discrete time representation, the distances between neurons are discrete parameters and then the parameter space is given by the Cartesian product of a subset of a finite dimensional Euclidean space and a finite set. This poses a few theoretical and numerical challenges, whose solution is the object of this study.

Loss of Superior Genotypes in Early Stages of Sugarcane Varieties Selection in Kenya

Peter Maina Wachira
Kenya Sugar Research Foundation, Kisumu, Kenya
peter.maina@kesref.org
Coauthors: E. Onginjo (Kenya Sugar Research Foundation, Mtwapa, Kenya), E. Ndindi (Mumias Sugar Company, Kenya) and R. Simwa (University of Nairobi, Kenya)

Sugarcane selection is used for the identification and release of elite varieties from the original seedling population. In Kenya, only 9 varieties that outperform the checks have been released since 1980. The varieties are tested in five stages with stage 1 being the original seedlings population consisting of about 35,000 individual seedlings. Moderate selection intensity is performed among families and individual seedlings. This results to about 3,500 (10%) seedlings being selected to stage 2 where the test varieties are tested on a single site single row trial. We hypothesized that the large number of varieties tested in stage 2 results to loss of superior genotypes. To determine the extent of this loss we selected 30 elite varieties from MS 2008 seedling stages for testing in 3 sites on augmented row-column design with single row plot. The 30 elite varieties were also included in stage 2 of the current selection program. Data was analysed and the elites varieties that significantly out-performed the checks were compared with those advanced to stage 3 in the current selection program. The results shows that the current selection method results to loss of superior genotypes since only 4 varieties were advanced to stage 3 out of 12 elite varieties that outperformed the checks. We evaluated analysis of stage 2 data by including data obtained in stage 1 which resulted to improved efficiency with 9 out of 12 elite varieties being selected. We propose the use of Bayesian approach in analysis of sugarcane varieties selection data.