61^{st} Annual Conference of the New Zealand Statistical Association

held in conjunction with the

Statistical Methodologies Symposium celebrating the work of Professor Chin-Diew Lai

MASSEY UNIVERSITY

Palmerston North, New Zealand
Tuesday 29 June to Thursday 1 July, 2010

Jonathan Godfrey (Editor)
June 30, 2010

Contents

Welcome Message **4**

Organizing Committee **5**

Sponsors **6**

Speakers (in alphabetical order) **7**

Ali, Abdul 7
Alzaid, Abdulhamid 9
Anwar, Nafees 10
Ardalan, Arash 11
Aubry, Jean-Marie 12
Balakrishnan, N. (Bala) 13
Barton, James 14
Bebbington, Mark 15
Betz-Stablein, Brigid 16
Bilton, Penelope 17
Chan, Ping Shing Ben 18
Chee, Chew-Seng 19
Chen, Chen 20
Cheng, Ching-Shui 21
Cook, Len 22
Costilla, Roy 22
Davis, Walter 23
Ehlers, René 24
Fernando, Sarojinie 25
Filus, Lidia 26
Ganesalingam, Ganes 28
Haslett, Steve 29
Haywood, John 30
Hazelton, Martin 31
Jones, Geoff 32
Kachapova, Farida 33
Kale, Hazel 34
Khoo, Michael 35
Lai, Chin-Diew 35
Lin, Gwo Dong 36
Mao, Tian 36

McGirr, Rebecca & Hawkes, Tim 37
Miller, Arden 38
Nagatsuka, Hideki 39
Namay, Rico 41
Ng, Hon Keung Tony 42
Noble, Alasdair 43
Olkin, Ingram 44
Ong, Hong Choon 45
Ong, Seng Huat 46
Richens, Andrew 46
Rodado, Armando 47
Rohan, Maheswaran 48
Sampson, Allan 49
Scott, Alastair 50
Tang, Boxin 51
Tularam, Anand & Roca, Eduardo 52
Turner, Rolf 53
van Koten, Chikako 54
Walker, Lyndon 55
Wang, Ting 56
Wang, Yuancheng 57
Westbrooke, Ian 58
Willink, Robin (1 of 2) 59
Willink, Robin (2 of 2) 60
Wood, Graham 61
Yee, Thomas 62
Zheng, Guan Yu (Fish) 63
Zitikis, Ričardas 64

Biography of Professor Chin-Diew Lai **64**

Programme **65**

Welcome Message

The organizing committee for the 61^{st} Annual Conference of the New Zealand Statistical Association wish to welcome delegates to Palmerston North and Massey University in particular. We hope that you enjoy the varied talks and a relaxed atmosphere that encourages networking with colleagues from around New Zealand and overseas.

As this conference also celebrates the career of Massey University's own Professor Chin-Diew Lai, we have been very fortunate to attract a large number of delegates from overseas this year. We have speakers from universities in more than ten different countries, but digging a little deeper shows us that these people come from an even broader range of countries.

Please do not hesitate to ask the locals about what Palmerston North and New Zealand have to offer you while you are here.

Alasdair Noble	Jonathan Godfrey
Organizing Committee Chair	Programme Chair

Organizing Committee

The following people formed the organizing committee for this conference:

Chair:	Alasdair Noble	Institute of Fundamental Sciences, Massey University
Programme:	Jonathan Godfrey	Institute of Fundamental Sciences, Massey University
Committee:	Mark Bebbington	Institute of Fundamental Sciences, Massey University
	Brigid Betz-Stablein	Institute of Fundamental Sciences, Massey University
	Colleen Blair	Institute of Fundamental Sciences, Massey University
	Siva Ganesh	Institute of Fundamental Sciences, Massey University
	Duncan Hedderley	Plant & Food
	Chin-Diew Lai	Institute of Fundamental Sciences, Massey University
	Andrew McLachlan	Plant & Food
	Zaneta Park	AgResearch

Special thanks are owed to Professor Balakrishnan who helped promote the conference to many of our international speakers.

Our Sponsors

We are grateful for the financial assistance offered by our sponsors:

MASSEY UNIVERSITY

Taylor & Francis Group
an informa business

Speakers (in alphabetical order)

Ali, Abdul

Ko te Anga te Hoto — Structure is the Link

Abdul Ali
School of Computing and Information Technology,
Manukau Institute of Technology

The link between Statistics and Computing needs to be explored within the larger framework of Information Structures, which should include at least Mathematics and importantly Philosophy as well. Particular links may be good or bad. Data Modelling and Data Analysis exist in both Statistics and Computing. The aim would be to have a generic database and hence a generic module for data entry with various modules for manipulation. This would seem to be possible because Grouping Variables in Statistics and Primary Keys in Computing have similar functions. The link between Statistics and Computing is obscured because of unconnected nomenclature. There should be a deliberate convergence in nomenclature, not only between these two disciplines but also within Information Structures generally. This would require a regulatory body. The testing of software has occasioned the application of Statistics to Computing. Various factors might be operating system, type of data file, type of display, and different hardware components. The use of statistical experimental design to limit the number of combinations tested fails to appreciate fully the philosophical framework for such designs. These were developed in an agricultural context where continuity existed between the experimental units themselves as well as frequently between the levels of treatment factors. However one would think the effective continuity between various operating systems for example would be minimal. This would be the position with other factors as well. Thus the use of efficient designs is inappropriate, as interpolation (let alone extrapolation) between levels of a factor does not have a physical basis. The way ahead would seem to be to employ the concept of orthogonality as then the number of treatment combinations required to be individually tested grows in a linear rather than multiplicative fashion. This approach would require common standards in the computing industry and hence a regulatory body. The opportunity then could be taken to reorganise the way binary is interpreted by the computer. ASCII was developed with the English language in mind. Unicode is an extension of this and treats each writing system in isolation. All languages could be coded using a minimal number of common characters based

on phonetics. This may be achieved without common glyphs, although that could be a desirable later development.

Alzaid, Abdulhamid

Binomial Difference Distribution

Maha A. Omair
King Saud University, Riyadh, Saudi Arabia

Abdulhamid Alzaid
King Saud University, Riyadh, Saudi Arabia

Om Alssaad Aodah
King Saud University, Riyadh, Saudi Arabia

The difference of two independent binomial distribution having the same number of trials is defined. Distributional properties of the binomial difference distribution were derived. We obtained moment estimators and Maximum likelihood estimate and compared them via simulation study. Hypothesis testing using likelihood ratio test was considered.The binomial difference distribution is useful in modeling the change in number of events In this paper we use the binomial difference distribution to model the change on the price a stock share from the Saudi stock share market and the change in the bed occupancy of the nursery intensive care unit.

Anwar, Nafees

Measurement and Visualization of Data Complexity for Classification Problem

Nafees Anwar
Massey University, Palmerston North

Siva Ganesh
Massey University, Palmerston North

Geoff Jones
Massey University, Palmerston North

Ganes Ganesalingam
Massey University, Palmerston North

Visualization of high dimensional data sets traditionally focuses on graphical representation of information in lower dimension. They tend to provide an aid to users in tackling complex knowledge discovery task. Users usually face several problems such as: 1) hard to discover valuable information when too much data is visualized, 2) discoveries based on the visual exploration alone may lack accuracy and 3) they have no convenient access to the important knowledge learned by the other users. To tackle these problems it is recognized that analytical tools must be used into visualization system.

In this presentation, we use a nearest neighbour approach for characterizing the complexity of classification problems. We studied the comparative advantages of two methods, Euclidean distance and proximity measure by Random Forest, for constructing nearest neighbours. We investigated a collection of two-class problems from the UCI repository, and observed that there are strong correlations between classifier accuracies and class overlap. The experiments also demonstrated that the similarity measure by Random Forest is compatible with Euclidean distance for numerical data but has obvious advantages for data consisting of categorical or mixed type of variables where Euclidean distances cannot be computed directly.

We used Multi dimensional scaling to visualize dissimilarity in data sets (in low dimension), and to see the effectiveness of our proposed technique. These insights, hopefully, help us to understand imbalanced data sets and to devise a promising approach in comparison to other standard approaches that deal with imbalanced data sets. An Evaluation of Mahalanobis Taguchi System (MTS) for imbalance data sets was also carried out, and various results are reported in this presentation.

Ardalan, Arash

A Generalized Normal-Laplace Distribution: Properties, Estimation and Applications

Arash Ardalan
Shiraz University, Iran

S.M. Sadooghi-Alvandi
Shiraz University, Iran

A.R. Nematollahi
Shiraz University, Iran

H.A. Mardanifard
Shiraz University, Iran

In this article we are focused on the Two-Pieces Normal-Laplace (TPNL) distribution, an important member of Asymmetric Exponential Power Distributions (AEPD) proposed by Zhu and Walsh (2009). TPNL distribution is a split skew distribution consists of a normal part (short tail) and a Laplace part (heavy tail). We have independently, from Zhu and Walsh (2009), generalize this kind of distribution to three parameters, which is more flexible and can be applied in many statistical theories such as centile regression and etc. An algorithm for computing maximum likelihood estimators is given. Consistency and asymptotic normality of maximum likelihood estimators of the parameters, which has not been considered by Zhu and Walsh (2009), are established. Properties of this distribution along with Bayesian calculations are presented. Applications are made and flexibility of this distribution, in comparing to other recently introduced asymmetric normal or Laplace distributions, along studying three real data examples and a simulation study, is discussed.

Reference: Zhu, D. and Zinde-Walsh, V. (2009). "Properties and estimation of asymmetric exponential power distribution." *Journal of Econometrics* **148**, 86–99.

Aubry, Jean-Marie

Large Deviations for Quasi-arithmetically Self-normalized Random Variables

Jean-Marie Aubry
University of Paris East, France

Marguerite Zani
University of Paris East, France

We introduce a family of convex (concave) functions called sup (inf) of powers, which are used as generator functions for a special type of quasi-arithmetic means. Using these means we generalize the large deviation result that was obtained in the homogeneous case by Shao (1997) on self-normalized statistics. Furthermore, in the homogenous case, we derive the Bahadur exact slope for tests using self-normalized statistics.

Reference:

Shao, Q.-M. (1997) "Self-normalized large deviations". *Ann. Probab.* **25(1)**, 285–328.

Balakrishnan, N. (Bala)

Some Cure Rate Models and Associated Inference and Application to Cutaneous Melanoma Data

N. Balakrishnan
McMaster University
Hamilton, Ontario, Canada

In this talk, I will first describe the basic cure rate models and then explain their limitation. I shall introduce some new cure rate models based on weighted Poisson distributions and explain some of their properties and features. Next, I will discuss the associated inferential procedures.

Finally, I will illustrate the use of these models by applying them to a cutaneous melanoma dataset, and discuss the findings and their significance.

Professor Balakrishnan's attendance is generously supported by our principal sponsor:

Barton, James

Uncertainty in New Zealand's Greenhouse Gas Inventory

James P. Barton
James P Barton & Associates,
Wellington

The annual submission of national greenhouse gas (GHG) inventories is a requirement for those nations who have ratified the Kyoto Protocol. These inventories are a compilation from various sources of sectoral estimates of GHG emissions and removals. Some sectors are inherently more complex to estimate emissions and removals for than others and rely on models for the annual estimates.

A requirement of the Kyoto Protocol is that countries follow good practice guidance and seek to reduce uncertainties in the estimates as better information becomes available.

This talk traces the ways in which uncertainties in New Zealand's GHG inventory have been reducing as better information becomes available. It specifically discusses methods used to reduce and estimate uncertainties in the key Land Use, Land-use Change and Forestry (LULUCF) sector for New Zealand.

Bebbington, Mark

Analyzing Treatment Effects on Distributions with Complex Structure

Mark Bebbington
Massey University, Palmerston North

Marcel Voia
Carleton University

Ričardas Zitikis
University of Western Ontario

Comparing treatment effects using, for example, the average treatment value is not particularly informative about specific regions of the treatment and control distributions, especially when the distribution of the outcome variable of interest is complicated in nature. For this reason, we consider a suite of nonparametric tests based on entire distributions, which provide a more informative analysis of treatments and their effectiveness. The performance of the tests is illustrated on one treatment from the Pennsylvania Bonus Experiment (PBE), and the approach compared with one involving detailed parametric modelling of the distribution.

Betz-Stablein, Brigid

Disease Mapping Techniques Applied to Glaucoma Visual Field Datasets

Brigid D. Betz-Stablein
Massey University, Palmerston North

Martin L. Hazelton
Massey University, Palmerston North

William H. Morgan
Lions Eye Institue, Perth, Western Australia

Glaucoma is the second leading cause of blindness in the world. It is caused by a buildup of fluid in the eye, which can lead to irreparable vision loss. Glaucoma severity can be measured by testing a subject's visual field over a grid like structure, and the progression of the disease determined by studying sequences of visual field test results taken over time. By borrowing tools from the disease mapping literature, we develop models for longitudinal sets of visual field data that account for the correlation structure generated by the spatial configuration of visual field measurements across the retina. We describe model fitting using MCMC methods. Preliminary results are very encouraging — our approach appears to be able to identify clinically significant glaucoma progression earlier than is possible using existing methods.

Bilton, Penelope

A Statistical Model to Characterize the Naphthenic Acid Component of Petroleum

Penelope Bilton
Massey University, Palmerston North

Martin L. Hazelton
Massey University, Palmerston North

Peter J. Derrick
Massey University, Palmerston North

Naphthenic acids are organic acids found particularly in heavier types of crude oil. The corrosive effects of these substances are well documented, but in more recent times the environmental risk they pose has also become evident. The toxicity of naphthenic acids is related to their molecular structure, but characterization of individual acids is difficult because the naphthenic acid component of petroleum can comprise a complex mix of hundreds of different molecules. Fourier transform mass spectrometry (FT-MS) has proved the best tool to distinguish between individual naphthenic acids, since it provides identification of explicit chemical formulae. A complicating factor in characterizing naphthenic acids is the process of dimerization, the aggregation of acid monomers to form dimers. The research task is to develop a statistical method for identifying individual acid monomers when dimerization has occurred. Since this is an ill-posed problem the technique of linear inverse modelling is applied to FT-MS data to investigate possible models for the formation of naphthenic acids dimers.

Optimal Allocation for Multi-level Stress Testing with Extreme-value Regression

Chan, Ping Shing Ben
The Chinese University of Hong Kong

Ng, Hon Keung Tony
Southern Methodist University, Texas, USA

Ka, Cheuk Yin
The Chinese University of Hong Kong

Balakrishnan, N.
McMaster University, Canada

Optimal allocation problems in a Weibull (extreme value) multi-level regression model under Type I and Type II censoring are discussed. The maximum likelihood estimators and their Fisher information and asymptotic variance-covariance matrix are derived. Different optimality criteria are used to discuss the optimal allocation problem. Optimal allocation of units, both exactly for small sample sizes and asymptotically for large sample sizes, for two- and four-stress-level situations are determined numerically. Conclusions and discussions are provided based on the numerical studies.

Chee, Chew-Seng

Mixture-based Nonparametric Density Estimation: Maximum Likelihood vs. Least Squares

Chew-Seng Chee
University of Auckland

Yong Wang
University of Auckland

Mixture models constitute a flexible family of statistical distributions and are potentially very useful for nonparametric density estimation. In this talk, we describe how to use mixture models for nonparametric density estimation that is developed based on the maximum likelihood or the least squares method. We will contrast the performance of these two approaches using different performance criteria, and compare them with the popular kernel-based density estimator.

Confidentiality for the 2011 Census: Statistical Thinking Applied

Chen Chen
Statistics New Zealand, Christchurch

Mike Camden
Statistics New Zealand, Wellington

Statistics New Zealand is reviewing its census confidentiality methods in preparation for the 2011 Census. We considered simplifications to the 2006 rules, tested the options for 2011 on existing data to assess the information loss and risk, and are planning a new confidentiality system. In this presentation, we focus on the roles that we took as statistical methodologists in designing the proposed system. These included formulating new options for rules, and providing evidence from 2006 data for how the options would affect utility and safety. They included applying statistical thinking to clarify our basic principles and goals, to revise the logic behind the rules, and to create a plan that targeted each data product appropriately. Also in this presentation, we compare the 2011 confidentiality plans in United Kingdom, Australia and New Zealand, and consider the possibilities for the future of New Zealand census confidentiality.

Multistratum Fractional Factorial Designs

Ching-Shui Cheng
University of California, Berkeley

Multistratum experiments refer to those with multiple sources of errors. Multiple strata arise, e.g., when some treatment factors require larger experimental units than others since their levels are more difficult to change, or in experiments with multiple processing stages, the levels of the treatment factors are assigned at different stages. Nelder (1965a&b) developed a unified theory for the analysis of randomized experiments with what he called simple block structures. Speed and Bailey (1982) and Tjur (1984) further extended the theory to cover the more general orthogonal block structures. In this talk, I will revisit some recent works in the analysis and construction of multistratum fractional factorial designs, and show how they can be studied under the general framework provided by the theory of orthogonal block structures. An optimality criterion for selecting multistratum fractional factorial designs taking the stratum variances into account is proposed.

References:

Nelder, J.A. (1965a). "The analysis of randomized experiments with orthogonal block structure. I. block structure and the null analysis of variance". *Proceedings of the Royal Society A* **283**, 147–162.

Nelder, J.A. (1965b). "The analysis of randomized experiments with orthogonal block structure. II. Treatment structure and the general analysis of variance." *Proceedings of the Royal Society A* **283**, 163-178.

Speed, T.P. and Bailey, R.A. (1982). "On a class of association schemes derived from lattices of equivalence relations". *Algebraic Structures and Applications* 55–74, Marcel Dekker, New York.

Tjur, T. (1984). "Analysis of variance models in orthogonal designs". *International Statistical Review* **52**, 33–81.

Cook, Len

Various Stages in the Evolution of Official Statistics in Britain, their Influences on New Zealand, and Differences in the Development of Official Statistics Here

Len Cook

The influence of social reformers in the early 19th century played a critical part in establishing the place of statistics in British public life, which continued to be reflected in the way that government statistics developed, even to the present. The New Zealand experience had many parallels, but was founded on a very different place for government in social reform. Some contemporary NZ and UK issues are discussed in the context of this comparative analysis.

Costilla, Roy

Estimating Cancer Survival in Subpopulations with a Small Number of Cases: A Parametric Approach

Roy Costilla
Ministry of Health

Tony Blakely
University of Otago

Estimating survival for people diagnosed with cancer is an important and challenging task with potential policy and budget implications. A number of methods have been applied to estimate it in New Zealand: cancer specific survival, and relative survival amongst others. Nonetheless, they all share a caveat; they may produce highly imprecise and/or counterintuitive estimates in presence of small numbers of cases. This talk will propose the use of parametric models in these situations, namely Poisson regression of the excess mortality (i.e. the difference in observed and expected mortality, the latter being taken from population lifetables). Relative advantages of this modelling in this context will be discussed.

Davis, Walter

Fisher's Rank & Order Conditions and Instrumental Variables: Connections and Implications

Walter R. Davis
Statistics New Zealand

Fisher's rank and order conditions (1966) are necessary and sufficient conditions of identification for non-recursive systems of linear equations. Underrecognized is that Fisher's rules are an application of identification via instrumental variables (see also Lee 2008) and therefore can be viewed as necessary and sufficient tests for the existence of instrumental variables implied by the model. This approach can be combined with the more traditional instrumental variable method of identification to develop a general test of identification. This talk will demonstrate the equivalence of the approaches and extend Fisher and Lee and suggest ways Fisher's tests might be useful in the identification of structural equations with latent variables (under a broad set of conditions).

References:

Fisher, F. (1966). *The Identification Problem in Econometrics.* New York, McGraw-Hill.

Lee, M. (2008). "Method-of-moment view of linear simultaneous equation systems", *Statistica Neerlandica* **62(2)**, 230–238.

Ehlers, René

Triply Non-central Extended Bivariate Dirichlet Type I Distribution

René Ehlers
University of Pretoria, South Africa

Andriëtte Bekker
University of Pretoria, South Africa

The enriched triply non-central extended bivariate Dirichlet type I distribution is introduced. This distribution is constructed from independent chi-squared random variables by using the variables-in-common (or trivariate reduction) technique. The marginal density, product moment and the distribution of the product of the correlated components of this distribution are also derived. The effect of the additional parameters on the shape of the density functions and the correlation between the variables is shown. Special cases are highlighted to position this distribution in the bivariate Dirichlet distributions context.

Fernando, Sarojinie

Spatio-temporal Modelling of Relative Risk

Sarojinie Fernando — **Massey University, Palmerston North**

Martin L. Hazelton — **Massey University, Palmerston North**

Ganes Ganesalingam — **Massey University, Palmerston North**

The geographical relative risk function is a useful tool for investigating the spatial distribution of disease based on case and control data. We first discuss the usual way of estimating this function as the ratio of bivariate kernel density estimates constructed from the locations of cases and controls respectively. While this method is good at identifying disease hotspots when there are point sources of risk, it is less well suited to larger scale trends in risk such as might be expected line sources of risk. Motivated by this estimation, we propose a local linear model of the log-relative risk function. We show how this can be fitted using local likelihood methods, and we provide the asymptotic properties of the resulting estimator. We examine finite sample performance through numerical studies. Secondly, we extend relative risk function in space to spatio-temporal estimation through the use of suitable temporal kernel functions since time-scale is an important context when estimating disease risk. We use two numerical simulations to compare bandwidth selection methods using density ratio method and to compare density ratio over local linear method with respect to the best bandwidth selection method. I conclude with the outline of my further research.

Filus, Lidia

Weak Stochastic Dependence and Semi-pseudonormal Probability Distributions

Jerze K. Filus
Department of Mathematics and Computer Science, Oakton Community College

Lidia Z. Filus
Department of Mathematics, Northeastern Illinois University

In many reliability and biomedical applications of one- or multi-variate probability distributions, some random variables, say $X_1, \ldots, X_k$, that serve as explanatory to a variable of main interest such as the widely understood 'lifetime', say T, are the subject of an increasing attention of researchers in the area of stochastic modeling. We start with the actuary problem of modeling a (given his/her age) human (random) residual lifetime T that is assumed to depend on a random amount X of a stress, that the given (or a randomly chosen) human was subjected to. Such a stress may be, for example, a length X of a time the person smokes(ed) tobacco or admitted any other harmful substance, possibly more than only one. The main goal of our presentation is, roughly speaking, description of **stochastic dependence(s)** of general random variable, say T on an explanatory random variable X or on a set of such variables (in particular, stresses) $X_1, \ldots, X_k$. There are, at least two possibilities for these kinds of the dependence description. One, more frequently occurring in literature, is to define and then verify, a proper (algebraic) transformation: $T = f(X_1, \ldots, X_k)$ by means of a continuous function $f(\,)$ that belongs to a specified parametric family of functions. The shortcoming of such an approach may be the fact that, in many cases, a given realization $(x_1 \ldots, x_k)$ of the random vector $(X_1, \ldots, X_k)$ actually determines (with the probability one !) the corresponding unique realization t of the variable of interest T. This seldom is the case in most of the modeled realities. For example, the human lifetime has (besides being influenced by the stresses) 'its own' strength and time randomness, so the outer stresses only partially determine his/her residual lifetime. The mostly unrealistic algebraic dependence from explanatory quantities, we call 'strong dependence', in contrast to the essentially stochastic 'weak dependence'. The latter we define as an impact of the explanatory values $x_1, \ldots, x_k$ on the T variable's probability distribution, rather than directly on its particular realization t. That dependence definition, in turn, is obtained by means of specifying proper conditional distri-

butions of the main variable T, given any realizations $x_1, \ldots, x_k$ of the explanatory variables $X_1, \ldots, X_k$.

More specifically, we assume that the given set of the values $x_1, \ldots, x_k$ (stresses) impacts the (baseline) probability density $g(t;\theta)$ of T by changing its scalar or a vector parameter θ, in a unique way. Graphically, the difference between the strong and the weak transformation can shortly be express as $(X_1, \ldots, X_k) \to T$, for the algebraic transformation and as $(X_1 \ldots, X_k) \to g(t;\theta)$ for the weak. More precisely, the weak transformation can be expressed as a relation $(X_1, \ldots, X_k) \to \theta$, that is defined by a continuous function, say $\theta(x_1, \ldots, x_k)$. That function takes a unique value (of the parameter) for every elementary random event $(X_1, \ldots, X_k) = (x_1, \ldots, x_k)$. The above enables us to define the conditional density (or cdf) of T, given $(x_1, \ldots, x_k)$. That conditional pdf appears to be expressed in the form: $g(t|x_1, \ldots, x_k) = g(t;\theta(x_1, \ldots, x_k))$. The latter 'trick' exhibits an easy method for constructing the conditional distributions and that method is central to the emerging theory. Taking as the initial distribution $g(t;\theta)$, that may actually belong to any arbitrary class of pdfs (or cdfs) one can define a bunch of the models like the conditional distributions described above. Of our special attention is the class of semi-pseudonormal joint distributions, say $(T, X_1, \ldots, X_k)$ obtained as the arithmetic products of the normal, in t, conditionals $g(t|x_1, \ldots, x_k)$ and a joint distribution of the random vector $(X_1, \ldots, X_k)$. In applications, the marginals $X_1, \ldots, X_k$ may often be considered independent. The joint pdf of the $(T, X_1, \ldots, X_k)$ may, in a very special case, be the classical $(k+1)$ dimensional normal. Dropping the linearity of the conditional expectations $\theta(x_1, \ldots, x_k) = E[T|x_1, \ldots, x_k]$ in the normal models, one obtains some $(k+1)$-variate pseudonormal. What is the main novelty in this presentation is extending the class of the pseudonormal distributions by relaxing the assumption of the normality or pseudonormality of the explanatory random vectors $(X_1, \ldots, X_k)$, while preserving the normality of the conditional distribution $g(t|x_1, \ldots, x_k)$ in t. In such a way, one obtains a huge range of the semi-pseudonormal models that, hopefully, most of them can find the real life applications just as the univariate conditional normals.

An Analytical Expression for the Misclassification Error Rates Associated with the QDF in Discriminating Two Normal Populations

S. Ganesalingam
Massey University, Palmerston North

Siva Ganesh
Massey University, Palmerston north

A. Nanthakumar
Department of Mathematics, SUNY Oswego, USA

The estimation of the misclassification error rates is of vital importance in classification problems, as this is used as a basis to choose the best discriminant function; (ie) the one with a minimum misclassification error.

Consider the problem of statistical discrimination involving two multivariate normal populations with different covariance matrices. Traditionally a quadratic discriminant function (QDF) is used to separate two such populations. The error rates are estimated only via sample based misclassification.

In this paper, an approximate analytical expressions for the misclassification error rates associated with the QDF are derived for a very general p-dimensional normal population. An illustration is given for a 2 dimensional case.

Haslett, Steve

Data Cloning for Confidentiality and Data Encryption

S.J. Haslett
Massey University,
Palmerston North

K. Govindaraju
Massey University,
Palmerston North

It is possible to change the data in a linear model and still get the same parameter estimates. There are a range of techniques that can be used for such data cloning. This talk will give a brief summary, and then focus on one version of data cloning that seems to have wide application for confidentialising and encrypting data bases, even where the direct interest is not linear models.

References:

Govindaraju, K. and Haslett, S. (2008) "Illustration of regression toward the mean", *International Journal of Mathematical Education in Science and Technology* **39**, 544–550.

Haslett, S. and Govindaraju, K. (2009) "Illustration of regression toward the mean", *Australian and New Zealand Journal of Statistics* **51**, 499–504.

Haywood, John

Improved Multi-step Forecasting via a New Test of MLE Robustness

John Haywood

Victoria University of Wellington

Granville Tunnicliffe Wilson

Lancaster University

We propose a general test of whether a time series model, with parameters estimated by minimising the single-step forecast error sum of squares, is robust with respect to multi-step prediction, for some specified lead-time. The test may be applied to a, possibly seasonal, autoregressive integrated moving average (ARIMA) model using the parameters and residuals following maximum likelihood estimation. It is based on a score statistic, evaluated at these estimated parameters, that measures the sensitivity of the multi-step forecast error variance with respect to the parameters. We derive the large sample properties of the test and show by a simulation study that it has acceptable small sample size properties for higher lead times when applied to the integrated moving average or IMA(1,1) model that gives rise to the exponentially weighted moving average predictor. We investigate the power of the test when the IMA(1,1) model has been fitted to an ARMA(1,1) process. Further, we demonstrate the high power of the test when an autoregression is fitted to a process generated as the sum of a stochastic trend and cycle plus noise. We use frequency domain methods for the derivation and sampling properties of the test, and to give insight into its application. The test is illustrated on real time series, and an R function for its general application is available from http://msor.victoria.ac.nz/Main/JohnHaywood.

Hazelton, Martin

Shape Constrained Semiparametric Regression

Martin L. Hazelton
Massey University, Palmerston North

Berwin A. Turlach
University of Western Australia

Standard linear models can prove inadequate when the responses are correlated, or when there are nonlinear relationships between the response and one or more predictors. It has long been recognized that linear mixed models (LMMs) provide a means of tackling dependencies in the data, but it has only more recently been appreciated that the LMM framework is convenient for handling P-spline representations of nonlinear regression functions. In principle the P-splines have the capacity to capture very complex functional relationships between variables, but in some cases they may offer too much flexibility. For example, we may not know the precise form of the relationship between the response and a given predictor, but we may be certain that it is has some property like monotonicity or convexity. In this talk we describe a Bayesian MCMC approach to fitting semiparametric regression models incorporating such shape constraints. We pay particular attention to generation of candidate values for the coefficients of the P-spline, describing how to construct a truncated multivariate normal proposal distribution with (typically) high acceptance rate. We illustrate our methods on an example involving monotonic growth curves for trees under different experimental conditions.

Jones, Geoff

Inferring Infection History from Repeated Measures of Multiple Diagnostic Tests

Geoff Jones
Massey University, Palmerston North

Wes Johnson
University of California, Irvine, U.S.A.

Daan Vink
Massey University, Palmerston North

Nigel French
Massey University, Palmerston North

For many diseases the infection status of individuals cannot be observed directly, but can only be inferred from biomarkers that are subject to measurement error. Diagnosis based on observed symptoms can itself be regarded as an imperfect test of infection status. The temporal relationship between infection and disease may be complex, especially for recurrent diseases where individuals can experience multiple bouts of infection. Given repeated measures of a biomarker for infection and apparent disease status of a number of individuals at multiple time points, together with relevant covariates, we propose and estimate a model in which the unobserved infection status is a correlated latent process. This model can be used to investigate the temporal dynamics of infection, and to evaluate the usefulness of the biomarker for monitoring purposes. Our work is motivated and illustrated by a longitudinal study of Bovine Digital Dermatitis on commercial dairy farms in Cheshire, UK.

Population Monotony Coefficient

Farida Kachapova
Auckland University of Technology

Ilias Kachapov
University of Auckland

A measure λ^{**} briefly introduced by Reimann for positively quadrant dependent random variables is generalized to a measure ρm of monotone dependence of arbitrary random variables. The value of ρm is calculated for several bivariate distributions by applying Hoeffding lemma. Properties of ρm are described; in particular, random variables X and Y are increasingly (decreasingly) dependent if and only if $\rho m(X, Y) = 1(-1)$.

Kale, Hazel

Exploratory Data Analysis for Statistical Data Confidentiality

Hazel Kale
Statistics New Zealand, Wellington

Mike Camden
Statistics New Zealand, Wellington

When designing confidentiality processes, a statistical office aims for a proper balance between utility and safety for its data products. Measures for both of these properties exist, but they often result in one number for a table, dataset or variable. However an exploratory data analysis approach can reveal a wealth of useful information about the confidentiality processes. The effects of these processes on utility can be seen for each cell in a table and comparisons can be made between the noise added by the processes and the noise that is inherited from a sample survey. We attempt to show how the level of safety can vary across the respondents in a data collection. There is a surprisingly wide set of situations within confidentiality, where a data visualisation approach can reveal new and useful information. These visualisations can lead us to better confidentiality processes, and better guidance for researchers and other users about the validity of their results. We give examples from this set, and assess the strengths and weaknesses of the approach.

Khoo, Michael

Univariate Synthetic Control Charts for Variables Data: A Review

Michael B.C. Khoo
School of Mathematical Sciences, Universiti Sains Malaysia

Control charts are powerful tools used in the monitoring of the quality of a process. Univariate control charts can be grouped under two broad categories, based on the type of data, namely charts for variables data and charts for attribute data. A variables chart is used to monitor a process whose quality characteristic can be measured on a continuous scale while an attribute chart is employed for the monitoring of a quality characteristic which can only be classified as either conforming or nonconforming. A univariate synthetic chart consists of an integration of the basic univariate chart with the conforming run length (CRL) chart. This paper attempts to provide an overview of the main works on univariate synthetic charts for variables data, proposed in the last decade.

Lai, Chin-Diew

Distributions for Late Life Deceleration Phenomenon

Chin-Diew Lai
Massey University, Palmerston North

It has been observed that for several biological species, including humans, acceleration of mortality slows down after reaching a certain age. Several possible definitions for mortality deceleration are given and some ageing distributions are proposed and studied.

Lin, Gwo Dong

Maximum Correlation for Baker's Bivariate Distributions with Fixed Marginals

G. D. Lin
Institute of Statistical Science, Academia Sinica, Taiwan

J. S. Huang
Department of Mathematics and Statistics, University of Guelph, Canada

We investigate Baker's bivariate distributions with fixed marginals, which are based on order statistics, and find conditions under which the correlation converges to the maximum for Fréchet-Hoeffding upper bound as the sample size tends to infinity. The convergence rate of the correlation is also investigated for some specific cases.

Mao, Tian

Classification Trees for Poverty Mapping

Tian Mao
Massey University, Palmerston North

Geoff Jones
Massey University, Palmerston North

Siva Ganesh
Massey University, Palmerston North

Steve Haslett
Massey University, Palmerston North

Measuring differences in poverty levels within a country is important for aid allocation. Small area estimates of poverty incidence can be found by combining census and survey data. The usual method uses multiple regression, but an intuitive alternative is to build a classification tree for classifying households as poor or non-poor. This talk presents some preliminary results using this method, and compares them to the traditional regression method.

McGirr, Rebecca & Hawkes, Tim

A New Output Geography for Offence Statistics

Rebecca McGirr
Statistics New Zealand, Christchurch

Tim Hawkes
Statistics New Zealand, Wellington

Geographical variation in crime is often reported to lay audiences using police administrative boundaries — i.e. police districts and areas. These do not align well with other administrative boundaries such as territorial authorities and regional councils or cities. However crime data also exists at police station level, meaning there is potential for building new classifications that more closely line up with territorial authorities, regional councils and cities. We present the rules we developed to assign police stations to form three new crime reporting classifications. We also present our assessment of the quality of the new classifications. Finally we illustrate the new classifications using an example of violent crime reports.

MDS-optimal Supersaturated Designs

Arden Miller
University of Auckland

Boxin Tang
Simon Fraser University, Canada

A minimal dependent set (MDS) is a set of vectors that are linearly dependent but if any one of them is removed the resulting subset is independent. This talk will discuss the relationship between the minimal dependent sets of the column vectors of the design matrix for a 2-level supersaturated design and the resolvability of the design. It will introduce the concepts of MDS-resolution and MDS-aberration as criteria for comparing supersaturated designs. Results concerning supersaturated designs that have minimum MDS-aberration will be presented.

Nagatsuka, Hideki

Consistent Method of Estimation for Distributions with Unknown Origin

Hideki Nagatsuka
Tokyo Metropolitan University, Japan

N. Balakrishnan
McMaster University, Canada

Let a probability density function (pdf) be of the form

$$f(x;\gamma,\eta,\theta) = \frac{1}{\eta} f_0\left(\frac{x-\gamma}{\eta};\theta\right), \qquad \gamma < x, \tag{1}$$

where γ is a single unknown location parameter, η is a single unknown scale parameter and θ is a scalar or vector of parameters other than β and γ. Distributions of the type (1) include many well-known distributions used in many applications arising in life testing, reliability analysis, meteorology, hydrology and economics wherein it is reasonable to assume that there is a non-zero origin below which no event can occur, such as the Weibull distribution, gamma distribution, lognormal distribution and inverse Gaussian distribution.

The maximum likelihood (ML) method has the desirable properties of being consistent, asymptotically normal and asymptotically efficient under general conditions. However, it is well known that the classical regularity conditions for ML estimation of distributions of the type (1) are not satisfied since the support of the pdf depends on the unknown location parameter. The non-regular problem can emerge the ML estimation fails to produce solutions and the ML estimators have not the desirable asymptotic properties. For this reason, various alternative methods have been sought. However, in spite of many papers dealing with the parameter estimation for distributions with unknown origin, there does not appear to be any work wherein it is proved mathematically that, for the entire parameter space, estimates always exist uniquely and the estimators have consistency over the entire parameter space.

In this talk, we present our method of estimation for the parameters of distributions with unknown origin. The method is based on a data transformation. The transformation enables us to avoid the unbounded likelihood problem. In this method, under very mild assumptions, the estimates always exist uniquely over entire parameter space, and the estimators have consistency as well over the entire parameter space, which are proved mathematically. Through Monte

Carlo simulations, we further show that our method performs well compared to some known existing methods in terms of bias and root mean squared error for the specified distributions.

Namay, Rico

Population-preserving Propensity Score Stratification for Survey Non-response Modelling

Rico Namay
Inland Revenue Department,
Wellington

This presentation shows how propensity scores are used to model non-response in Inland Revenue's survey of SME tax compliance costs while preserving subgroup and overall population counts.

Traditional weighting class adjustment methods for survey non-response group 'similar' respondents and non-respondents in weighting classes. Usually these classes are formed by combinations of demographic variables. When numerous variables have to be considered to form the weighting classes, the number of cells can exponentially increase possibly producing many small or nil-sized weighting classes.

This limitation is addressed by propensity modelling stratification as one may consider as much relevant auxiliary information as necessary to create the non-response model. Because of the complex sampling scheme for the SME tax compliance cost survey however, propensity model strata will possibly overlap with several sampling design strata. This results in the non-preservation of the sampling design strata sizes.

The approach in this paper shows that via propensity modelling, the use of auxiliary information to form weighting classes need not be restricted to a few variables. At the same time, the presentation shows how the problem of subgroup and population counts preservation can be overcome.

Ng, Hon Keung Tony

Parametric Inference for System Lifetime Data with Signatures Available

H. K. T. Ng
Southern Methodist University, Dallas, Texas, USA

J. Navarro
University of Murcia, Spain

N. Balakrishnan
McMaster University, Canada

In this talk, the statistical inference of the lifetime distribution of component based on observing the system lifetimes with signature available is discussed. A general proportional hazard rate model for the lifetime of the components is considered, which includes some commonly used lifetime distributions. Different estimation methods for the proportional parameter are discussed. Monte Carlo simulation study is used to compare the performance of these estimation methods and recommendations are made based on these results.

Noble, Alasdair

Using Statistical Models to Combine Existing Data Sources to Produce Sounder, More Detailed, and Less Expensive Official Statistics

Alasdair Noble
Massey University, Palmerston North

Stephen Haslett
Massey University, Palmerston North

Geoff Jones
Massey University, Palmerston North

Dimitris Ballas
University of Sheffield, United Kingdom

The three techniques of small area estimation, spatial microsimulation and mass imputation are generally not seen as being similar. They are not all used within the same discipline — the technical literature on the first and last is mostly in Statistics, while the second has mostly been used by Human Geographers. There is however the similarity that, in certain forms, all aim to provide predictions, which can be amalgamated to form estimates for subgroups. In an effort to better understand the links and differences between the three techniques, an extensive simulation study was carried out. The results of this study will be reported and I will discuss the approaches to the simulations which helped gain insight into deeper similarities between what seemed, initially, very different methods.

This work was supported by a Statistics New Zealand Official Statistics Research Grant.

Olkin, Ingram

Life Distributions in Survival Analysis and Reliability: Structure of Semiparametric Families

Ingram Olkin
Professor Emeritus of Statistics and Education, Stanford University

Albert W. Marshall
Professor Emeritus of Statistics, University of British Colombia

Semiparametric families are families that have both a real parameter and a parameter that is itself a distribution. A number of semiparametric parametric families suitable for lifetime data in survival or reliability are introduced: scale, power, frailty (proportional hazards), age, moment, and others. Interesting results on stochastic orderings are obtained for these families. The coincidence of two families provides a characterization of the underlying distribution. Some of the characterization results provide a rationale for the use of certain families. In this talk we provide an overview of these semiparametric families, and present several characterizations.

Professor Olkin's attendance is generously supported by our principal sponsor:

Statistics New Zealand
TATAURANGA AOTEAROA

Ong, Hong Choon

Modelling the Aids Epidemic in Penang, Malaysia

Hong Choon Ong
School of Mathematical Sciences, Universiti Sains Malaysia

Lay Fong Sin
School of Mathematical Sciences, Universiti Sains Malaysia

Li Ling Tan
School of Mathematical Sciences, Universiti Sains Malaysia

This study briefly look at some of the statistical methods that have been developed to model the HIV/AIDS epidemic and also use the back calculation method to estimate the HIV infection rate in Penang. The back calculation program has been chosen to model the underlying HIV/AIDS epidemic in Penang, Malaysia because it makes use of the AIDS incidence data which is more reflective of the epidemic as compared to the number of HIV infected recorded which is known only if tests are conducted. The AIDS incidence data collected however have some limitations and uncertainties due to censoring of the data, reporting delay, under reporting and the changes in the AIDS surveillance definition. The back calculation method is used to reconstruct and estimate the number of people who have been infected previously in Penang using the AIDS incidence data obtained and an estimate of the incubation period distribution. The simulation shows the presence of under and delay reporting in the AIDS incidence data obtained especially in the early stages.

Ong, Seng Huat

Properties and Application of the Inverse Trinomial Distribution

Seng Huat Ong
University of Malaya, Kuala Lumpur, Malaysia

Kian Wah Liew
Universiti Tunku Abdul Rahman, Kuala Lumpur, Malaysia

This paper considers the inverse trinomial distribution derived by Shimizu and Yanagimoto (1991) as a one-dimensional random walk distribution. Various formulations for the inverse trinomial distribution have been derived. In particular the generalized and mixed Poisson formulations are considered. For the inverse trinomial as a mixed Poisson distribution, the mixing distribution has been obtained. Probabilistic properties like infinite divisibility and discrete self-decomposability are also examined. Examples of application to empirical modeling are also given.

Reference:

Shimizu, K. and Yanagimoto, T. (1991). "The inverse trinomial distribution." *Jap. J. Appl. Stat* **20(2)**, 89–96. (in Japanese)

Richens, Andrew

Using Score Functions to Identify Key Firms in Statistics New Zealand Surveys

Andrew Richens
Statistics New Zealand

One of the biggest costs for Statistics New Zealand surveys is following up non respondents. To help minimise these costs, each survey has a list of 'key firms' that help prioritise follow up. Until now, each survey has developed its own list of influential unit. This talk describes a standard approach to identifying key firms based on score functions. The objective is to find a standard method to maximise survey coverage while minimising collection costs.

Rodado, Armando

Selecting Central American Volcanoes for an Empirical Bayes Analysis

Armando Rodado
Massey University, Palmerston North

Mark Bebbington
Massey University, Palmerston North

Alasdair Noble
Massey University, Palmerston North

Shane Cronin
Massey University, Palmerston North

Gill Jolly
GNS, Auckland

An empirical Bayes analysis (EBA) of volcanic eruptions has been used for inference of the eruption rate of a volcano under a Poisson process before by Solow. To implement this methodology however several decision need to be taken. In particular, selecting meaningful volcanoes is an important step that need to be explicitly considered. Our approach to deal with the analytical and computational challenges that are in the core of this methodology are studied and exemplified in a Central American volcano context.

Using Finite Mixtures to Compute Robustified Statistics for Regression Parameters

Maheswaran Rohan
University of Waikato & Department of Conservation, Hamilton

When data is contaminated, robust methods are commonly used to compute robust statistics for the model parameters. In contrast with the classical statistical modelling, the methodology of robust statistics is often ad hoc and lacks a unified approach. Hence we are developing a unified method for making likelihood based statistical modelling more robust by employing finite mixture models and the EM algorithm. In this talk, I will explain a general approach for estimating the linear regression parameters and computing influence functions of the estimates. I will use the well known Belgium telephone data in an example, showing how to compute statistics for the parameters based on our method and influence functions of these estimates. Finally, I will compare the results obtained by our method with M-estimators such as Huber estimator.

Sampson, Allan

Multivariate Modelling Issues for Multiple Outcomes in Post-mortem Tissue Studies

Allan R. Sampson
University of Pittsburgh

Qiang Wu
East Carolina University

Josephine Asafu-Adjei
University of Pittsburgh

Post-mortem brain tissue studies are employed for a number of psychiatric diseases to identify cellular biomarkers which distinguish subjects with the disease from normal controls. We have extensively collaborated with neuroscientists and psychiatrists who have conducted numerous such studies concerning schizophrenia. These studies typically have used a matched subject-control design for sampling and tissue processing. In this talk we focus on several projects with differing purposes which integrate results concerning schizophrenia across multiple studies. There are specifics of the available data that require new approaches for these projects in terms of structured multivariate models, mixture modeling, missing data, and discriminant techniques. The goal of our presentation is to provide an overview of these issues and outline some of our methodology.

Scott, Alastair

Pseudo Likelihood-Ratio Tests for Survey Data

Alastair Scott
Department of Statistics, University of Auckland

Thomas Lumley
Department of Biostatistics, University of Washington

Traditional survey methods are largely concerned with estimating simple descriptive quantities such as population means and proportions. Increasingly, however, data from complex surveys are being used to build the same sort of explanatory and predictive models used in the rest of statistics. Unfortunately complexities such as the use of variable selection probabilities and correlations induced by the hierarchical structure of multi-stage sampling mean that the assumptions underlying standard statistical methods for model-building are not even approximately valid for survey data. The problem of parameter estimation has been largely solved through the use of weighted estimating equations, and software for many standard statistical procedures is now available in the major statistical packages. Then, given an estimate of a parameter vector along with an estimate of its covariance matrix, we can use the Wald statistic to test hypotheses and to construct appropriate confidence regions. This has the usual problems associated with the Wald test. For example, the statistic is not invariant under nonlinear transformations of the parameter, the tests often have poor small sample behavior, and the confidence regions often contain invalid values of the parameter. There is an additional problem with survey data: the degrees of freedom for the estimated covariance matrix are often very small, depending on the number of primary sampling units rather than the number of observations, and the inverse of the estimated covariance matrix can be very unstable. Ideally we would prefer to use a likelihood ratio test which is invariant and usually has better small sample properties. Unfortunately there is no likelihood function for survey data. However, it is possible to construct a pseudo likelihood-ratio statistic which has many of the same properties. We develop these properties and show that the asymptotic null distribution is a linear combination of chi-squared random variables. The coefficients are eigenvalues of a matrix product that does not involve the inverse of the estimated covariance matrix. We compare the performance of the test with the Wald test using data from some well-known health surveys.

Tang, Boxin

Robust Designs Through Partially Clear Two-Factor Interactions

Ryan Lekivetz
Simon Fraser University

Boxin Tang
Simon Fraser University

Orthogonal arrays with clear two-factor interactions are very attractive for industrial experiments as they provide a class of designs that are robust to nonnegligible two-factor interactions. If prior knowledge suggests that some two-factor interactions are negligible before running the experiment, then designs can be found that allow additional factors to be studied while remaining robust to the nonnegligible two-factor interactions. This is done through *partially clear* two-factor interactions. We study the existence and construction of such robust designs in this paper. Several construction results are presented and examples are provided for illustration. We also prove a result that establishes an upper bound on the maximum number of clear two-factor interactions in an orthogonal array.

Tularam, Anand & Roca, Eduardo

An Investigation of the Relationship Between Socially Responsible Investment Markets Based on the Dynamic Conditional Correlation Methodology

Gurudeo Anand Tularam
Griffith University, Australia

Eduardo Roca
Griffith University, Australia

Victor S.H. Wong
Griffith University, Australia

This paper investigates the relationship of the Australian Socially Responsible Investment (SRI) market with other SRI markets worldwide during the period 1994 to 2009 based on the dynamic conditional correlation multivariate GARCH model (DCC-MVGarch, Engle, 2002). In the DCC method, the multivariate conditional variance estimation is simplified by estimating univariate GARCH models for each asset. Using the transformed residuals resulting from the first stage, we can estimate a conditional correlation estimator. The standard errors for the first stage parameters remain while the standard errors for the correlation parameters are modified. The study examines the relationship of the Australian SRI market with fourteen other markets - Canada, Denmark, France, Germany, Hong Kong, Ireland, Japan, Netherlands, Norway, South Africa, Sweden, Switzerland, United Kingdom and the United States. Over the last ten years, there has been a phenomenal growth in the amount of funds placed in SRI globally, now estimated to be US$6.5 trillion, with around US$55 billion in the Australian market. Knowledge regarding the correlation of the Australian SRI market with other SRI markets overseas is highly important for Australian SRI investors in their quest for international portfolio diversification. Portfolio diversification theory posits that the lower (higher) the correlation between markets, the higher (lower) the gains to be made. Our results reveal that as expected, the Australian market experienced a surge in correlation with all other markets during the global financial crisis. During the period of study, the correlation of Australia with Canada, Denmark, Norway, and the United Kingdom increased over time while its correlation with other countries remained stationary. This implies that the Australian SRI market is becoming more integrated with those of Canada, Denmark, Norway and the United Kingdom and therefore these overseas markets provide less portfolio diversification benefits to Australian SRI investors while the other markets will do.

Turner, Rolf

Renyi's Theorem and Poisson Processes for the Uninitiated

T. Rolf Turner
Starpath Project,
University of Auckland

In introductory statistics courses which have a bit of a "theoretical" component, it is conventional to introduce constant intensity Poisson processes by saying that they satisfy a number of simple conditions (including, e.g., $\Pr(N(0,h)=1) = \lambda h + o(h)$) and then deriving the formula for the Poisson probability function (e.g. via a system of differential equations). The fact that the Poisson formula comes out of these vague "$o(h)$" conditions has a magical feel, and is very satisfying. Nevertheless it is well known that the "$o(h)$" conditions are "overkill" and it has always seemed to me that one should be able to do better. E.g. if one assumes that the process is simple, has constant intensity and satisfies the "independence condition" then one should be able to obtain the Poisson probability formula (even more magically). If the condition of stationarity is added then one can indeed derive the Poisson formula — but is the stationarity assumption necessary? Renyi's Theorem (as discussed in Kingman's book on Poisson processes) shows that it is *not* necessary. However I still cannot get at the derivation of the Poisson formula, sans stationarity assumption, except by going via the rather convoluted and rather deep route provided by Renyi's Theorem. Any hints would be gratefully received.

van Koten, Chikako

An Analysis of Power Approach to Time Series with a Known Periodic Input

Chikako van Koten
AgResearch, Lincoln

In some designed experiments we obtain a set of time series of measurements as output of a known periodic input, and our main interest is to compare the output series from different groups to examine if the group means are different. I came across such an experiment in which fixed mechanical disturbance applied to independent samples of two different floorings, resulted in output time series of dust particle concentration levels. I analysed data from this experiment using a frequency domain method called Analysis of Power (ANOPOW).

One way to model the relation between the input and output series in this type of experiment is to use a regression form:

$$y_t = \sum_{r=-\infty}^{\infty} z_{t-r}\beta_r + v_t$$

where y_t and z_t are the output and input series and v_t is the error series. The error series is assumed to be a zero-mean, stationary, normal process. In spectral analysis, we can approximate this form as a frequency domain regression model using the discrete Fourier transform of the series:

$$Y(\omega_k) = Z(\omega_k)B(\omega_k) + V(\omega_k)$$

ANOPOW is then performed to test the null hypothesis that the regression coefficient $B(\omega_k) = 0$, by taking the ratio of regression and error power spectral components, in a way analogous to ANOVA.

If our interest is a simple equality of group means test, we can use ANOPOW to compare the full model with different group effects to the reduced model under the null assumption that group means are equal. ANOPOW is simple but very effective for testing equality of group means in this type of experiment. In this presentation, I intend to introduce ANOPOW and show how the test is done using the experiment I mentioned above as an example.

Walker, Lyndon

Analysing Ethnic Partnership Matching Using a Grid-Based Evolutionary Algorithm

Lyndon Walker
Unitec & University of Auckland, Auckland

The Modelling Social Change (MoSC) project is a Marsden funded study that investigated changes in the social structure of New Zealand by examining patterns of inter-ethnic partnering in married and de-facto relationships. One component of this study was a social simulation model of partnership formation that was populated with unit-level data from the New Zealand Census. The simulation was written in Java and run on the Auckland cluster of the BeSTGRID computer network (www.bestgrid.org). The processing power of the cluster allowed the simulation to be run at a city level, with unit-level data that provided demographic information for all of the single eighteen to thirty year olds listed in the census in the Auckland, Wellington and Canterbury regions. This presentation will explain how social simulation was used to supplement the traditional statistical analysis in examining partnership patterns. It will then discuss how the BeSTGRID cluster was used for the parallel processing of the simulation model, in order to use an evolutionary optimisation algorithm to search for optimal combinations of the partnering parameters.

Wang, Ting

Markov-modulated Hawkes Process with Stepwise Decay

Ting Wang
Massey University, Palmerston North

Mark Bebbington
Massey University, Palmerston North

David Harte
Statistics Research Associates Ltd, Wellington

We proposed a new model — the Markov-modulated Hawkes process with stepwise decay (MMHPSD) — to investigate the variation in seismicity rate during a series of earthquake sequences including multiple main shocks. The MMHPSD is a self-exciting process which switches among different states, in each of which the process has distinguishable background seismicity and decay rates. Parameter estimation is developed via the Expectation Maximization algorithm. The model is applied to data from the Landers-Hector Mine earthquake sequence, demonstrating that it is useful for modelling changes in the temporal patterns of seismicity. The states in the model can capture the behaviour of main shocks, large aftershocks, secondary aftershocks and a period of quiescence with different background rates and decay rates. The state transitions can indicate if there is any seismicity shadow or relative quiescence.

Wang, Yuancheng

Assessing the Performance of Matrix Representation with Parsimony

Yuancheng Wang
University of Canterbury, Christchurch

James H. Degnen
University of Canterbury, Christchurch

Phylogenetics is the research of ancestor-descendant relationships among different groups of organisms, for example, species or populations of interest. The datasets involved are usually sequence alignments of various subsets of taxa for various genes.

A major task of phylogenetics is often to combine estimated gene trees from many loci sampled from the genes into an overall estimate species tree. Eventually, one hopefully can construct the tree of life that depicts the ancestor-descendant relationships for the whole known species around the world. If there is missing data or incomplete sampling in the datasets, then supertree methods can be used to assemble gene trees with different subsets of taxa into an estimated overall species tree.

In this study, we assume that the gene tree conflict is solely due to incomplete lineage sorting under the multispecies coalescent model. On top of that, we examine the performance of the most commonly used supertree method, Matrix Representation with Parsimony (MRP) to explore its statistical properties in this setting. In particular, we show that MRP is not statistically consistent. That is, an estimated species tree other than the true species tree can be more likely to be returned by MRP as the number of gene trees increases. For some situations, using longer branch lengths, randomly deleting taxa or even introducing mutation can improve the performance of MRP so that the matching species tree is recovered more often. In conclusion, MRP is a supertree method that is able to handle large amounts of conflict in the input gene trees. However, MRP is not statistically consistent, when using gene trees arise from the multispecies coalescent model to estimate species trees.

Westbrooke, Ian

R with Menus — R Commander Software for Teaching Statistics to Non-statisticians

Ian Westbrooke
Department of Conservation, Christchurch

Maheswaran Rohan
Department of Conservation, Hamilton

R Commander is a package for R that allows users to create and run R code using menus that are similar to many point and click statistical packages. This provides a useful tool for non-statisticians in various fields to carry out statistical analyses. At the Department of Conservation, we have only two statisticians for a staff of 1800 plus, including hundreds of graduates in science, technical and field roles. Many require significant statistical skills ranging from design to applying and interpreting statistical models. An effective way of improving their statistical skills is through training. Last year we modified our course teaching linear models and glms, changing from R scripts to R Commander. We will look at the advantages of R Commander and demonstrate some of the material we teach, including importing data, creating graphs and carrying out statistical modelling. We will then briefly describe the results of a recent survey of DOC staff who attended last year's courses along with the information we have about the use of R Commander at local universities. We have found that R Commander works well for introducing our group of non-statisticians to using R.

Willink, Robin (1 of 2)

Some Statistical Advances Originating in Measurement Science

Robin Willink
Industrial Research Ltd,
Lower Hutt

Probabilistic and statistical analysis is important in the field of metrology, i.e. measurement science. I will draw on a decade of experience of working and publishing alongside scientists from the Measurement Standards Laboratory of New Zealand to describe several statistical advances that have been made in this field but which possess potential for wider application. These advances include (i) a Welch-type approximation for the 'effective number of degrees of freedom' when estimating a linear combination of the mean vectors of multivariate normal distributions, (ii) a moment-based method of approximating a function of several random variables by a variable from the Pearson system and explicit approximations to useful tail quantiles of this Pearson variable, and (iii) a means of approximating a random variable by a different variable with a simple quantile function (inverse distribution function) for easy sampling in a Monte Carlo study. These developments have been made against the background of international confusion surrounding various concepts of 'measurement uncertainty', where the idea of describing a fixed quantity by a probability distribution is often wrongly deemed to be compatible with orthodox statistical analysis. This issue will be briefly discussed.

Willink, Robin (2 of 2)

A Confidence Interval for the Bounded Normal Mean from a Small Sample: an Adaptation of the Feldman-Cousins Interval

Robin Willink
Industrial Research Ltd,
Lower Hutt

The usual confidence-interval procedure for estimating the mean of a normal distribution can generate empty intervals when the mean is known to be bounded, as in a practical measurement problem, for example. An exact interval that is never empty has been given by Feldman and Cousins for the case where the variance is known (Physical Review D, volume 57, number 7, 3873-3889, 1998). We describe this interval and present a similar interval for the situation where the variance is unknown and the sample is small. The interval ultimately developed is designed for practical use, so it is constructed to have a simple explicit form.

Normalization of Ratio Data

Graham Wood
Macquarie University, Sydney

Jamie Sherman
University of California, San Francisco

Mark Malloy
Australian Proteome Analysis Facility, Sydney

Joseph Descallar
Macquarie University, Sydney

Braddon Lance
Macquarie University, Sydney

Pauliina Uitto
Macquarie University, Sydney

In searching for proteins that might be associated, for example, with ovarian cancer, it is common to consider the ratio of the amount of a particular protein in a given quantity of the cancerous cell over the amount in the same quantity of a healthy cell. Ensuring that the quantities are equal in the experimental process is difficult, so it is generally necessary to include so-called "housekeeping" proteins in the study, assumed to display only minimal changes in abundance between samples under comparison. These are used to normalize the ratio data. A standard way of normalizing is to divide by the geometric mean of the housekeeping ratios, equivalent to centring the log transformed data. In this talk, normalization will be set in a broad context, showing that the standard practice just described has desirable properties, but that improvements are sometimes possible.

Yee, Thomas

Parameter Estimation in Many Statistical Distributions

Thomas W. Yee
University of Auckland

The classes of vector generalized linear and additive models (VGLMs and VGAMs; Yee and Wild (1996), Yee and Hastie (2003)) enables maximum likelihood estimation of many models and distributions including scores of standard and nonstandard univariate and continuous distributions, categorical data analysis, survival analysis, time series and nonlinear least-squares models. Usually Fisher scoring is used for these. Many common statistical distributions have been implemented in the author's VGAM R package, such as the negative binomial, zero-inflated and zero-altered Poisson and negative binomial, the zeta and Zipf distributions, etc. This talk will give an overview of the VGLM/VGAM framework, and the performance and usage of the software for estimating the parameters of many statistical distributions.

References:

Yee, T.W. and Hastie, T.J. (2003). "Reduced-rank Vector Generalized Linear Models." *Statistical Modelling* **3**, 15–41.

Yee, T.W. and Wild, C.J. (1996). "Vector Generalized Additive Models." *Journal of the Royal Statistical Society. Series B. Methodological* **58**, 481–493.

Zheng, Guan Yu (Fish)

Empirical Study of Extreme Values on Seasonal Adjustments in Time Series Analysis

Guan Yu Zheng
Statistics New Zealand

Statistics New Zealand is interested in detection and treatment of extreme values in its time series as these may lead to bias and unreliability of some of its time series outputs. Time series outputs in Statistics New Zealand are include descriptive series for short term (e.g. seasonally adjusted) and long term movements (e.g. trend). An extreme value is a one-off event which leads to an atypically large value for a particular time period. It is well-known that extreme values can affect the fitted time series models, particularly trend estimates. Therefore some modification of the recorded value should be applied to reduce the impacts of extremes. Statistics New Zealand has access to all the unit values that contribute to any time series value, and thus has the facility to prior adjust by removing individual extreme values rather than modifying the value. Empirical results are presented to illustrate various approaches to identifying and managing extreme unit values. Possible methods for incorporating the economic information into the trend estimate are also discussed.

Zitikis, Ričardas

Weighted Distributions, Insurance Premiums, and Extreme Events

Ričardas Zitikis
University of Western Ontario,
London, Canada

Dedicated to Chin-Diew Lai on the occasion of the publication of our 20th joint paper

The conditional tail expectation (CTE) has become a standard risk measure in Insurance. The risk measure is closely related to the mean residual life (MRL) function, which naturally leads to weighted distributions and, in turn, to general weighted insurance premiums and capital allocations, recently introduced by E. Furman (Toronto, Canada) and the present speaker. After a little excursion through other related problems of Economics, Finance, and Mathematics, we shall come back to the "simple" CTE risk measure and discuss some of the statistical challenges in the area. In particular, we shall note the crucial role of the extreme value theory and with it associated problems, which, when constructing an empirical CTE estimator, have recently been tackled by A. Necir and A. Rassoul (Biskra, Algeria) together with the present speaker.

Biography of Professor Chin-Diew Lai

Professor Chin-Diew Lai graduated with a Masters degree from Auckland, and a PhD from Victoria University of Wellington, the latter under the supervision of David Vere-Jones. In his career at Massey since 1979, he has published over 100 peer-reviewed journal articles, proceedings papers and book chapters. He has also co-authored 4 books, with such luminaries as M. Xie and N. Balakrishnan, which have become standard reference works on bivariate distributions and on ageing in reliability.

Professor Lai is a recognized authority on mathematical reliability theory, particularly in its extension to the more 'gritty' (or less idealized) real-world form where component failures are no longer independent. As such he has been invited to speak at numerous conferences, from Australia to Spain, and the USA, by way of Russia! His reputation is higher off-shore than in his own country, something that this conference seeks to overturn.

Programme

Tuesday 29 June

Time	Stream 1	Stream 2	Stream 3
9:00		**Welcome**	
9:30	Balakrishnan, N. (Bala): 13 "Some Cure Rate Models and Associated Inference and Application to Cutaneous Melanoma Data"		
10:30		**Morning Tea**	
11:00	Miller, Arden: 38 "MDS-optimal Supersaturated Designs"		Lin, Gwo Dong: 36 "Maximum Correlation for Baker's Bivariate Distributions with Fixed Marginals"
11:30		Bilton, Penelope: 17 "A Statistical Model to Characterize the Naphthenic Acid Component of Petroleum"	Wood, Graham: 61 "Normalization of Ratio Data"
12:00		**Lunch**	
1:00	absent due to illness	Jones, Geoff: 32 "Inferring Infection History from Repeated Measures of Multiple Diagnostic Tests"	Ong, Seng Huat: 46 "Properties and Application of the Inverse Trinomial Distribution"
1:30	van Koten, Chikako: 54 "An Analysis of Power Approach to Time Series with a Known Periodic Input"	Betz-Stablein, Brigid: 16 "Disease Mapping Techniques Applied to Glaucoma Visual Field Datasets"	Ehlers, René: 24 "Triply Non-central Extended Bivariate Dirichlet Type I Distribution"

2:00	Zheng, Guan Yu (Fish): 63 "Empirical Study of Extreme Values on Seasonal Adjustments in Time Series Analysis"	Wang, Yuancheng: 57 "Assessing the Performance of Matrix Representation with Parsimony"	Alzaid, Abdulhamid: 9 "Binomial Difference Distribution"
2:30	free	Costilla, Roy: 22 "Estimating Cancer Survival in Subpopulations with a Small Number of Cases: A Parametric Approach"	Ardalan, Arash: 11 "A Generalized Normal-Laplace Distribution: Properties, Estimation and Applications"
3:00	**Afternoon Tea**		
3:30	Willink, Robin (1 of 2): 59 "Some Statistical Advances Originating in Measurement Science"	Young Statisticians' Session	Ng, Hon Keung Tony: 42 "Parametric Inference for System Lifetime Data with Signatures Available"
4:00	Turner, Rolf: 53 "Renyi's Theorem and Poisson Processes for the Uninitiated"	continued	
4:30	Ali, Abdul: 7 "Ko te Anga te Hoto — Structure is the Link"	continued	

Wednesday 30 June

Time	Stream 1	Stream 2	Stream 3
9:00	**Welcome**		
9:10	Olkin, Ingram: 44 "Life Distributions in Survival Analysis and Reliability: Structure of Semiparametric Families"		
10:10	**Morning Tea**		
10:30	Kale, Hazel: 34 "Exploratory Data Analysis for Statistical Data Confidentiality"	Ganesalingam, Ganes: 28 "An Analytical Expression for the Misclassification Error Rates Associated with the QDF in Discriminating Two Normal Populations"	Wang, Ting: 56 "Markov-modulated Hawkes Process with Stepwise Decay"
11:00	Chen, Chen: 20 "Confidentiality for the 2011 Census: Statistical Thinking Applied"	Walker, Lyndon: 55 "Analysing Ethnic Partnership Matching Using a Grid-Based Evolutionary Algorithm"	Filus, Lidia: 26 "Weak Stochastic Dependence and Semi-pseudonormal Probability Distributions"
11:30	Haslett, Steve: 29 "Data Cloning for Confidentiality and Data Encryption"	Anwar, Nafees: 10 "Measurement and Visualization of Data Complexity for Classification Problem"	Kachapova, Farida: 33 "Population Monotony Coefficient"
12:00	**Lunch**		

1:00	Willink, Robin (2 of 2): 60 "A Confidence Interval for the Bounded Normal Mean from a Small Sample: an Adaptation of the Feldman-Cousins Interval"	Aubry, Jean-Marie: 12 "Large Deviations for Quasi-arithmetically Self-normalized Random Variables"	Zitikis, Ričardas: 64 "Weighted Distributions, Insurance Premiums, and Extreme Events"
1:30	McGirr, Rebecca & Hawkes, Tim: 37 "A New Output Geography for Offence Statistics"	Davis, Walter: 23 "Fisher's Rank & Order Conditions and Instrumental Variables: Connections and Implications"	Yee, Thomas: 62 "Parameter Estimation in Many Statistical Distributions"
2:00	Namay, Rico: 41 "Population-preserving Propensity Score Stratification for Survey Non-response Modelling"	Tularam, Anand & Roca, Eduardo: 52 "An Investigation of the Relationship Between Socially Responsible Investment Markets Based on the Dynamic Conditional Correlation Methodology"	Bebbington, Mark: 15 "Analyzing Treatment Effects on Distributions with Complex Structure"
2:30	Scott, Alastair: 50 "Pseudo Likelihood-Ratio Tests for Survey Data"	Rohan, Maheswaran: 48 "Using Finite Mixtures to Compute Robustified Statistics for Regression Parameters"	Nagatsuka, Hideki: 39 "Consistent Method of Estimation for Distributions with Unknown Origin"
3:00	**Afternoon Tea**		

3:30	Haywood, John: 30 "Improved Multi-step Forecasting via a New Test of MLE Robustness"	Chee, Chew-Seng: 19 "Mixture-based Nonparametric Density Estimation: Maximum Likelihood vs. Least Squares"	Tang, Boxin: 51 "Robust Designs Through Partially Clear Two-Factor Interactions"
4:00		Fernando, Sarojinie: 25 "Spatio-temporal Modelling of Relative Risk"	Khoo, Michael: 35 "Univariate Synthetic Control Charts for Variables Data: A Review"
4:30	Rodado, Armando: 47 "Selecting Central American Volcanoes for an Empirical Bayes Analysis"	Hazelton, Martin: 31 "Shape Constrained Semiparametric Regression"	Chan, Ping Shing Ben: 18 "Optimal Allocation for Multi-level Stress Testing with Extreme-value Regression"
5:00		NZSA AGM	

Thursday 1 July

Time	Stream 1	Stream 2	Stream 3
9:00	**Welcome**		
9:05	Cheng, Ching-Shui: 21 "Multistratum Fractional Factorial Designs"		
10:05	**Morning Tea**		
10:30	Sampson, Allan: 49 "Multivariate Modelling Issues for Multiple Outcomes in Post-mortem Tissue Studies"	Cook, Len: 22 "Various Stages in the Evolution of Official Statistics in Britain, their Influences on New Zealand, and Differences in the Development of Official Statistics Here"	Richens, Andrew: 46 "Using Score Functions to Identify Key Firms in Statistics New Zealand Surveys"
11:00	Ong, Hong Choon: 45 "Modelling the Aids Epidemic in Penang, Malaysia"	Noble, Alasdair: 43 "Using Statistical Models to Combine Existing Data Sources to Produce Sounder, More Detailed, and Less Expensive Official Statistics"	Mao, Tian: 36 "Classification Trees for Poverty Mapping"
11:30	Lai, Chin-Diew: 35 "Distributions for Late Life Deceleration Phenomenon"		
12:00	Wrap-up		
12:15	**Lunch**		
1:15		Westbrooke, Ian: 58 "R with Menus — R Commander Software for Teaching Statistics to Non-statisticians"	
1:45		TBA	
3:15	**Afternoon Tea**		