New Zealand Statistical Association 63rd Annual Conference

29-30 November 2012
University of Otago, Dunedin

Wednesday 28 November	
18:00 - 19:00	Welcome Reception (drink and canapés) and Poster Session (Castle Lecture Theatre Foyer)

Note: (S) denotes a student talk

Thursday 29 November			
8:45	Welcome and 'housekeeping'		
9:00	Alastair Scott (Castle 2) Chair: David Fletcher *Fitting regression models with response-dependent samples*		
	Modelling 1 (Castle 2) Chair: Russell Millar	Software (Burns 1) Chair: David Fletcher	Sampling (Burns 2) Chair: Austina Clark
10:00	Khair Jones (S) *Shape constrained penalised spline regression for generalised linear models*	Andrew Cathie *Some Stories from the Field*	Kylie Maxwell *Identifying respondent groups in a longitudinal business survey sample: an application of dual latent trajectory modelling*
10:20	Daniel Turek (S) *A new approach to model-averaged confidence intervals*	John Koolaard *Package Predictmean: Further inference from linear models*	Haizhen Wu *Design of variables acceptance sampling plans for non-normally distributed quality characteristics*
10:40	Break		
	Bayes (Castle 2) Chair: David Fletcher	Environment (Burns 1) Chair: Ian Westbrooke	Social (Burns 2) Chair: Len Cook
11:00	Russell Millar *A modified DIC for comparison of Bayesian state-space models*	Maryann Pirie *Do kauri trees experience old age?*	Penelope Bilton (S) *Decision tree models for poverty mapping*
11:20	Beatrix Jones *A model-list approach to using decomposable models for Gaussian (inverse-)covariance selection*	Darcy Webber (S) *Broad scale management in spatially heterogeneous fisheries - does it matter?*	Janine Wright *Modelling Maori Language*
11:40	Brigid Betz-Stablein (S) *Predicting glaucoma progression: A novel spatial analysis*	Shirley Pledger *Capture-recapture models for estimating breeding return times and abundance.*	Barry McDonald *Taking the numbers on faith?*
12:00	Lyndal Henden (S) *Estimation of Population Size for Small Communities in Indonesia*	Graham McBride *A sensible weight-of-evidence from three simultaneous hypothesis tests?*	Elena Moltchanova *LEGO figurines as a tool of assessing emotions*
12:20	Lunch		
13:20	Roger Payne (Castle 2) Chair: Neil Cox *Hierarchical generalized models - theory and practice*		
	GLMs (Castle 2) Chair: Neil Cox	Testing 1 (Burns 1) Chair: Ken Dodds	Mixtures (Burns 2) Chair: Vanessa Cave
14:20	Murray Jorgensen *The sex life of untagged wetas*	David Scott *The tail probabilities of the generalized inverse Gaussian distribution*	Daniel Fernandez (S) *Likelihood-based finite mixture models for ordinal data*
14:40	David Fletcher *A new method for estimating overdispersion in generalized linear models*	Lingyun (Larry) Zhang *The exact likelihood ratio test for equality of two normal populations*	Zoe van Havre (S) *Investigating the number of components in overfitted Gaussian mixture models.*

Thursday 29 November			
15:00	Break		
	Geosciences (Castle 2) Chair: Martin Hazelton	Official Statistics 1 (Burns 1) Chair: Mike Camden	Probability (Burns 2) Chair: Mark Holmes
15:20	Julie Bessac (S) *Gaussian state-space model for wind fields in the North-East Atlantic*	Frances Krsinich *Decomposing property price indexes into land and structure components*	Charles Newman *Coarsening Models*
15:40	Claudia Seibold (S) *Box-Counting: How to describe complexity?*	GuanYu Zheng *Measuring regional industrial structure and economic concentration in New Zealand*	
16:00	Emily Kawabata (S) *Modelling thickness variability in tephra deposition*	Evan Caygill *Work experience at Statistics New Zealand*	Andrea Collevecchio *Generalized preferential attachment schemes*
16:20	Peter Green (S) *Faster palaeoclimate reconstruction using monotone RegEM*	Amanda Hughes *Data visualisation at Statistics New Zealand*	Yevhen Mohylevskyy (S) (16:30) *Ergodicity and percolation for variants of one-dimensional voter models.*
16:40	Brendon Brewer *Probabilistic catalogues in astronomy*	Patrick Graham *Bayesian methods for population based microsimulation models*	
17:00	AGM, until 18:00 (Castle 2)	Young Statisticians' Meeting, until 19:00 (Common Room, Mathematics and Statistics, Third Floor, Science III)	
19:30	Conference Dinner (Cumberland College) – please arrive around 19:15		

Friday 30 November			
9:00	Patty Solomon (Castle 2) Chair: Peter Herbison *Statistical analysis of hospital performance: understanding the uncertainty*		
	Medical 1 (Castle 2) Chair: Sheila Williams	Testing 2 (Burns 1) Chair: Austina Clark	Stats Education 1 (Burns 2) Chair: John Harraway
10:00	Roy Costilla *Estimating the healthcare costs of smoking: A proposal for the case of New Zealand*	Thomas Rolf Turner *A remark on Monte Carlo p-values*	Sharleen Forbes *The coming of age of statistics education in New Zealand*
10:20	Myron Chang *Designs for Randomized Phase II Clinical Trials*	Robert Davies *Hypothesis testing when a nuisance parameter is present only under the alternative*	Doug Stirling *The future of textbooks and evolution of an e-book*
10:40	Break		
	Medical 2 (Castle 2) Chair: Claire Cameron	Official Statistics 2 (Burns 1) Chair: TBA	Stats Education 2 (Burns 2) Chair: TBA
11:00	Alain Vandal *Dimensional reduction for automated classification of Alzheimer's Disease MRI brain volumes*	Anna MacDonald & Lena Rodnyanskiy *Transforming census*	Jeanette Chapman *Genstat for Teaching and Learning (GTL) in the high school setting*
11:20	Martin Hazelton *Estimation of spatial relative risk by local smoothing*	Mike Camden *Confidentiality for Statistics NZ's Integrated Data Infrastructure prototype*	John Harraway *Teaching bootstrapping visually: a teacher's perspective.*
11:40	Tilman Davies *Modelling dichotomously-marked muscle fibre configurations*	Laura O' Sullivan *Data integration and the IDI (Integrated Data Infrastructure) at Statistics New Zealand*	Stephanie Budgett *Dynamic visualisations and the randomisation test*
12:00	Priya Parmar *Polymorphisms in genes within the IGF-axis influence antenatal and postnatal growth*	Asheel Ramanlal *Improving access to microdata: enhancing data utility & safety*	Austina S S Clark *Undergraduate Retention and Completion Rates and Factors for Maori Students at University of Otago*
12:20	Lunch		
13:20	Thomas Lumley (Castle 2) Chair: Ken Dodds *Two million t-tests: issues in genome-wide association*		
	Genetics (Castle 2) Chair: Ken Dodds	Modelling 2 (Burns 1) Chair: David Fletcher	Reliability & Time Series (Burns 2) Chair: Tilman Davies
14:20	Benoit Liquet *A novel approach for biomarker selection and the integration of repeated measures experiments from two assays*	Siva Ganesh *Comparison of some statistical models for identifying critical source areas of nitrogen in cattle grazed hill pastures*	Richard Arnold *Inference for Multicomponent Systems with Dependent Failures*
14:40	Steffen Klaere *Do your data fit your phylogenetic tree*	Alasdair Noble *A modeling exercise with many permutations*	Peter Thomson *A hidden seasonal switching model for multisite daily rainfall*
15:00	David Bryant *How to lasso positively, quickly and correctly*	Geoffrey Jones *Trying to herd cats*	
15:20	Break / End		

Abstracts

Plenary talks

Thomas Lumley

University of Auckland

Two million t-tests: issues in genome-wide association

Genome-wide association studies measure hundreds of thousands of genetic markers and use them to find small regions of the genome where genetic variation is associated with disease or with other interesting biological variables. The typical analysis uses statistical methods from Stage 1 and Stage 2 introductory stats courses, but still provides interesting statistical challenges in asymptotics, model choice, sample spaces, and other issues.

Roger Payne

VSN International

Hierarchical generalized models - theory and practice

Hierarchical generalized linear models (HGLMs) extend the familiar generalized linear models (GLMs) by allowing you to include additional random terms in the linear predictor. However, they do not constrain these terms to follow a Normal distribution nor to have an identity link, as e.g. in generalized linear mixed models. So they provide a richer of class of models that may be more intuitively appealing. The methodology provides improved estimation methods that reduce bias, by the use of the exact likelihood or extended Laplace approximations. In particular, the Laplace approximations seem to avoid the biases that are often found when binary data are analysed by generalized linear mixed models.

The algorithm involves fitting two (or more) interlinked GLMs, firstly to estimate the fixed and random effects in the model that describes the mean, and secondly to model the dispersion of the random terms. So all the familiar model checking techniques are available. We can also exploit other GLM extensions such as prediction and the inclusion of nonlinear parameters in the linear predictor.

The theory will be explained, with examples using GenStat to illustrate its usefulness in practical data analysis.

Alastair Scott

University of Auckland

Fitting regression models with response-dependent samples

We are interested in fitting regression models to data from samples when we do not have complete information on all members of the sample. In particular, we look at situations where the probability of missing data for a unit depends, at least in part, on the value of the response of that unit. Case-control studies, where the selection probabilities depend directly on the outcome, are simple examples. We look at examples of related studies, as well as studies where the dependence on the response is more subtle.

When the chance of missing data depends on the response, the likelihood involves the distribution of the explanatory variables as well as the regression parameters. We certainly do not want to have to model this covariate distribution in general, so we look for semi-parametric methods that avoid the need for such modelling. We develop fully efficient semi-parametric methods for some situations and good, practical procedures for situations where full efficiency is not feasible.

Patty Solomon

University of Adelaide

Statistical analysis of hospital performance: understanding the uncertainty

Critical care is expensive with costs increasing inexorably as our populations age. Understandably, governments at all levels want accurate measures of hospital performance to provide a basis for planning, for accountability and to inform public debate. However, provider comparisons via league tables have proven to be methodologically challenging as well as politically controversial, as witness the recent inquiry into Australia's Bundaberg Base Hospital. Furthermore, analyses purporting to measure hospital performance often suffer from a number of serious deficiencies, ranging from inadequate adjustment for patient case-mix to no allowance for multiple comparisons. In this talk, I will discuss these issues and present our recent work on comparing hospital performance using the Australian and New Zealand Intensive Care Adult Patient Database, one of the largest databases of its kind in the world.

Contributed talks

Richard Arnold

Victoria University of Wellington

Inference for Multicomponent Systems with Dependent Failures

We present a general approach to inference in Independent Overlapping Subsystem models, where a component's failure time is the time of the earliest failure in all of the subsystems of which it is a part, and each of those subsystems has an independent failure process. We apply this method to observations of an IOS model that associates individual shock processes with sets of overlapping subsystems made up of groupings of components, giving examples for various system configurations (series, parallel, and other arrangements).

Julie Bessac

IRMAR

Gaussian state-space model for wind fields in the North-East Atlantic

We propose a stochastic space-time model for wind fields at regional scale in the North-East Atlantic. This work aims at developing stochastic models which can generate realistic wind conditions and be used to estimate various related risks (renewable energy, coastal erosion,...).

We use a gaussian linear state-space model in which the hidden state represents the mean circulation at the regional scale and the observation equation relates the regional conditions to the local ones. One of the goal of the model is to reproduce space-time motions of the meteorological systems as, for example, the propagation of a storm in the channel. The observation equation of the state-space model describes the spatial structure of the variables and time structure is contained in model equation.

The estimation strategy is based on maximum likelihood (EM algorithm). Estimation is done on reanalysis data from ECMWF. The model is validated by comparing statistics of the data with those computed from artificial realizations of the model.

This model does not allow to correctly reproduce the weather regimes existing in this area. The next step could be to add an extra layer of hidden variables with values in a finite state space to describe the various regional weather types.

Brigid Betz-Stablein

Massey University

Predicting glaucoma progression: A novel spatial analysis

Glaucoma is the second leading cause of blindness worldwide. Caused by an increased pressure within the eye, glaucoma 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. Our model is extended to include several physiological features such as adjacent loci on the visual field map not being adjacent on the optic disk, the presence of the blind spot, and large measurement error dependent on location. We employ conditional autoregressive priors, weighted to account for the physiological correlations in the eye, to describe spatial and spatio-temporal correlation in the mean response over the visual field, and we regard the discrete visual field responses themselves as being censored below at zero on a log-scale (to account for threshold effects in the measurement process). The models are fitted within a Bayesian framework and implemented using Metropolis-Hastings algorithms.

Penelope Bilton

Massey University

Decision tree models for poverty mapping

Elimination of extreme poverty by 2015, the first of the United Nation's Millennium Development Goals, is being addressed by the World Bank through expenditure of billions of dollars of aid in the poorest countries in the world. Optimal distribution of aid resources is ensured by employing a statistical model to provide estimates, at low geographical levels, of poverty measures, which are then incorporated into a poverty map. ELL, the current standard methodology for poverty mapping, utilises a linear regression model. The talk discusses application of decision tree techniques to poverty mapping as an alternative to ELL. The challenges inherent in the research include incorporating complex survey design elements, weighting, stratification and clustering. Variability of poverty estimates must be obtained indirectly through some type of variance estimation procedure, such as replication, jackknifing, bootstrapping or inverse sampling. A small area estimation procedure must also be included, to avoid imprecise estimates which can arise from small sample sizes. Utilising decision tree methodology with complex survey data could have applications in other research areas. The ultimate goal of the study, however, is more efficient modelling of poverty, to facilitate better allocation of the billions of dollars currently spent on aid funding.

Brendon Brewer

The University of Auckland

Probabilistic catalogues in astronomy

A fundamental problem in astronomy is the construction of catalogues from raw image data. Traditionally, a heuristic procedure is run on the raw images in order to produce a list of sources (e.g. stars) that are present in the image. However, the traditional approaches can fail when multiple sources overlap or are very faint. This motivates a probabilistic (Bayesian) approach to making catalogues, where the question is, "What is our state of knowledge about the objects in the image, given the data?" Instead of creating a single catalogue (a point estimate), the result is a posterior distribution over the space of possible catalogues. I will demonstrate this approach and contrast the results with those obtained from traditional methods. I will also discuss issues related to Bayesian computation in large problems.

David Bryant

University of Otago

How to lasso positively, quickly and correctly

The lasso (Tibshirani, 1996) is shrinkage method which is particularly popular because it provides an automatic method for variable selection. The LARS-Lasso algorithm of Efron et al. (2004) makes the lasso computational feasible: with little added computational cost one can construct lasso solutions for all ranges of the shrinkage parameter. In many applications, however, model considerations impose a non-negativity constraint on the variable coefficients, giving need for the "positive lasso". Efron et al (2004) propose a LARS type algorithm for the positive lasso, but it is not guaranteed to produce optimal solutions. Here we describe a modification to the LARS algorithm which does produce optimal positive lasso solutions, and illustrate the algorithm using an example from phylogenetics.

Stephanie Budgett

The University of Auckland

Dynamic visualisations and the randomisation test

This is a joint presentation with Maxine Pfannkuch. Hypothesis testing is recognised as a difficult area for students of introductory statistics. Changing to a new paradigm for learning inference through computer intensive methods rather than mathematical methods is a pathway that may be more successful. A large collaborative project explored ways of improving students' inferential reasoning at the Year 13 and Stage One university levels. Part of this project involved development of new learning trajectories and dynamic visualisations for the randomisation test. We report on student outcomes from a pilot study, modifications which were made in light of these outcomes, and some initial findings from the main study involving over 3000 students. We discuss how the randomisation test using dynamic visualisations clarifies some of the concepts underpinning inferential reasoning, and why the nature of the inferential argument still remains a challenge.

Mike Camden

Statistics New Zealand

Confidentiality for Statistics NZ's Integrated Data Infrastructure prototype

The Integrated Data Infrastructure prototype (IDI) is a major step forward in Statistics NZ's provision of access to microdata. The IDI data is integrated, longitudinal, and mostly from full-coverage administrative sources. Since the start of this year, a group of researchers from across the official statistics sector has been exploring IDI data and producing results from it. The Confidentiality Rules for IDI enables these results to be as useful as possible, and safe in terms of confidentiality.

We describe: the IDI, the Rules, the many issues that arose in building the rules, our solutions, and our outstanding concerns. A major issue is consistency in the way we treat outputs: across data from different sources, and across different teams that produce output in different contexts. An issue that produces surprises is the presence of social and business data in the same structure.

IDI output contains examples from a very large part of the output types that a statistical office produces. The project contains new challenges, and causes us to question how we think about all aspects of confidentiality. We outline conceptual frameworks for handling these challenges.

Andrew Cathie

SAS Institute

Some Stories from the Field

What are some of the applications that statisticians are doing in NZ and overseas? SAS Institute has a diverse range of customers in many industries. This talk briefly covers some of the interesting topics that we find our customers are doing in NZ and around our world. These cases are all forecasting at heart, however bring some other interesting features in to the mix.

Evan Caygill

Statistics New Zealand

Work experience at Statistics New Zealand

Statistics New Zealand - Tatauranga Aotearoa - is a government department and New Zealand's national statistical office. We have been publishing headline national statistics (e.g. GDP, unemployment rate) for nearly 120 years. We have three offices; one each in Auckland, Wellington, and Christchurch, with around 1,000 staff in total. As a new addition to these 1,000 or so staff, and as someone who was not a part of the graduate program I will discuss my experience working at NZ's national statistical office. More specifically, I will cover my working experience so far in the Statistical Methods section. We will look into what is Statistical Methods with reference to the Statistics New Zealand generic business process model and where we fit in as a part of the wider organisation.

Statistics 2020 - Te Kapehu Whetu - is our vision for creating a statistical system for the future, and is the programme of change that supports our vision, and our progress. Finally, we will look at the Statistics 2020 projects led by Statistical Methods with a view to implementing Statistics New Zealand's long term vision.

Myron Chang

University of Florida

Designs for Randomized Phase II Clinical Trials

The most common primary statistical endpoint of a phase II clinical trial is the categorization of a patient as either a 'responder' or 'non-responder'. The primary objective of typical randomized phase II anti-cancer clinical trials is to evaluate experimental treatments that potentially will increase response rate over a historical baseline and select one to consider for further study. We propose single- and two-stage designs for randomized phase II clinical trials, precisely defining various type I error rates and powers in order to achieve this objective. We develop a program to compute these error rates and powers exactly, and we provide many design examples to satisfy prefixed requirements on error rates and powers. Finally, we apply our method to a randomized phase II trial in patients with relapsed non-Hodgkin's disease.

Jeanette Chapman

Otago Girls' High School

Genstat for Teaching and Learning (GTL) in the high school setting

GTL has been used very successfully for three years at Otago Girls' High School. The program is used from Year 9 to Year 13 by all teachers in the Mathematics Department. Students have used GTL successfully to analyse the data for their own investigations. This presentation will cover the implementation of GTL at Otago Girls' High School. The advantages and disadvantages of using GTL will be discussed as well as suggestions for successful use in the school setting. Information about resources to help you implement its use will also be given. How GTL fits with the new re-aligned Achievement Standards at Years 11 to 13 will also be covered.

Austina S S Clark

University of Otago

Undergraduate Retention and Completion Rates and Factors for Maori Students at University of Otago

The success of undergraduate students in completing their degrees is one of the most important issues for universities. Reasons for this include 'wastage' (of human potential, and financial cost), concerns about the reputation of an institution, and concerns of not meeting broader educational responsibilities.

From the previous studies done at University of Otago (Clark et al, 2009), we realize that the Maori students has a lower completion rate compared to the overall completion rate for the University, here we try to establish the possible reasons behind this.

The main questions we sought to answer were: 1) what are the similarities and differences in completion rates between Maori students and the domestic European students at our university? 2) what are the most important factors contributing to the success of completing a degree for Maori students?

Andrea Collevecchio

Ca' Foscari University

Generalized preferential attachment schemes

We study a general preferential attachment model. At each step a new vertex is introduced, which can be connected to at most one existing vertex. If it is disconnected, it becomes a pioneer vertex. Given that it is not disconnected, it joins an existing pioneer vertex with a probability proportional to a function of the degree of that vertex. This function is allowed to be vertex-dependent, and is called reinforcement function. We prove that there can be at most three phases in this model, depending on the behavior of the reinforcement function. Consider the set whose elements are the vertices with cardinality tending a.s. to infinity. We prove that this set either is empty, or it has exactly one element, or it contains all the pioneer vertices.

Roy Costilla

Ministry of Health

Estimating the healthcare costs of smoking: A proposal for the case of New Zealand

This talk discusses methodologies to estimate the healthcare costs of smoking in New Zealand. After reviewing previous studies, it compares potential approaches, describes available data sources, and discusses strengths and limitations of the methodologies.

Two approaches are proposed. The prevalence approach, or annual excess costs of smoking, which measures the healthcare costs associated with smoking for a given year; and the incidence approach, or lifetime estimation of the healthcare costs of smoking, that aims to capture the costs trade-off between smokers and non-smokers, namely that smokers are more expensive to treat but have a shorter life expectancy.

It is proposed that healthcare costs are predicted using generalised-linear models where costs are a function of smoking status and several socioeconomic and health covariates. Health Survey-administrative linked data over several financial years (2006/07-2010/11) and life tables for smokers and non-smokers are proposed as sources for these estimations.

Robert Davies

Statistics Research Associates

Hypothesis testing when a nuisance parameter is present only under the alternative

One of the papers announcing the successful Higgs boson search says, "The global significance of a local 5.9 sigma excess anywhere in the mass range 110-600 GeV is estimated to be approximately 5.1 sigma ...". The relationship between these two sigma values seems to be derived from my 1977 paper on hypothesis testing. This application provides a good excuse for me to review my paper and some of its subsequent applications.

Tilman Davies

University of Otago

Modelling dichotomously-marked muscle fibre configurations

Human skeletal muscle consists of contractile elements (fibres) that may be differentiated according to their physiological and biochemical properties. The different types of fibre are distributed throughout each muscle, with the pattern of cell distribution being an important determinant of the functional properties of each muscle. It is well known that the proportions and distributions of muscle fibre types change with advancing age, but few studies have quantitatively investigated these changes. Several statistical methods designed to gauge the departure of a dichotomously-labelled muscle fibre distribution from that of a random fibre-type dispersal are discussed and tested. These methods are also applicable to a wide range of biological investigations in which the spatial distribution of cells or specimens underpins an important biological principle. This work includes the proposal of a novel technique, based on weighted kernel-smoothed density-ratios, which can account for the variable areas of the individual fibres. The methodology is illustrated using a number of real-data examples, and a comprehensive set of simulations is employed to assess the empirical power and false-positive rates of these tests.

Daniel Fernandez

Victoria University of Wellington

Likelihood-based finite mixture models for ordinal data

Many of the methods to deal with the reduction of dimensionality in matrices of data are based on mathematical techniques such as distance-based algorithms or matrix decomposition and eigenvalues. In general, it is not possible to use statistical inference with these techniques because there is no underlying probability model. Recent research has been developed a group of likelihood-based finite mixture models for a data matrix with binary or count data, using basic Bernoulli or Poisson building blocks (Pledger and Arnold 2012, under review). My current research intends to undertake an extension of that research and establishes likelihood-based multivariate methods for a data matrix with ordinal data. My talk will introduce the first results from this research, which applies fuzzy clustering via finite mixtures to a newly developed model for ordinal data: the ordered stereotype model. In addition, I will show the results of a simulation experiment which includes a variety of scenarios in order to test the reliability of the proposed model. Finally, I will show the results of the application of the model in a real example.

David Fletcher

University of Otago

A new method for estimating overdispersion in generalized linear models

I consider the problem of fitting a generalized linear model to overdispersed data, focussing on a quasi-likelihood approach in which the variance is assumed to be proportional to that specified by the model, and the constant of proportionality, ϕ, is used to obtain appropriate standard errors and model comparisons. It is common practice to base an estimate of ϕ on Pearson's lack-of-fit statistic, with or without Farrington's modification. I propose a new estimator that has a smaller variance, subject to a condition on the third moment of the response variable. I conjecture that this condition is likely to be achieved for the important special cases of count and binomial data. I illustrate the benefits of the new estimator using simulations for both count and binomial data.

Sharleen Forbes

Statistics New Zealand

The coming of age of statistics education in New Zealand

For some time, New Zealand has been leading the world in terms of the focus and scope of its statistics curriculum in schools. This curriculum is characterised by its data handling, and in more recent years, data visualisation approach. In 2013 bootstrapping and randomisation will be added to the senior secondary school. This paper gives an historical perspective of the people and groups that have influenced the development of the curriculum including Professors Jim Campbell and David Vere-Jones, Geoff Jowett, the Wellington group of Megan Clark, Mike Camden and the author through to the Auckland group led by Professor Chris Wild and Maxine Pfannkuch. The role of the New Zealand Statistical Association is also discussed as is the possible long-term impacts of the school curriculum on tertiary statistics teaching.

Siva Ganesh

AgResearch Limited

Comparison of some statistical models for identifying critical source areas of nitrogen in cattle grazed hill pastures

A critical source area (CSA) is an area, usually occupying small parts of a farm, with a large source of nutrient or faecal contaminants. For determining CSA, an important assumption is that there is a strong correlation between the time an animal spends lying in an area of the paddock and the proportion of the total daily urination events deposited there.

In this study, we have developed a process to predict where cattle will lie in a random paddock for which only slope, aspect, elevation, Northing and Easting values are known for each (5m x 5m) grid cell overlying the paddock. The training data came from a cattle grazed hill country pasture in which cows wore a GPS collar to track their movement. Cows were also fitted with a motion sensor to differentiate between walking, grazing, standing and lying activities.

This presentation summarises the results from fitting models such as Multiple Linear Regression (MLR), Geographically Weighted Regression (GWR), Random Forest Regression (RFR), k-Nearest Neighbour Regression (kNNR) and Generalized Additive Model (GAM) for predicting the time a cow spends lying in a grid cell.

Patrick Graham

University of Otago

Bayesian methods for population based microsimulation models

In health and social sciences microsimulation models are used to explore the potential impact of alternative policies and programmes. The characteristic feature of these models is the generation of a large number of individual life histories from a, probability model which often embodies a latent process such as a model of the natural history of disease. Methodology for microsimulation studies appears to have developed in an ad-hoc manner and largely outside of a formal statistical inference framework. In this paper I outline a general Bayesian framework for population based microsimulation studies. Viewed from this perspective, microsimulation studies are an exercise in posterior predictive inference. Accounting for parameter uncertainty is automatic in the Bayesian framework but treated by ad-hoc sensitivity analyses in some microsimulation models. Another issue that is clarified by the Bayesian perspective is model calibration which is simply posterior computation in the Bayesian framework but is approached using a variety of strategies in the existing microsimulation literature. However, Bayesian computation for complex microsimulation models presents some challenges which will be discussed in the context of developing a microsimulation model for colorectal cancer in New Zealand.

Peter Green

University of Otago

Faster palaeoclimate reconstruction using monotone RegEM

Many recent multiproxy palaeoclimate reconstructions have used the RegEM algorithm, which is an expectation-maximisation algorithm designed to find penalised maximum likelihood estimates given rank-deficient multivariate climate data.

The setup of palaeoclimate reconstruction problems means that the EM algorithm converges very slowly, making simulation studies of RegEM based reconstruction methods particularly difficult. However, the patterns of missing data in this kind of problem are are approximately monotone, which means that substantially faster algorithms are available.

Maximum penalised likelihood estimates have a natural interpretation in the Bayesian framework as maximum a posteriori estimate. RegEM estimates can therefore be used to construct approximate posterior distributions, allowing a Bayesian analysis of the uncertainties in past temperature estimates.

John Harraway

University of Otago

Teaching bootstrapping visually: a teacher's perspective

This is a joint work with Sharleen Forbes, Victoria University

In 2013 bootstrapping will be introduced to the senior secondary school Mathematics and Statistics curriculum. Auckland University have designed a teaching sequence that uses bootstrapping to introduce confidence interval concepts in the classroom. Dynamic visualization software, iNZSight, was also developed to enable students to see the resampling process, and the creation of the bootstrap distribution, in action. The authors of this paper are part of the research team that piloted this method in school classrooms, first year university courses and the workplace. Although the formal research results are not yet available the authors will introduce the visual approach and their separate experiences in teaching in first year statistics at Otago University and in Statistics New Zealand.

Martin Hazelton

Massey University

Estimation of spatial relative risk by local smoothing

The spatial relative risk function is an important tool for examining geographical patterns of disease. It is defined as the ratio of bivariate densities of the spatial coordinates of cases and controls for the disease of interest, and can be estimated using kernel density estimates constructed from case-control data. The estimated relative risk function is usually displayed on a log-scale, in part so that case and control densities are handled in a comparable manner.

The choice of smoothing regimen is critical to the performance of the estimated log-relative risk function. However, the development of effective, data-driven methods for doing some have proven challenging. In particular, fixed bandwidth kernel estimation of the log-relative risk often performs very poorly. No single bandwidth is likely to work well across the entirety of the estimation region, and moreover, available data-driven methods for finding the fixed bandwidth that is 'the best of a bad job' have unreliable performance.

We describe an alternative adaptive smoothing regimen, in which the bandwidths are local to the estimation point. We show that such an approach facilitates the development of data-driven bandwidth selectors with strong theoretical properties, and produces pleasing estimates of the spatial relative risk function in practice.

Lyndal Henden

Massey University

Estimation of Population Size for Small Communities in Indonesia

Indonesia is a maritime nation formed by over 17,000 islands that hosts a wide range of linguistic, ethnic and genetic diversity. We have samples from 50 small communities in Indonesia where the history of these villages remains largely unknown. Using Mitochondrial DNA sequence data, we aim to reconstruct whether these populations have expanded or contracted in the recent past by estimating their effective population size.

We are able to simulate the evolutionary process and resulting genetic data using a state-of-the-art coalescent model that assumes the communities are independent and of constant size. However, we are faced with the challenge that the resulting distribution theory is intractable. This means that, although we can simulate data, we are unable to use classical methods of parameter estimation, such as the method of maximum likelihood. One method for parameter estimation that is becoming increasingly popular is Approximate Bayesian Computation, which allows us to approximate the posterior distribution using the simulated data. Our work examines the implementation of Approximate Bayesian Computation to the Indonesian data as well as two other methods not widely used in population genetics, namely simulation-based total least squares and generalized least squares.

Amanda Hughes

Statistics New Zealand

Data visualisation at Statistics New Zealand

Over the past year Statistics New Zealand has made great progress in the area of data visualisation. Statistics New Zealand's project Telling Stories Using Official Statistics aims to "improve Statistics New Zealand's ability to tell statistical stories in words, pictures, and tables and present them in ways that are easily understood." This project, along with an increase in data visualisation activity from across the organisation, has led to great progress in this area.

My talk outlines work in data visualisation that took place in the organisation in the past year, current work, and future opportunities. I will speak about projects from across the organisation and show some of our best graphics.

Beatrix Jones

Massey University

A model-list approach to using decomposable models for Gaussian (inverse) covariance selection

Gaussian graphical models have some unique properties, and also some properties that are common to high dimensional model selection problems. In the Bayesian setting, one of their unique attributes is the relative ease of evaluating the marginal likelihood for a subset of the model space, the set of decomposable models. Even in this restricted class, as in other high dimensional problems, the model space is sufficiently large that true posterior sampling is quite difficult. Metropolis Hastings and other MCMC algorithms become merely heuristic search options. One approach in this situation is to retain a large set of high-posterior models, and consider the posterior conditional on this set. We will consider this approach when a decomposable model is fit, and the true underlying model is non-decomposable. There are both positive and negatives to the approach: the set of models taken together often points quite clearly to the true underlying non-decomposable model. However, this dance around the true model leads to a list of very similar models, which minimizes the model list's ability to represent model uncertainty.

Geoffrey Jones

Massey University

Trying to herd cats

In which our hero attempts to compare the fits of a nonlinear mixed model and a restricted b-spline mixed model on a small but annoying dataset. The use (or mis-use) of the nlme() and lmer() functions in R will feature strongly. Ruminations will be offered on the difficulty of finding an empirical model that realistically incorporates known or desired features. Trying to get the fitted curves to do what you want them to can be a bit like, well, trying to herd cats.

Khair Jones

Massey University

Shape constrained penalised spline regression for generalised linear models

In classical generalised linear models, numerical covariates in the linear predictor component usually appear as linear or low order polynomial terms. For more complicated relationships, the effect of such covariates on the linear predictor can be described using nonparametric techniques such as penalized splines. However, at times this approach may produce too much flexibility, and it may be desirable to constrain the shape of the splines in some way.

In this talk, the adaptation of shape constrained spline regression to generalised linear models will be considered. By using quadratic penalized splines we are able to express some common shape restrictions like monotonicity and convexity as linear constraints on model parameters. Our proposed method employs an MCMC independence sampler. The proposal distribution is a multivariate normal distribution derived from the sampling distribution of the parameter estimates from an unconstrained fitted model and then truncated to satisfying the aforementioned constraints.

To illustrate the method, we consider a model for predicting scores based on FIFA ratings in international football, where it is assumed that a larger positive difference in rating between a team and its opposition will lead to an increase in the number of goals it scores.

Murray Jorgensen

University of Waikato

The sex life of untagged wetas

A study of the use of artificial refuges by weta in a stand of Kahikatea in Hamilton presents some challenges for data analysis. The site was visited 45 times over 455 days and the number of adult male wetas and adult female wetas and adult female wetas in each of 80 refuges was recorded.

There were pronounced changes in the refuge habitation patterns over the period of the study. We describe a Generalized Linear Mixed Model approach to studying how the increase (and decrease) of the number of females (and males) between visits depends on the initial habitation pattern of the refuge. [Joint work with Priscilla M Wehi and Mary Morgan-Richards]

Emily Kawabata

Massey University

Modelling thickness variability in tephra deposition

One of the major hazards from volcanic eruptions is the dispersal of rocks and ash, collectively called tephra, at considerable distances from the volcano. The thickness of tephra fall deposits often decreases with distance from the source. A tephra attenuation model as a function of distance and direction from the source can be useful in estimating tephra thickness at a given location.

Isopachs, contours of tephra thickness, are generally fitted to the model in recent studies. However, drawing isopachs involves varying degrees of subjectivity. Here, we will fit our models to actual tephra measurements using maximum likelihood estimation. This way the variability (sampling error) in the thickness measurements can be expressed explicitly.

The lognormal, Weibull and gamma distributions have been considered to describe the variability for the 1973 Heimaey eruption. The Weibull and gamma fitted the data similarly well. The estimated 'effective volume', mean wind direction/s and attenuation rate/s from the two error models matched with the observations.

The Weibull distribution has been considered for the 1977 Ukinrek Maars eruptions. For multiple source eruptions, the model is implemented in a mixture framework to account for the multiple lobes and/or vents and identify the source and direction of tephra deposits.

Steffen Klaere

University of Auckland

Do your data fit your phylogenetic tree

Phylogenetic methods are used to infer ancestral relationships based on genetic and morphological data. What started as more sophisticated clustering has now become a more and more complex machinery of estimating ancestral processes and divergence times. One major branch of inference is maximum likelihood methods. Here, one selects the parameters from a given model class for which the data are more likely to occur than for any other set of parameters of the same model class. Most analysis of real data is executed using such methods.

However, one step of statistical inference that has little exposure to application is the goodness of fit test between inferred model and data. Recently, methods to detect sections of the data which do not support the inferred model have been proposed, and strategies to explain these differences have been devised. In this talk I will present and discuss some of these methods, their shortcomings and possible ways of improving them.

John Koolaard

AgResearch

Package Predictmean: Further inference from linear models

This is joint work with Dongwen Luo.

An R package 'Predictmean' will be introduced by examples. This package provides further inference and graphs for various linear models, including 'aov', 'lm', 'glm', 'lme', and 'lmer'. The core function 'predicted.means' calculates predicted means, SE of means, SEDs and LSDs between means, and means plots with SED or LSD bars. The package also has the facility to run specified post-hoc comparisons between means. In addition, for mixed-effects models, diagnostic residual plots can be produced. In summary, 'Predictmean', combined with R packages 'nlme' or 'lme4' provide a useful set of inferential tools for mixed effects models in R.

Frances Krsinich

Statistics New Zealand

Decomposing property price indexes into land and structure components

New Zealand's Quotable Value House Price Index (QVHPI) uses a sales price appraisal ratio (SPAR) method to adjust the prices of houses sold for the changing composition of houses sold each quarter. The data is rich in characteristics of the houses sold and Statistics New Zealand has been using this characteristics information to produce benchmark hedonic indexes against which the SPAR method can be compared. In the course of this work we have also been looking at the potential for utilising the land and structure components of the valuation data, along with sale prices and characteristics of houses sold, to estimate separate price indexes for land and structures.

Benoit Liquet

MRC Biostatistics Unit

A novel approach for biomarker selection and the integration of repeated measures experiments from two assays

High throughput 'omics' experiments are usually designed to compare changes observed between different conditions (or interventions) and to identify biomarkers capable of characterizing each condition. We consider the complex structure of repeated measurements from different assays where different conditions are applied on the same subjects.

We propose a two-step analysis combining a multilevel approach and a multivariate approach to reveal the effects of conditions within subjects separately from the biological variation between subjects. The approach is extended to two-factors designs and to the integration of two matched data sets. It allows internal variable selection to highlight genes able to discriminate the net condition effect within subjects.

The approach was applied to an HIV-vaccine trial evaluating the response with gene expression and cytokine secretion. The discriminant multilevel analysis selected a relevant subset of genes while the integrative multilevel analysis highlighted clusters of genes and cytokines that were highly correlated across the samples. Our combined multilevel multivariate approach may help in finding signatures of vaccine effect and allows for a better understanding of immunological mechanisms activated by the intervention. The integrative analysis revealed clusters of genes that were associated with cytokine secretion. These clusters can be seen as gene signatures to predict future cytokine response.

Anna MacDonald

Statistics New Zealand

Transforming census

This is a joint work with Lena Rodnyanskiy.

Next year Statistics New Zealand will be running the 33rd New Zealand Census of Populations and Dwellings. Over the last 150 years the way in which census forms have been delivered and collected has remained mostly unchanged. The current model requires recruitment of a large temporary labour force. The significant cost associated with this model, changes in society and international best practice mean that the way we collect census information needs to evolve. In this presentation we will give a history of the census and the developments we are considering for the future.

Kylie Maxwell

Statistics New Zealand

Identifying respondent groups in a longitudinal business survey sample: an application of dual latent trajectory modelling

Latent trajectory models (Jones et al., 2001, Nagin & Tremblay, 2001) can be used to classify different patterns of growth and decline across subjects over time. We apply these models to better understand the survey response behaviour of New Zealand businesses over time. Based on model results we classify businesses into: Great Responders, Good Responders, Learners, Over-time Drop-offs and Bad Responders. We find response behaviour in the first five months of a longitudinal survey is a strong predictor of future response behaviour. We also show how business characteristics contribute to predicting response behaviour.

Understanding these patterns is useful in the field when collecting longitudinal survey data. Our analysis provides insights into early detection of problematic responders, targeting of non-response intervention and a possible dynamic sample design in a longitudinal context.

References:

Jones, B. L., Nagin, D. S., and Roeder. K., 2001. A SAS Procedure Based on Mixture Models for estimating Developmental Trajectories. *Sociological Methods & Research* 29:374-93.

Nagin, D. S. and Tremblay, R. E., 2001. Analyzing Developmental Trajectories of Distinct but Related Behaviors: A Group-Based Method. *Psychological Methods* 6 (1): 18-34.

Graham McBride

NIWA

A sensible weight-of-evidence from three simultaneous hypothesis tests?

In environmental science it is still common for referees and editors to request results for tests of "nil" hypotheses. These are generally false a priori, yet a test's P-value is often used as a weight-of-evidence for the existence of an effect. Some have responded by proposing a "three-valued logic" in which test outcomes concern inference about the direction of change, not its magnitude: (i) Confidence that the change is in the positive direction; (ii) Confidence that the change is in the negative direction; (iii) Insufficient data to infer direction. While pleasing, this leaves unanswered questions associated with the magnitude of change. To address this we suggest accompanying this three-valued logic with two simultaneous "equivalence tests" (for the equivalence hypothesis and for the inequivalence hypothesis). Each of these has two permissible outcomes, so that the combined procedure could have twelve (but four of them are impossible). A table of permissible combinations has been constructed to facilitate judgement about the weight-of-evidence conferred by the collected data. It can also indicate whether further sampling is likely to strengthen the overall conclusion, which would be rather helpful to the funders of sampling programmes.

This is joint work with Ian Westbrooke, DOC, Christchurch

Barry McDonald

Massey University

Taking the numbers on faith?

This non-technical talk will discuss some sources of religion data, and possible demographic trends.

Religions in New Zealand and worldwide are undergoing dramatic demographic changes. In Europe, North America and the Pacific non-Religion is on the increase, largely at the expense of Christianity. By contrast in Africa and Asia, Islam and Christianity are both growing strongly. Demographic changes are taking place within Christianity as well. Protestantism is rising in Latin America at the expense of Catholicism, while in New Zealand churches are becoming more conservative and more fragmented. Will these factors eventually arrest or reverse the decline in religion?

Russell Millar

University of Auckland

A modified DIC for comparison of Bayesian state-space models

The deviance information criterion (DIC) has become very popular in applied Bayesian research, including applications where state-space population dynamics models are fitted to animal abundance data. DIC of state- space models is routinely calculated by measuring the fit of the latent states to the observations, however the structure of candidate models being compared often differs at the process (population dynamics) level. This mismatch can result in the stochastic components of the process model compensating for inadequacies in models, leading to erroneous model choice. Increasingly this phenomenon is occurring in practice. Here, a form of partial deviance is developed, and its advantages are demonstrated using simple examples, and using a population dynamics model encountered in the modeling of coho salmon.

Yevhen Mohylevskyy

University of Auckland

Ergodicity and percolation for variants of one-dimensional voter models

We study variants of one-dimensional q-color nearest-neighbor voter models in discrete time. In addition to the usual voter model transitions in which a color is chosen from the left or right neighbor of a site there are two types of noisy transitions. One is bulk nucleation where a new random color is chosen. The other is boundary nucleation where a random color is chosen only if the two neighbors have distinct colors. We prove under a variety of conditions on q and the magnitudes of the two noise parameters that the system is ergodic, i.e., there is convergence to a unique invariant distribution. The methods are percolation-based using the graphical representation of the model which consists of coalescing random walks combined with branching (boundary nucleation) and dying (bulk nucleation).

Elena Moltchanova

University of Canterbury

LEGO figurines as a tool of assessing emotions

Previously, we have asked the study participants to rate the emotional expression of 94 different LEGO Minifigure faces in an online questionnaire. These data have been used to produce a LEGO-based scale of emotional assessment for the 6 basic emotions (anger, disgust, fear, happiness, sadness, surprise), each rated from weak (1) to intense (5). The aim of this new study has been to validate the newly constructed scale. In order to do this, volunteers were asked to rate photographs of human faces with various expressions using either (i) the newly constructed LEGO-scale, (ii) its stylized version or (iii) the standard numeric scale.

The comparisons were made using multinomial logistic regression fitted within a Bayesian framework. We will present the results as well as discuss various interpretation problems such as the absence of 'ground truth' and the aspects of analyzing categorical data.

Charles Newman

New York University – Courant Institute

Coarsening Models

Coarsening models are continuous time Markov processes whose states are the assignments of one of two possible values (say +1 or -1) to the vertices of some (usually infinite) graph like Z^d (with nearest-neighbor edges) or a homogeneous tree. The transition rules (which are the zero temperature limit of stochastic Ising models) are that at rate one each vertex updates by adjusting to agree with a strict majority of its neighbors or in the event of a tie, tosses a fair coin. One is often interested in an initial state in which sites choose values independently with probability p of being +1. These models have been or can be used to study evolution in time of spatial structure in materials or in voting preferences. Among the questions of interest are, for $p = 1/2$, whether sites change preference infinitely often and, for $p > 1/2$, whether all sites are eventually +1, and how the answers to these questions depend on the underlying graph. We will review some old results about Z^d for $d \leq 2$ and recent somewhat unexpected results (jointly with Michael Damron and Vladas Sidoravicius) about two dimensional slabs. There are many open problems.

Alasdair Noble

Plant and Food Research

A modeling exercise with many permutations

A developmental stage in wheat is hypothesized to be controlled by a combination of two genes each of which responds to temperature. An experiment had been carried out some time ago that produced data which could be used to investigate the relationships between two developmental variables and temperature. This was not the intention of the experiment so the data did not cover the ranges of the variables particularly well. A range of models were tried and the combinations of base models and relationships with temperature ensured a time consuming process. I will discuss the process which resulted in a useful outcome though not in the way I had expected.

Laura O'Sullivan

Statistics New Zealand

Data integration and the IDI (Integrated Data Infrastructure) at Statistics New Zealand

Data integration is defined broadly as the combination of data from different sources about the same or a similar individual or unit. This definition includes linkages between survey and administrative data, as well as between data from two or more administrative sources. One of the current priorities at Statistics New Zealand is to make as much use of administrative data as possible. As part of this initiative the Integrated Data Infrastructure (IDI) was founded.

I will present an overview of the IDI with focus on the linking of student loans and allowances data to ministry of education data. These datasets will be described along with the variables used for linking. The importance of getting to know your data will also be demonstrated. The linkage methodology will be summarised and some of the challenges and successes that arose when linking these two datasets will be presented. The results of the linking will be shown including some discussion of quality.

Priya Parmar

Auckland University of Technology

Polymorphisms in genes within the IGF-axis influence antenatal and postnatal growth

Two pregnancy cohorts were used to investigate the association between SNPs in genes within the insulin-like growth factor (IGF) axis and antenatal and postnatal growth from birth to adolescence. Longitudinal analyses were conducted in the Raine pregnancy cohort ($n = 1,162$) using repeated measures of fetal head circumference (HC), abdominal circumference(AC) and femur length (FL) from 18-38 weeks gestation and eight measures of postnatal height and weight (1-17 years). Replications of significant antenatal associations were undertaken in the Generation R Study ($n = 2,642$). Of the SNPs within the IGF-axis genes, 40% ($n = 58$) were associated with measures of antenatal growth ($p \leq 0.05$). The majority of these SNPs were in receptors; IGF-1R (23%; $n = 34$) and IGF-2R (13%; $n = 9$). Fifteen SNPs were associated with antenatal growth in Raine ($p \leq 0.005$): five of which remained significant after adjusting for multiple testing. Four of these replicated in Generation R. Associations were identified between 38% ($n = 55$) of the IGF-axis SNPs and postnatal height and weight; 21% in IGF-1R ($n = 31$) and 9% in IGF-2R ($n = 13$). Twenty-six SNPs were significantly associated with both antenatal and postnatal growth; 17 with discordant effects and nine with concordant effects. New analytic methods are required to better understand this key metabolic pathway integrating biologic knowledge about the interaction between IGF-axis genes.

Maryann Pirie

University of Auckland

Do kauri trees experience old age?

Visual inspection of tree-ring series for kauri trees shows that old trees tend to have more narrow growth rings. This leads to a concern that kauri may have the potential to experience old age, where trees nearing the end of their lives struggle to grow. The implication of this aging effect is that the variation within ring widths may be influenced by age and is not related to local climate. If an aging effect is present, data from these outer rings of old trees should be removed before reconstructing past climates.

I will first give an overview of how kauri trees are used to understand past climates and then describe a Bayesian multilevel hierarchical approach for estimating the age of kauri trees. Finally, I compare the common signal and climate response for mature and very old kauri trees.

Shirley Pledger

Victoria University of Wellington

Capture-recapture models for estimating breeding return times and abundance

Many animals exhibit breeding site fidelity. They may also have intermediate non-breeding years when they do not return to the site. Existing temporary emigration methodology provides estimates of birth, death and temporary emigration parameters, provided Pollock's robust design is used for the sampling. However, for some populations only a simple (Jolly-Seber type) sampling scheme is possible.

This is the case with lake sturgeon in Black Lake MI, USA which return to Black River in the years when they are spawning. They are caught on only one occasion per year, at the time of spawning, and only at the spawning site in the river. No information from their non-breeding location in the lake is available.

We show that even with such limited information, it is possible to build likelihood-based hidden-Markov models, to do model selection and to obtain plausible estimates of return times and abundance. These models will be described using the sturgeon data for illustration.

Asheel Ramanlal

Statistics New Zealand

Improving access to microdata: enhancing data utility & safety

Parliament has recently approved a key amendment to the Statistics Act 1975. Now a wide range of researchers will be able to apply for access to microdata through our data laboratory for valid research or statistical purposes. This is a way for us to create more opportunities for data users to access valuable data and develop new insights into New Zealand's economic and social structure.

We will discuss what microdata access means to the data laboratory; take a high-level view of what microdata means to Statistics NZ, such as the organisation's confidentiality standard for microdata access, the confidentialised unit record file system, and the synthetic unit record file system; discuss the organisation's framework of the 'five safes'; look at recent and exciting developments in the data laboratory such as the organisation's move to a high-trust environment, the accredited researcher scheme, the use of remote access.

We will look at work being done by Dr Barry Milne. This looks at constructing a micro simulation system for policy makers by creating a synthetic birth cohort of individuals from 2006 Census data. This sits outside our standard processes and we will look how we manage the issues in terms of the '5 safes'.

David Scott

University of Auckland

The tail probabilities of the generalized inverse Gaussian distribution

Slevinsky and Safouhi (2010) used the G-transformation to approximate the incomplete Bessel function and hence the tail probabilities of the generalized inverse Gaussian distribution. However there are cases where the method fails, due in part to the way the incomplete Bessel function is specified in the literature. Instead, it is better to work the other way: obtain the G-transformation for the tail probabilities of the generalized inverse Gaussian distribution and use that to evaluate the incomplete Bessel function.

I will describe the G-transformation and the improvements I have made to Slevinsky and Safouhi's work which result in a far more reliable algorithm.

Claudia Seibold

University of Canterbury

Box-Counting: How to describe complexity?

The most important characteristics of a fractal object, its self-similarity and non-integer dimension, are useful for describing complex patterns. Compared to mathematically generated artificial fractal objects which can show exact self-similarity, most objects in the real world are only statistically self-similar. Research in analysing complex real world objects in terms of fractals are motivated by the possibility of modelling the complexity and reflecting possible changes on different scales. The method for calculating fractal dimensions considered in this presentation is box-counting which is applicable not only to geological objects, e.g. river networks, but also to other complex spatial patterns such as traffic networks in civil engineering, the brain in neuroscience or time series of stock indices in economics. It is intuitive and therefore popular. However, box-counting includes a drawback: its simplicity might be the reason for a lack of standards on how to apply this method. There are several well-thought-out programs provided but their resulting fractal dimensions differ. Divergent fractal dimensions result from decisions made in three steps: converting data, counting boxes and calculating fractal dimension. We introduce the history of fractals and their application, and focusing on the box-counting we discuss not only the advantages but also the disadvantages of applying this method.

Doug Stirling

Massey University

The future of textbooks and evolution of an e-book

This talk identifies problems with existing paper-based textbooks that are caused by their static format. Publishers are moving towards delivery of textbooks on tablets and laptops, but the requirement to profit from them imposes constraints that usually result in books with the static nature and inflexibility of their paper versions.

CAST started life as a single e-book whose initial goals were to use interactive diagrams for active learning within a textbook and to use dynamic diagrams to explain concepts more clearly. Various changes since then have turned CAST into a flexible framework that allows a customised e-book to be created for any group of students.

The core of CAST now includes alternative versions of pages with examples from different application areas, summary versions of pages, and the ability to print complete chapters. Versions of pages with videos are currently being prepared. A customisation tool allows anyone to generate an e-book with arbitrary content and ordering, including user-supplied pages.

Examples from CAST will be used to illustrate its evolution, and possible future directions will be discussed.

Peter Thomson

Statistics Research Associates

A hidden seasonal switching model for multisite daily rainfall (joint with T. Carey-Smith, NIWA, and J. Sansom, NIWA)

A hidden seasonal switching model for daily rainfall over a region is proposed where season onset times are stochastic and can vary from year to year. The model allows seasons to occur earlier or later than expected and have varying length. This additional seasonal variation leads to increased rainfall variability over and above that explained by the standard fixed seasons.

In essence, the model dynamically classifies daily rainfall into seasons whose onsets vary from year to year and within which the model parameters are assumed to be time homogeneous. A variety of non-seasonal models could have been used to describe daily rainfall within seasons. Here a generalisation of the Richardson model is adopted and it is further assumed that it is only the dynamics of rainfall states that vary from season to season.

A suitable estimation strategy based on maximum likelihood and the EM algorithm is developed for fitting the model across a region. This strategy is validated on simulated data and various forms of the model are fitted to daily rainfall measurements from selected New Zealand sites. These results are discussed and compared to those from fitting standard fixed season models.

Daniel Turek

University of Otago

A new approach to model-averaged confidence intervals

When model-averaging, use of a standard Wald confidence interval for the parameter of interest can be problematic, as estimation of the standard error of a model-averaged estimate is difficult. I propose a new approach to constructing model-averaged confidence intervals. Confidence limits are defined as the values such that the model-weighted average of the nominal error rates from each candidate single-model interval are equal to the required nominal rate. This technique can also be used to construct model-averaged profile likelihood confidence intervals. I present the results of a simulation study comparing the coverage rate and width of these intervals against existing model-averaged intervals, both frequentist and Bayesian.

Thomas Rolf Turner

University of Auckland

A remark on Monte Carlo p-values

There are many hypothesis testing settings in which one can calculate a "reasonable" test statistic, but the null distribution of this statistics is unknown and/or completely intractable. In turn, for many such situations, it is possible to simulate values of the test statistic under the null hypothesis, in which case one can determine a Monte Carlo p-value (which is "exact" in a certain sense) provided that there are no ties in the data. I was recently interested in a scenario in which there are lots of ties - the null distribution of the test statistic has a point mass. It turns out that one can modify the usual procedure for calculating a Monte Carlo p-value to handle this setting. Dealing with this procedure leads to an intriguing identity involving the binomial probability function and its derivatives. I will briefly explain the modified procedure and discuss simulation studies which demonstrate its efficacy.

Alain Vandal

AUT University

Dimensional reduction for automated classification of Alzheimer's Disease MRI brain volumes

Alzheimer's Disease (AD) is hard to diagnose using clinical observation. Better screening and diagnostic procedures are sought. Classifiers such as discriminant analysis and support vector machines are applied to dimensionally reduced images, stabilising the classifier by getting rid of noise. When data sport high within-group variability, yet large common structures, reducing dimensions by keeping only the first principal components may bury or jettison subtle differences between groups.

To reduce the dimensions of brain images from subjects with mild cognitive impairment and those with probable AD, we used Krzanowski's angle (JASA, 1979) to identify and remove components generating similar subspaces in both groups, either because they are similar in structure (first PCs) or noisy (last PCs). We use a simple thresholding mechanism based on the sequence of angles created. Peeling components from both ends in this fashion yields two sets of minimally similar principal components.

We compared this method, dubbed SimMinPCA, with PCA and logistic dimensional reduction. The combination of SimMinPCA for dimensional reduction and a cost-optimised support vector machine for classification outperformed all other combinations based on the criteria of area under the ROC curve and the Youden index.

Joint work with Anoukh van Giessen, Geurt Jongblood and D. Louis Collins.

Zoé van Havre

Queensland University of Technology & Universite Paris Dauphine

Investigating the number of components in overfitted Gaussian mixture models

Finite mixture models with an unknown number of components pose a complex mathematical and computational challenge, yet are commonly encountered in today's increasingly complex datasets and research problems. When too many components are included the true parameters fall within an unidentifiable subset of the larger parameter space making estimation difficult, and this escalates as the number of components and dimensions increases.

Recent advances in the asymptotic theory of overfitted mixture models by Rousseau & Mengersen (2011) led us to investigate the impact of the prior on the weights in overfitted Gaussian mixture models. As a result of this investigation we propose a new technique to estimate the number of components in Gaussian mixtures using overfitting with a dynamic prior, and show that it provides a very effective and comprehensible solution to estimate the number of components as well as their parameters. This method is simple to implement using a Gibbs sampler and bypasses any need for model selection or complicated trans-dimensional steps, as well as automatically dealing with the label-switching problem.

We present this method and apply it to a range of datasets, with a focus on its performance on a selection of increasingly complex case studies from various fields.

Darcy Webber

Victoria University of Wellington

Broad scale management in spatially heterogeneous fisheries - does it matter?

In fisheries science, stock assessment models are used to estimate the current and historic size of fish populations. These models often assume the population to be discrete and spatially homogenous. In reality, fish populations are far from spatially homogeneous and although the importance of accounting for spatial population structure in stock assessments is acknowledged, it is not yet fully understood due to the complexities inherent in modelling fisheries dynamics and biological systems. However, before complex spatial models are developed we must ask, at what point does space begin to matter? Or, do our current models do an adequate job? To test these hypotheses we use Antarctic toothfish (Dissostichus mawsoni) data to inform an operating model from which we can simulate the "truth" and generate pseudo data sets with known parameter values. We then apply standard stock assessment methods similar to those used to manage many of New Zealand's fish stocks. Comparisons of model estimates with the known parameter values are made to probe the performance of standard stock assessment models. Initial results indicate that not taking spatial heterogeneity into account can result in biased estimates of the population size, an outcome that might result in poor management decisions.

Janine Wright

Otago University

Modelling Māori Language

In New Zealand, recent reviews have indicated that there is a declining percentage of the New Zealand population speaking Māori. On the basis of these reviews, a Māori Language Strategy is being developed and as part of this strategy, one proposal is that there is a national language target of 80% of Māori speaking Māori language by 2050.

Following a developing international trend to model changes in language use statistically, we have developed a model specific to the New Zealand situation that allows us to examine how various language policy choices affect language usage over successive generations. Our aim is to show whether such a language target is achievable, how it might be measured and which language policy choices are most likely to maintain or potentially grow inter-generational Māori language use.

We hope that a model such as ours will provide evidence to assist policy makers to consider the potential effects of their current choices on future Māori language speakers.

Haizhen Wu

Massey University

Design of variables acceptance sampling plans for non-normally distributed quality characteristics

Acceptance sampling is established to be cost optimal in the presence of measurement uncertainty when compared to none or all inspection strategies. Variables acceptance sampling plans are usually considered superior to their attributes counterparts in terms of requiring smaller sample sizes, but they need an extra assumption on the distribution of target quality characteristics.

The current industry practices and standards are based on the normal distribution. This assumption is not legitimate for many quality characteristics, in particular those supported in half line or bounded intervals. Apart from the distribution support, other distributional features of the real data such as skewness and kurtosis cannot be described by normal distribution and a popular remedy of applying normalizing transformation method is also proven to be invalid. This fact drives the demand for developing variables acceptance sampling plans for non-normal underlying distributions.

In this presentation, we will reveal that the key difficulty of developing non-normal based variables acceptance sampling plans comes from the dependence of the operating characteristic curves on the shape parameters. We will show that, but replacing the acceptance/reject decision rule from the traditional k-method approach to the estimated proportion nonconforming approach, the impact of shape parameters can be partially eliminated.

Lingyun (Larry) Zhang

BNU-HKBU United International College

The exact likelihood ratio test for equality of two normal populations

Testing the equality of two independent normal populations is a perfect case of the two sample problems, yet it is not treated in the main text of any textbook or handbook. In this talk, I will shed new light on the solution of this two sample problem.

GuanYu Zheng
New Zealand Productivity Commission

Measuring regional industrial structure and economic concentration in New Zealand

The spatial organisation of economic activity has a potentially important influence on a country's rate of economic growth, innovation and productivity. As such, understanding the New Zealand economy from a spatial perspective could be important in identifying the causes of New Zealand's poor economic performance relative to a number of other OECD countries. As part of that work, this paper provides a descriptive analysis on industrial structure and agglomeration in New Zealand regional economies at the Territorial Local Authority level. This is a useful starting point for examining differences in regional economic performance and possible reasons underlying these differences, including the role of agglomeration forces. Results offer insights into the following questions: i. how diverse are New Zealand's regional economies?; ii. Which New Zealand industries are geographically concentrated?; iii. Which New Zealand regions are more specialised?

Posters

Benoit Liquet (Cambridge University); Philippe Delorme (University of Montréal); Pierre Lafaye de Micheaux (University of Montréal); Riou Jérémy (Danone Research)

Power and sample size computations for multiple-endpoints: finding at least r among m significant hypotheses.

Generally, in multiple endpoints situations we want to reject all hypotheses or at least only one of them. For some time now, we see emerge the need to answer the question : "Can we reject at least r hypotheses?" However, the statistical tools to answer this new problem are rare in the literature. We decide to develop general power formulas for the standard procedures : Bonferroni's, Hochberg's and Holm's procedures. We also develop an R package for the sample size calculation for multiple endpoints, when we want to reject at least r hypotheses. We limit ourselves to the case where all the variables are continuous and we present four different situations depending on the structure of the data's variance-covariance matrix.

M.F. Parry (University of Otago), G.J. Gibson (Heriot-Watt University), S. Parnell (Rothamsted Research), T.R. Gottwald (United States Department of Agriculture), M.S. Irey (U.S. Sugar Corporations), T. Gast (U.S. Sugar Corporations) & C.A.Gilligan (University of Cambridge)

Inference for spatiotemporal models of an arboreal epidemic in the presence of disease control

The in-orchard spread of Huanglongbing (HLB) is used as a case study for modelling an emerging epidemic in the presence of disease control measures. Specifically, the spread of the disease is modelled as a spatially explicit SEIDR epidemic, where the exposure and infectious times are not observed, detection times are censored, and removal times are known. Control measures are accounted for via time dependence of the infectious process and seasonal and host-age effects are included in the model of the latent period. Parameters are estimated in different sub-regions of a large commercially cultivated orchard, and are used to establish a temporal pattern of invasion, age dependence of the dispersal parameters, and a close to linear relationship between primary and secondary infectious rates. The results support an exponential dispersal kernel and an annual sinusoidal variation in the latent period. The resulting model can be used to simulate HLB epidemics to assess economic costs and potential benefits of planned control strategies.

Kate Richards, Martin Hazelton, Mark Stevenson (Massey University)

Assessment of the effectiveness of intervention strategies to control animal diseases using the inhomogeneous K-function

Foot-and-mouth disease (FMD) can affect all types of cloven-hoofed animals. It is one of the world's top 10 agricultural diseases, with great economic impact. Intervention strategies play a major part in the controlling of outbreaks. For the control and eradication in the 2001 UK epidemic, a movement ban was put in place for all susceptible species and a strategy of slaughtering, burning and burial of all FMD-susceptible livestock on infected premises and on farms within a 1.5 km radius of infected premises was implemented. Other countries (usually those where FMD is endemic) use vaccination campaigns. The question then arises as to how we should assess the relative performance of multiple different methods of disease control.

If a disease control method is working well we would expect spatial clustering of infected premises to be short lived, while for poor methods clustering would typically persist. We can therefore learn about the effectiveness of control strategies by examining the temporal changes in the pattern of disease clustering. To this end, we generated simulated spatio-temporal datasets of FMD outbreaks in Ireland with a variety of control strategies applied. We then examined temporal changes to the inhomogeneous K-function and pair correlation function.

Alison Sefton (Massey University)

Novel methods of assessing viability of small area estimation in poverty

Small area estimation is an important analytical tool, which is used to estimate the level of poverty at finer levels of aggregation than is possible with traditional methods or sample surveys alone. A consumption based methodology of small area estimation was developed by Elbers, Lanjouw and Lanjouw (ELL) (2003) by combining sample survey and census data, providing a greater level of precision of estimates at aggregated levels. However it has the disadvantage of being very computer intensive due to the number of predictions needed to determine the level of poverty of each individual household. The core aim at the modelling stage is to find good diagnostics to test whether it is possible to remove variables at the modelling stage, that (were the full computational procedure have been carried out) would have later proved least relevant to generating sound small area estimates. Therefore I investigate novel methods to assess the selection of variables in the ELL methodology using the Census and Socio- Economic survey data from Cambodia. This considers not only the auxiliary variables ability to explain the variation in the expenditure of an individual, but also its ability to distinguish between the relative levels of poverty in the small areas.

Jimmy Zeng (University of Otago)

Model averaging for regression models with count data

The most common complication when faced with modeling count data is the potential of overdispersion. The default method for dealing with overdispersion is either to adapt a quasi-likelihood approach, or to fit the data explicitly using negative binomial models. We compare the performances of these two approaches in the model-averaging setting, where model-averaging is commonly used to make allowance for model uncertainty in parameter estimation.