Abstracts are ordered in session order for oral presentations (including professional development workshops) followed by poster presentations

1.1 Contributed - Statistics & the Law

Tuesday 2 September 9.40am - 10.40am

Statistical evaluation of forensic DNA evidence

Roberto Puch-Solis
LGC Forensics, Tamworth, Staffs., UK

Forensic DNA is one of the most powerful evidence types nowadays. The technology for producing profiles is continually improving and, at present, DNA systems are very fast and sensitive. A profile can be produced within a few hours and from a fragment of a cell. For example, profiles are obtained from touch DNA, i.e. from DNA contained in organic material left when a person touches an object.

The sensitivity of DNA systems posits interpretation challenges for several reasons: many profiles are mixtures, that is profiles coming from more than one person; artefacts of the process; insufficient DNA for producing all the components of a person's profile (dropout); small amount of spurious DNA sometimes present in laboratory consumables (dropin) such as swabs and tubes; and multiple profiles need to be interpreted simultaneously (replicates).

Part of the evaluation of the DNA evidence in a court of law consists of quantifying the evidential value of a profile obtained from the scene of a crime together with the profile of a defendant. In this talk a statistical method for the quantification of DNA evidence is presented. The method consists of existing and novel models that can simultaneously take account of the phenomena described above. The method combined the rarity of the profile components and the quality of the crime scene profile. The introduction of this method into casework will have an impact on the quality of evidential evaluation of DNA and on the number of cases presented in court.

1.1 Contributed - Statistics & the Law

Tuesday 2 September 9.40am - 10.40am

Using chain event graphs to address asymmetric evidence in legal reasoning

Anjali Mazumder, James Smith
University of Warwick, Coventry, UK

In forensic science, evaluation of evidence is expressed using the likelihood ratio. Barristers are tasked with expressing unfolding events to relate individuals to the evidence. Recently Bayesian networks (BNs) have been useful in providing a graphical representation of the problem, calculating marginal and conditional probabilities of interest, and making inferences particularly addressing propositions about the source of an evidential-sample. To address propositions relating to activities, there is a need to account for different plausible explanations of suspect/perpetrator's actions and events as it relates to the evidence. We propose the use of chain event graphs (CEGs), exploiting event tree structures to depict unfolding events as postulated by each side (defence and prosecution) and differing explanations. Different explanations can introduce different sets of relevant information affecting dependence relationships between variables and symmetric structure. Deriving a Kullback-Leibler measure, we can assign a score to different sub-trees comparing separation relationship among vertices to assess ambiguous explanations. CEGs are a flexible class of graphical models which can model the asymmetric story structure directly in its topology; because of its graph modular structure it inherits many benefits of BNs which are not always rich enough to incorporate all obtainable information. Using complex case examples (involving transfer and persistence) we show how CEGs can be useful in addressing activity level propositions by directly supporting the barrister's argument within the topology of a graph and in pairing and development of propositions by addressing the uncertainty in evidence evaluation and asymmetric unfolding of events to better assist the courts.

1.1 Contributed - Statistics & the Law

Tuesday 2 September 9.40am - 10.40am

Calibrated Probabilities and the Investigation of Soft Fraud in Automobile Insurance Claims

Alex Lenkoski, Ingrid Hobæk Haff, Linda R. Neef, Anders Løland
Norwegian Computing Center, Oslo, Norway

The prevalence of soft fraud in automobile insurance claims is non-negligible and, if undetected, can lead to a significant additional cost for the insurer. From an operational perspective, it is desirable for an insurance company to verify the validity of claims and clear claimants as quickly and cost-effectively as possible. Insurers therefore require reliable probabilities that a given claim is valid. In this work, we discuss our experience collaborating with a major Nordic insurance company to develop such a robust system to verify claim validity. We construct an ensemble of statistical models that proves more accurate and--critically--more reliable than individual methods. Furthermore, we borrow results from recent research in the combination of forecasts from complex numerical models, and consider a non-linear pooling method to improve calibration over individual models.

In particular, we demonstrate how the beta linear pool can be used to join and recalibrate probability forecasts from a number of statistical methods. We conclude with an assessment of our framework's performance from a cost-benefit perspective. This shows that combined probabilities outperform individual methods and further adds to the evidence that beta linear pooling improves over the standard linear pool.

1.2 Contributed - Spatial Analysis

Tuesday 2 September 9.40am - 10.40am

Markov Chain Monte Carlo Methods for Bayesian Spatial Survival Analysis

Benjamin Taylor
Lancaster University, Lancaster, UK

This talk concerns the analysis of survival outcomes where the data are spatially referenced. Such data are of importance in the health sciences as (i) they can be used to identify spatial areas in which survival outcomes are poorer than might be expected and (ii) taking into account potential correlation between observations reduces bias in estimates of putative covariate effects. Bayesian inference for spatial survival models is challenging because routine Markov chain Monte Carlo (MCMC) methods used to deliver inference are computationally O(n^3) where n is the number of observations. In this talk we introduce a model-based solution to the issue of computational complexity.

Our model assumes a constant frailty for individuals belonging to a small grid cell, itself a member of a set of spatially-correlated auxiliary frailties on a regular grid. This allows us to exploit discrete Fourier transform techniques for the handling of matrix operations, thereby reducing the computational complexity to O(mlog m) where m is the number of cells on the grid. The main assumption of our model is that frailties arise as a result of exposure to environmental risk factors, rather than being attributable directly to individuals. We present an adaptive MCMC scheme for sampling from the posterior of our target and demonstrate our methods in a case study in which we use a parametric proportional hazards model for the survival outcomes.

1.2 Contributed - Spatial Analysis

Tuesday 2 September 9.40am - 10.40am

Simulating and Modeling Spatial Binary Processes

Gabrielle Kelly, Renhao Jin
University College Dublin, Dublin, Ireland

Objectives: The majority of methods for generating correlated binary data have limitations with respect to generating such data with non-constant mean, or algorithmic limitations, and cannot be extended to generate spatially correlated binary data. Two procedures considered here that do not have these shortcomings are a conditional approach and a marginal copulas method that is totally new. The procedures enable models for fitting spatial binary data to be compared.

Methods and Models: The conditional method is based on the definition of a conditional GLMM and involves generating spatial random effects from a multivariate Gaussian distribution with a spatial covariance structure.
The copulas method involves generating random variables from a multivariate Gaussian distribution with a spatial covariance structure, transforming them to uniform random variables, and generating binary variables with specified mean based on these uniforms.
In a simulation study, processes simulated from both methods are fitted using both conditional (GLMM) and marginal spatial (GLM) models and the results compared using several criteria including parameter estimation.
An example related to bovine TB incidence illustrates the models.

Results and Conclusions: Simulation results and the numerical example indicate that both types of processes are best modeled by spatial GLM's rather than GLMM's. Conditional models fitted to processes generated by the copulas method did not converge. The conditional method is not suitable for generating spatial binary processes with specified mean and variance while the copulas method also has a drawback in that a correlation structure cannot be specified in advance.

1.2 Contributed - Spatial Analysis

Tuesday 2 September 9.40am - 10.40am

Inference for spatial extremes using elliptical Pareto processes

Emeric Thibaud[1], Thomas Opitz[2], Anthony C. Davison[1]
[1]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, [2]Université Montpellier II, Montpellier, France

Complex extreme events, such as heat-waves and flooding, have major effects on human populations and environmental sustainability, and there is growing interest in modelling them realistically. For risk assessment based on spatial environmental processes, it is necessary to properly model the dependence among extremes at several locations, and understand how the dependence varies spatially.

During the last decade, max-stable processes have been used to model spatial maxima and threshold exceedances of random processes. In this talk we shift the focus to Pareto processes (Ferreira and de Haan, 2014, Bernoulli; Dombry and Ribatet, 2014, Statistics and Its Interface) which extend Generalized Pareto distributions and appear as the limits of threshold exceedances of random processes. We discuss inference for so-called elliptical Pareto processes which appear as the limits of certain elliptical processes. These processes provide a large class of asymptotically justified models for spatial extremes controlled by correlation functions. Inference for them can be based on a censored likelihood function, which is shown to be efficient and to improve the composite likelihood approach currently used for max-stable processes. If we have time, we will discuss the construction of a Bayesian model and an application to real data.

1.3 Contributed - Household security, Housing and Private Schools

Tuesday 2 September 9.40am - 10.40am

The Association between Burglary and Effective Household Security: Individual, Household and Area Mediating Factors.

Andromachi Tseloni, Rebecca Thompson
Loughborough University, Loughborough, UK

Domestic burglary has fallen in England and Wales by over 60 per cent since 1995 according to estimates based on the Crime Survey for England and Wales (CSEW). This paper forms part of a wider ESRC Secondary Data Analysis Phase 1 funded project which utilises CSEW data from 1992 to 2011/12 to explore the role of security in declining burglary rates. The project's objective is to address: 'Which burglary security devices work for whom and in what context'. The 2008/09 to 2011/12 CSEW data sets merged with the 2001 Census are analysed here via bivariate multilevel logit models in order to investigate the association between burglary risk and availability of effective security configurations across specific population subgroups and areas (LSOA's) in England and Wales. A number of socio-demographic, lifestyle and area characteristics denoting routine activities and social disorganisation theories are employed for improving both our theoretical understanding of crime patterns and the practical allocation of security improvements. After accounting for the theoretical mediating factors of the link between burglary risk and availability of effective security, any residual association is partitioned between households and between areas in England and Wales.

1.3 Contributed - Household security, Housing and Private Schools

Tuesday 2 September 9.40am - 10.40am

Using Statistics to Inform Policymaking: Housing in England

Jeremy Hicks, Natalie Cartwright
Department for Communities and Local Government, London, UK

Everyone needs somewhere to live, making the subject of housing a popular topic often seen in the media. Perhaps because most of us either own a home or aspire to home ownership, we are keenly aware of reports of house price changes and levels of supply. The task of developing policy on housing is a challenging task not least because of our rapidly increasing population, depressed house building in recent years and the difficulty of saving a deposit acting as a barrier to home ownership for many first-time buyers.

The knowledge we have about housing in England is directly attributable to the variety and quality of statistics that are collected on social, demographic, housing and economic variables. Timely and accurate statistics are vital to effective governance. In the policy environment statistics are increasingly important to help policymakers grasp the status quo, identify potential avenues of action and evaluate policies that have been implemented.

With the welcomed emphasis on evidence-based policy in recent years, policymakers in the Department for Communities and Local Government increasingly utilise available data in the formulation of policies. Within DCLG there is a strong team of analysts who produce statistics on various areas of housing, including: housing supply; property sales and prices; professionalising the private rented sector; Government schemes to assist home ownership; and homelessness and rough sleeping. I will illustrate how statisticians in the Department influence policy by producing and disseminating quality evidence to policy colleagues.

1.3 Contributed - Household security, Housing and Private Schools

Tuesday 2 September 9.40am - 10.40am

Are private schools for the poor actually better? Evidence from India

Sunil Mitra Kumar
King's College London, London, UK

Much recent policy advocacy in developing countries focuses on the apparent benefits of private schooling for children from poor households, given the often dismal performance of government schools. We use detailed panel data from India that tracks a cohort of children over 7 years to examine this question, comparing the achievement scores of those enrolled in private schools with those enrolled in government schools. Using difference-in-difference matching estimators based on the propensity score that take into account several socioeconomic household-level and community-level indicators, we compare both the long-run differences in scores depending on the type of school a child studies in, as well as the differences due to switching from a government to private school over a shorter period.

Unlike simple unadjusted or regression-adjusted comparisons which show children enrolled in private schools as having significantly higher scores than those enrolled in government schools, we find that difference-in-difference matching estimators based on the propensity score show slight and occasionally negative, statistically insignificant differences in their achievement scores. In other words, that once socioeconomic characteristics are suitably adjusted for, learning levels in private schools are no different from those in government schools. This finding is at odds with a significant part of the policy advocacy that focuses on private schooling for the poor, but is in consonance with a smaller literature of qualitative studies based on classroom observation that explain why children do not actually learn well in either type of school.

1.4 Contributed – Industry & Commerce

Tuesday 2 September 9.40am - 10.40am

All grown up? The fate of the quarter of a million UK firms born in 1998 after 15 years

Michael Anyadike-Danes[1,2]
[1]Aston Business School, Birmingham, UK, [2]Enterprise Research Centre, Coventry, UK

Modelling firm growth has attracted considerable attention. However, much of the literature has gone little further than testing whether it is possible to reject the hypothesis that firm growth is random (Gibrat's "Law"). We take a different approach using firms' 'life histories' to tame the extraordinary heterogeneity in performance. Making use of the ONS' Business Structure Database, which records basic data on all firms in the UK, we explore the evolution of a birth cohort (the 250,000 firms born in 1998) up to age 15. Survival and growth are very heavily age-dependent: the hazard of 'death' is very large for young firms (only 11% of the cohort survive to age 15); and job growth is highly concentrated in the early years (75% of 15 year job growth occurs by age 5). Size is also critical. Given age, larger firms typically have a better chance of surviving but record rather slower job growth. Finally, it is useful to distinguish firms which grow by adding jobs at their founding workplace from those which add workplaces. Although the central concern is a deeper understanding of firm and job dynamics, our analysis serves a broader purpose by contributing to the improvement of business support policy evaluation. Often it is thought impractical (or too expensive) to collect data on an untreated control group, but conditioning on age, size and workplace provides a simple (and cheap) method of constructing baseline controls for measuring the effects of interventions on firm survival and job growth.

1.4 Contributed - Industry & Commerce

Tuesday 2 September 9.40am - 10.40am

Non-parametric modelling of SME performance

MENG MA, Jake Ansell, Galina Andreeva
The University of Edinburgh Business School, Edinburgh, UK

Small and Medium sized Enterprises are a staple for most economies. Collectively they provide contribution to countries GDP and employment. Hence they are encouraged and supported by governments. This paper will explore modelling of SMEs' performance. In most credit scoring models, the underlying assumptions are that variables are normally distributed and linearly correlated with dependent variables. These assumptions have received considerable criticism, especially during financial crisis. The aim of the paper is then to establish whether alternative models will be better able to explain SME performance in terms of regressors during a time of financial distress (credit crunch). Both general additive models and non-parametric models will be considered as well as other functional models. The data consists of a large sample over the period from 2007 till 2010 covering. Since it is acknowledged that there is a difference between recently established (start-up) and established (non-start up) SMEs the analyses these separately. It is found there are different non-parametric effects for the two. The implications for SME lending and for government policy towards SMEs are considered.

1.4 Contributed - Industry & Commerce

Tuesday 2 September 9.40am - 10.40am

Efficient Modelling and Forecasting of the Electricity Spot Prices

Florian Ziel, Rick Steinert
Viadrina European University, Frankfurt(Oder), Germany

The raising importance of renewable energy, especially solar and wind power, led to new impacts on the formation of electricity prices. Hence, this talk introduces an econometric model for the hourly time series of electricity prices of the EEX which incorporates specific features like renewable energy.

The model consists of several sophisticated and established approaches and can be regarded as a periodic VAR-TARCH with wind power, solar power and load as influencing time series. It is able to map the distinct and well-known features of electricity prices in Germany and Austria. The used multivariate model has multi-seasonal/periodic structures within the mean, conditional mean, variance and conditional variance. The seasonal components are modelled using periodic B-splines. Furthermore it covers the impacts of public holidays and daylight saving time such as the volatility asymetry. Statistical tests underline various effects that we can quantify such as the change of price induced by building new wind or solar parks.

An efficient iteratively reweighted LARS-lasso algorithm is used as estimation technique. Moreover, it is shown that several existing models are outperformed regarding the forecasting performance.

1.5 Contributed - Environment & Ecology

Tuesday 2 September 9.40am - 10.40am

ESTIMATION OF STORM PEAK AND INTRA-STORM DIRECTIONAL-SEASONAL DESIGN CONDITIONS IN THE NORTH SEA

Philip Jonathan[1,4], David Randell[1,5], Graham Feld[2], Kevin Ewans[3], Yanyun Wu[1,4]
[1]Shell Projects & Technology, Manchester, UK, [2]Shell Projects & Technology, Aberdeen, UK, [3]Sarawak Shell, Kuala Lumpur, Malaysia, [4]Lancaster University, Lancaster, UK, [5]Durham University, Durham, UK

Specification of environmental design conditions for marine structures is of fundamental importance to their reliability. Design conditions are typically estimated by extreme value analysis of time series of measured or hindcast significant wave height, Hs.

Analysis is complicated by two effects. Firstly, Hs exhibits temporal dependence. Therefore, time series must be de-clustered into observations of (independent) storm peak significant wave height HsSP, and (intra-storm) directional dissipation of Hs conditional on HsSP. Extreme value analysis is then performed on HsSP providing a mechanism to simulate storm peak events for arbitrary return periods. Design values for Hs (for an arbitrary storm sea-state) are next estimated by incorporation of dissipation effects. Design distributions for individual maximum wave height Hmax (and other variables of interest) can then be estimated by marginalisation using the known conditional distribution for Hmax given HS. Secondly, HsSP is non-stationary with respect to multiple covariates, particularly wave direction and season. Failure to accommodate non-stationarity can lead to incorrect estimation of design values. Covariate effects in peaks over threshold of HsSP are modelled using non-stationary models for extreme value threshold, rate and size of occurrence of threshold exceedances. Model parameters are described as smooth functions of covariates using multidimensional penalised B-splines. Optimal parameter smoothness is estimated using cross-validation.

In this work, we develop directional-seasonal design values for HsSP, Hs and Hmax for a location in the North Sea. Attention is paid to the assessment of model bias and quantification of model parameter and design value uncertainty using bootstrap resampling.

1.5 Contributed - Environment & Ecology

Tuesday 2 September 9.40am - 10.40am

Estimating trends for bat populations; which method is best?

Steve Langton[2,1]
[1]*Defra, York, UK,* [2]*Freelance statistician, York, UK*

Data on British bird populations have been systematically collected since the 1960s and are currently analysed using Poisson GAMs, with confidence limits calculated by bootstrapping (Fewster et al, 2000). By contrast, a comprehensive national approach to monitoring bat populations only started in 1998 with the inception of the National Bat Monitoring Programme (NBMP). This uses a combination of roost counts, hibernation surveys and field surveys to produce population trends using the same statistical methodology used for birds.

Bird populations have also been widely monitored in mainland Europe and these country indices are combined to produce European indices. These use a different methodology based on Generalised Estimating Equations (GEE) and asymptotic confidence limits. A specialist package TRIM (http://www.cbs.nl/en-GB/menu/themas/natuur-milieu/methoden/trim/default.htm) has been written to perform the analysis. Recently, prototype European bat indices have also been produced, also using TRIM.

Following the statistical methodology of the more mature bird schemes is clearly a sensible starting point. However, bat data differs in its statistical properties from bird data, most notably in its high level of over-dispersion compared to a Poisson distribution. This paper compares the properties of the GAM and GEE approaches, using real and simulated data.

For some datasets the two approaches produce very similar results but for others the asymptotic GEE confidence limits are considerably narrower. Simulations suggest that this is because coverage of the GEE limits can be seriously non-conservative.

Fewster,R.M.; Buckland,S.T.; Siriwardena,G.M.; Baillie,S.R.; Wilson,J.D. (2000) Ecology 81 1970-1984

1.5 Contributed - Environment & Ecology

Tuesday 2 September 9.40am - 10.40am

Cross-validatory extreme value threshold selection and uncertainty with application to off-shore engineering

Paul Northrop, Nicolas Attalides
University College London, London, UK

When designing marine structures it is important to quantify the stochastic behaviour of extreme sea states. An important quantity is significant wave height (Hs), a measure of sea surface roughness. For a suitably high threshold asymptotic theory suggests that excesses of a threshold u may be modelled by a generalized Pareto (GP) distribution. However, inferences can depend strongly on the value of u chosen. Too low a threshold incurs bias due to model misspecification, too high a threshold reduces the number of excesses and thus increases the variance of parameter estimators. Existing methods of threshold selection do not address directly this bias-variance trade-off.

Objectives

To address the bias-variance trade-off by considering out-of-sample predictive performance at extreme levels, and to enable improved inferences by combining inferences appropriately from multiple thresholds.

Methods/models

Inferences are based on a Binomial model for the number of exceedances a high threshold and a GP model for sizes of threshold excesses. Bayesian inferences are performed using references priors. Leave-one-out (Bayesian) cross-validation is used to quantify extremal predictive performance based on different thresholds. Bayesian model averaging is used account for uncertainty in the choice of threshold.

Results and Conclusions

A simulation study shows that averaging inferences appropriately over many thresholds improves prediction of future extremes, when compared with inferences from a single threshold. Of two datasets, from the North Sea and the Gulf of Mexico, analyses suggest that the Gulf of Mexico has much greater potential to experience extremely stormy seas.

1.6 Contributed – Medical/bioinformatics

Tuesday 2 September 9.40am - 10.40am

The Monotone Splines Lasso: Nonparametric additive monotone regression for high-dimensional data

Linn Cecilie Bergersen, Kukatharmini Tharmaratnam, Ingrid K. Glad
Department of Mathematics, University of Oslo, Oslo, Norway

The problems of variable selection and estimation in nonparametric additive regression models for high-dimensional data will be addressed. Several methods have been proposed to model nonlinear relationships when the number of covariates exceeds the number of observations by using spline basis functions and group penalties. Nonlinear **monotone** effects on the response play a central role in many situations, in particular in medicine and biology. The monotone splines lasso (MS-lasso) is constructed to select variables and estimate effects using monotone splines (I-splines). The additive components in the model are represented by their I-spline basis function expansion and the component selection becomes that of selecting the groups of coefficients in the I-spline basis function expansion. A recent procedure, called cooperative lasso, is used to select sign-coherent groups, that is, selecting the groups with either exclusively non-negative or non-positive coefficients. This leads to the selection of important covariates that have nonlinear monotone increasing or decreasing effect on the response. An adaptive version of the MS-lasso reduces both the bias and the number of false positive selections considerably. The MS-lasso and the adaptive MS-lasso are compared with other existing methods for variable selection in high dimensions by simulation and the methods are applied to relevant genomic data sets. Results indicate that the (adaptive) MS-lasso has excellent properties compared to the other methods both by means of estimation and selection.

1.6 Contributed – Medical/bioinformatics

Tuesday 2 September 9.40am - 10.40am

Approaches for parametric time-to-event analysis with informative entry times, with application to an HCV study of time to cirrhosis from infection

Brian Tom, Vernon Farewell, Sheila Bird
MRC Biostatistics Unit, Cambridge, UK

In this talk, we examine maximum and pseudo score approaches for the analysis of prevalence data arising from a referral cohort where entry into the cohort is dependent on a subject's residual fraction of time remaining to the event of interest, and inference on the incident population is required. Such data are believed to occur in hepatitis C virus (HCV) studies conducted in tertiary care settings, where HCV patients are more likely to be referred to specialist clinics at later stages of disease. The conventional truncation likelihood approach which simply conditions on the time of entry into the cohort does not work here as the referral time and time to the event are correlated. The ignoring of this referral bias has led to higher rates of progression to cirrhosis being reported in studies in specialist clinics compared to those in community-based settings. As cirrhosis linked to HCV infection is a major epidemic of the 21st century, it is therefore extremely important to get an accurate picture of the present and future disease burden facing affected regions in order to inform public health decisions and actions.

1.6 Contributed – Medical/bioinformatics

Tuesday 2 September 9.40am - 10.40am

Accounting for nonignorable missingness: A simulation study comparing method performance under model misspecification

Finbarr Leacy[1,2], Ian White[1], Sian Floyd[3], Tom Yates[4]
[1]MRC Biostatistics Unit, Cambridge, UK, [2]University of Cambridge, Cambridge, UK, [3]Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK, [4]Research Department of Infection and Population Health, University College London, London, UK

Objectives: This work seeks to compare the performance of maximum likelihood, Bayesian full probability modelling and multiple imputation approaches for handling nonignorable missingness in a single covariate under misspecification of the missingness and/or covariate models and assuming that the analysis model is a generalised linear model. It further aims to provide a framework for structured sensitivity analysis in this context.

Methods: We present results from a simulation study comparing the performance of the following methods: EM by the Method of Weights, Bayesian full probability modelling, multiple imputation with delta-adjustment and multiple imputation with re-weighting. We consider settings in which the covariate and missingness models are both correctly specified as well as settings in which at least one of these models is incorrectly specified. We assume that the analysis model is always correctly specified. We illustrate the methods using data from the Zambia South Africa Tuberculosis and AIDS Reduction trial, exploring the impact on inference of nonignorable missingness in the HIV test result variable across a range of sensitivity analyses.

Results: Misspecification of the covariate or missingness models can be associated with significant bias in estimates of the analysis model coefficients. Sensitivity analyses based on misspecified models can produce inferences that are inconsistent with those obtained under the true model.

Conclusions: As the true missingness mechanism cannot be determined on the basis of the observed data alone, sensitivity analyses that consider possible misspecification of the covariate and/or missingness models should form a central component of all practical analyses of incomplete data.

Plenary 1 - Campion (President's Invited) Lecture

Tuesday 2 September 10.45am - 11.30am

The Challenges of Operating at Scale

Richard Allan
Facebook

Facebook operates a service that is used by over 1.3 billion around the world. This creates novel challenges in terms of organising information so that it is available and useful to such a large community. Complex processes have been developed across many aspects of the service from how to present the most relevant stories in a small screen to being able to detect and prevent abuse of the service.

Richard Allan will share insights from his five years at Facebook into how these challenges present themselves and are addressed.

2.2 Invited - New advances in multivariate modelling techniques

Tuesday 2 September 12 noon - 1.20pm

Current challenges in multivariate modelling

Marta García-Fiñana, Gabriela Czanner
University of Liverpool, Liverpool, UK

Multivariate modelling techniques are often applied to analyse complex data sets (e.g., measurements collected over time and from different locations). Variable reduction procedures, classification methods, multilevel latent class modelling and cluster analysis are some of the techniques used. They can be applied to address research questions such as '*is there a natural clustering in the dataset?*' and '*can we predict whether a person will develop a particular condition/disease by looking at a range of clinical variables over time*?

In this introductory talk, we discuss several current challenges in multivariate modelling. We illustrate how the discriminatory strength of a classifier can improve when longitudinal information is taken into account, and when an existing heterogeneity of the variance-covariance between groups is not ignored. We also briefly sketch an approach for sample size calculation within the framework of linear discriminant analysis.

2.2 Invited - New advances in multivariate modelling techniques

Tuesday 2 September 12 noon - 1.20pm

Multilevel latent class modelling of colorectal cancer survival status, incorporating stage of disease

Wendy Harrison[1], Mark S Gilthorpe[1], Amy Downing[2], Paul D Baxter[1]
[1]Division of Epidemiology & Biostatistics, University of Leeds, Leeds, UK, [2]Cancer Epidemiology Group, University of Leeds, Leeds, UK

Previous studies investigating survival from colorectal cancer have typically considered stage of disease as a potential confounder. Stage however may lie on the causal path and statistical adjustment with stage as a confounder can introduce bias known as the reversal paradox. Classification of stage may also be imprecise and incomplete. Modelling using Latent Class Analysis (LCA) may minimise bias by including covariates on the causal path as 'class predictors' and by accommodating uncertainty associated with confounder values explicitly via the latent class part of the model. We construct multilevel latent class models to allow for the multilevel structure of the data: patients nested within NHS Trusts.

We use a dataset of patients in a large UK regional population diagnosed with colorectal cancer between 1998 and 2004. Death within three years is the outcome. Following exclusions 24,640 records were available for analysis. The optimum number of latent classes at patient and Trust level is determined with reference to likelihood-based model-fit criteria and classification error.

The three-patient five-Trust class multilevel LCA model was chosen. Patients were apportioned into either a good prognosis, reasonable prognosis or poor prognosis group. The stage profile differed across the patient classes. Socioeconomic background and older age were clearly associated with increased odds of death in all patient classes. Females had significantly decreased odds of death compared with males in the good prognosis class.

The five Trust classes identified outlying Trusts, indicating that the standard multilevel model would not have been sufficient to model these data.

2.3 Invited - How are official statistics used in policy evaluation?

Tuesday 2 September 12 noon - 1.20pm

Evaluation of Automatic Enrolment

Alison Cousley
Department for Work and Pensions

Automatic Enrolment is the largest change to private pensions for a generation. This new law requires all employers to enrol their eligible jobholders into a workplace pension scheme and make a contribution. This will result in millions of people saving for their retirement for the first time.

The government, industry, employers and directly impacted employees all have a massive interest in how the implementation is progressing. This talk outlines the Government's evaluation approach, and discusses the range of official data sources, success measures and simulation models that are used and brought together to produce evaluation updates.

It is early days, but as more data becomes available the initial signs are promising. There is still a way to go before its full roll out in 2018, and so this talk will touch on some of the challenges to come.

2.3 Invited - How are official statistics used in policy evaluation?

Tuesday 2 September 12 noon - 1.20pm

How are official statistics used in policy evaluation?

Jon Simmons
Head of Migration and Border Analysis, Home Office Science Directorate

Migration is one of the most hotly contested areas of government policy and the use of statistics and evidence is often challenged and challenging. However, migration management at the national, regional, and local levels requires decision-making that is informed by empirical research as well as continued monitoring and evaluation. This presentation will explain how official statistics and statistical evidence are used in the evaluation of UK migration policy. It will provide some examples of the use of statistics, and how their interpretation can often be contested. It will provide an overview of a number of different modes of evaluation, as they are applied in practice. It will be of interest to statisticians and researchers interested in public policy and the use of statistics by government.

2.4 Invited - Statistics in Industry

Tuesday 2 September 12 noon - 1.20pm

Data mining public and private energy statistics

Shirley Coleman
Newcastle University, Newcastle upon Tyne, UK

What is the purpose of data mining? This talk will consider the continuous improvement cycle for company data, moral vs exploitative monetisation of data as a company asset, innovative uses for public data and novel analyses of public energy statistics. The talk will include practical examples and thoughts about the challenges of turning data into information. The talk should be of interest to statisticians dealing with people interested in making more of their data.

2.4 Invited - Statistics in Industry

Tuesday 2 September 12 noon - 1.20pm

Making sense of sensor data

Idris Eckley

Lancaster University

The use of sensors for data collection is now ubiquitous in the modern business and industrial setting. Such sensors can unobtrusively record data at potentially very high rates. As such they are a rich source of data, though sensor datasets are not without their statistical challenges! These can arise due to the volume of data, data structure or the environment in which the data is collected. However this has the pleasant benefit of inspiring the development of statistical methodology to tackle these challenges. In this talk I will review some recent work in the area of non stationary time series which has been inspired by such data.

This research has been motivated by collaborations with various industrial partners, drawing on a diverse set of applications (from aerosol design to offshore logistics!).

2.4 Invited - Statistics in Industry

Tuesday 2 September 12 noon - 1.20pm

Statistical tools for enhancing automated manufacturing processes

Keith Harris, Jeremy Oakley, Kostas Triantafyllopoulos, Eleanor Stillman
University of Sheffield, Sheffield, UK

The accurate prediction of cutting forces and temperatures is critically important in designing manufacturing processes to ensure a high quality finished product whilst also keeping costs down and avoiding premature tool breakage. These predictions are typically obtained from computationally expensive finite-element models that simulate the metal cutting process. This limited data must then be used to choose the geometry of the cutting tool and other cutting parameters (like cutting speed, the depth of cut and the feed rate), whilst simultaneously taking into account the uncertainty in the properties of the metal being cut. Our solution to this problem is to build a fast surrogate statistical model (or emulator) of the finite-element simulation model to facilitate optimisation of the metal cutting process. In particular, we use the popular approach of Gaussian process (GP) emulation, which can be thought of as an advanced probabilistic form of regression. In our talk, we will explain our GP approach for enhancing the optimisation of a manufacturing process and illustrate it using a case study of titanium alloy milling (a machining process used in the manufacture of landing gears) from the Advanced Manufacturing Research Centre (AMRC) in Sheffield. Later stages of the project will involve developing a statistical process control (SPC) strategy for the identification of assignable causes of poor process performance and to control overall process variability, and also constructing automatic feedback mechanisms that take corrective action if the metal cutting process is out of control.

2.5 Invited - Bayesian spatio-temporal methodology for predicting air pollution....

Tuesday 2 September 12 noon - 1.20pm

Air Quality Modelling for Health Impacts Studies

Paul Agnew, Lucy Neal, Gerd Folberth, Mohit Dalvi, Fiona O'Connor, Rachel McInnes, Christophe Sarran, Deborah Hemming
Met Office, Exeter, Devon, UK

'HealthAir' is a collaborative project between the Universities of Southampton and Glasgow, and the Met Office, with objectives of quantifying the links between air quality and chronic health impacts under both present-day and possible future conditions. A key requirement is the provision of realistic air quality data. The Met Office has developed a skillful integrated meteorology and air quality forecast model, AQUM. This presentation will describe the key features of AQUM and give an assessment of its performance under a wide range of conditions. The model has been used to generate a dataset of hourly air quality estimates over the whole of the UK for 2007 to 2011 for use in HealthAir. A new capability is being developed as part of HeathAir to nest the high resolution AQUM model within regional and global scale climate models to provide improved consistency of UK air quality estimates under a future climate. The presentation will review both the present-day datasets and the future climate developments.

2.5 Invited - Bayesian spatio-temporal methodology for predicting air pollution....

Tuesday 2 September 12 noon - 1.20pm

Localised conditional-autoregressive models for capturing residual spatio-temporal structure in air pollution and health studies

Alastair Rushworth[1], Duncan Lee[1], Richard Mitchell[1], Sujit Sahu[2]
[1]University of Glasgow, Glasgow, UK, [2]University of Southampton, Southampton, UK

Ecological-level health and air pollution studies aim to estimate the negative impact of ambient air pollution exposure on related health outcomes such as respiratory deaths. It is important to account for those other confounding factors that could contribute to the health outcome and without which, the estimated effect of air pollution may be biased. Typically many of these risk factors are unmeasured and their influence cannot be modelled directly, but they are often correlated in space and as a result, they are often accounted for by introducing a set of spatially smooth random effects often with a smoothness-inducing prior distribution. However, some recent research has shown that in general, the standard choices for smoothing the random effects may not be appropriate, such as those based on conditional autoregressive (CAR) priors. This can be because the CAR prior might imply a global level of smoothness when the unobserved confounding might exhibit localised structure.

We describe the results of fitting a number of spatio-temporal models to long term health and air pollution data in the UK, which have been adapted from the literature designed to tackle these issues. In particular, we discuss the adequacy of some recent attempts to use `Wombling' approaches that try to explore localised structure in areal unit data by generalising the usual CAR priors.

3.1 Invited - Data visualisation – storytelling by numbers

Tuesday 2 September 2.30pm – 4pm

From data to wisdom? Urban data visualisation as a policy tool

Alasdair Rae
University of Sheffield, Sheffield, UK

This talk focuses on the current vogue for urban data visualisation and its potential usefulness as a tool to guide, inform and promote public policy. We now have more data than ever before, more powerful tools with which to analyse it, and the means for low cost mass communication of the results. Data champions would argue that this situation should lead to better understanding of policy problems and potentially better outcomes. Critics argue that it's all just fancy graphics and data nerds having fun. The truth is probably somewhere in between but before we make our minds up it's worth critically examining the issue.

In this presentation, I therefore focus on the much-cited links between data, knowledge, information and wisdom as a way to frame the debate. The objective is to provoke critical thinking on the ways in which 'big data' visualisation is currently used in an urban context and how we might do it better. This is approached from a visual perspective, using the presenter's work and that of other spatial data visualisation experts. There will be lots of maps. Although the focus is mainly on spatial data, the conclusions should resonate more widely with anyone involved in the analysis and visualisation of large datasets.

3.1 Invited - Data visualisation – storytelling by numbers

Tuesday 2 September 2.30pm – 4pm

'How Well Do You Know Your Area?' - Gamifying census results via participative visualisation

Alan Smith
Office for National Statistics, Titchfield, UK

Gamification is emerging as a powerful technique for engaging with users. For example, the most popular content item on the New York Times website in 2013 was not a news article - but an interactive app which asked users a series of questions relating to their own personal dialect. The result - a personalised heat map of where people are likely or unlikely to speak like you - exemplifies current thinking about 'participative' visualisations, content items that focus in equal measure on the Visual, Personal and Social.

In a similar vein, an interactive application developed by the ONS Data Visualisation Centre challenges users to find out 'How Well Do You Know Your Area?', via a personalised quiz of small area census data. Beyond being an interesting experiment for users to compare their own personal intuitions about an area with official data, this project has been an exercise in exploring the use of Web Application Programming Interfaces to create rich, personal content that can extend the outreach and impact of official statistics.

The paper discusses both the background to the project, the development process, publication and reaction, together with conclusions on future applications of this technique to official statistics.

References:

How Y'all, Youse and You Guys Talk retrieved 21/5/2014 from

http://www.nytimes.com/interactive/2013/12/20/sunday-review/dialect-quiz-map.html?_r=1&

Kahneman, D. (2011) *Thinking, Fast and Slow.* Farrar, Straus, and Giroux.

3.1 Invited - Data visualisation – storytelling by numbers

Tuesday 2 September 2.30pm – 4pm

What Will Happen To Data Without Statisticians? The Growth Of Data Visualisation In The Wrong Hands

David Lewis[1]
[1]Audiencenet Ltd., London, UK, [2]Market Research Society, London, UK

It is perhaps no coincidence that as the "140 characters or less" digital native generation has come of age, there has been an exponential growth in appetite for Data Visualisation, particularly in the commercial sector and the media. Data is suddenly sexy. Colourful, vibrant, data-rich, infographics and interactive dashboards are being shared, tweeted and pinned across the planet. Data is being interrogated by key decision makers on mobile phones between meetings and a whole industry is developing around the need to visualise data "stories".

One such Data Visualisation entrepreneur is David Lewis, Founder and CEO of Audiencenet and its subsidiary offering 'DataDesign'. With over 25 years experience of providing full-service consumer research to the music, entertainment and technology industries, David wishes to share his story of how the employment of a small team of graphic designers and animators, alongside his team of data experts, has seen his business take off dramatically in new areas, across both the public and private sector, as well as amongst senior government decision makers.

Central to David's message is his concern as to the resistance which many statisticians display towards the development of methods in data visualisation and the dangers and ethical considerations of data in the wrong hands, should statisticians fail to fully engage with this new world of visualised data.

3.2 Invited - Papers from the Journal of the Royal Statistical Society

Tuesday 2 September 2.30pm – 4pm

Diagnostics and Inference for Respondent-Driven Sampling Data

Krista Gile[1], Lisa Johnston[2,3], Matthew Salganik[4,5]
[1]University of Massachusetts, Amherst, Amherst, MA, USA, [2]Tulane University, New Orleans, LA, USA, [3]University of California, San Francisco, San Francisco, CA, USA, [4]Microsoft Research, New York, NY, USA, [5]Princeton University, Princeton, NJ, USA

Respondent-Driven Sampling is type of link-tracing network sampling widely used to study hard-to-reach populations. Beginning with a convenience sample, each person sampled is given 2-3 uniquely identified coupons to distribute to other members of the target population, making them eligible for enrollment in the study. This is effective at collecting large diverse samples from many populations.

Unfortunately, sampling is affected by many features of the network and sampling process. In this talk, we present advances in sample diagnostics for these features, as well as some advances in inference adjusting for such features.

3.2 Invited - Papers from the Journal of the Royal Statistical Society

Tuesday 2 September 2.30pm – 4pm

A non-parametric entropy-based approach to detect changes in climate extremes

Philippe Naveau
CNRS-LSCE, Gif-sur-Yvette, France

The talk focuses primarily on temperature extremes measured at 24 European stations with at least 90 years of data. Here, the term extremes refers to rare excesses of daily maxima and minima. As mean temperatures in this region have been warming over the last century, it is automatic that this positive shift can be detected also in extremes. After removing this warming trend, we focus on the question of determining whether other changes are still detectable in such extreme events. As we do not want to hypothesize any parametric form of such possible changes, we propose a new non-parametric estimator based on the Kullback–Leibler divergence tailored for extreme events. The properties of our estimator are studied theoretically and tested with a simulation study. Our approach is also applied to seasonal extremes of daily maxima and minima for our 24 selected stations.

3.2 Invited - Papers from the Journal of the Royal Statistical Society

Tuesday 2 September 2.30pm – 4pm

A dynamic probabilistic principal components model for the analysis of longitudinal metabolomics data

Gift Nyamundanda, Isobel Claire Gormley, Lorraine Brennan
University College Dublin, Dublin, Ireland

In a longitudinal metabolomics study, multiple metabolites are measured from several observations at many time points. Interest lies in reducing the dimensionality of such data and in highlighting influential metabolites that change over time. A dynamic probabilistic principal components analysis model is developed to achieve dimension reduction while appropriately modelling the correlation structure due to repeated measurements. This is achieved by assuming an auto-regressive model for some of the model parameters. Linear mixed models are subsequently used to identify influential metabolites that change over time. The model proposed is used to analyse data from a longitudinal metabolomics animal study. Influential and interesting metabolites are highlighted.

3.3 Invited - Data collection challenges in Establishment Surveys

Tuesday 2 September 2.30pm – 4pm

Causes, consequences and measurement of response burden: With implications for the design of business surveys

Jacqui Jones
Office for National Statistics, Newport, UK

Internationally each year survey organisations send out millions of questionnaires to businesses selected to participate in business surveys. Some businesses may only receive one survey questionnaire a year, whilst others receive numerous questionnaires for different surveys and for the different periods that the surveys are conducted e.g. monthly, quarterly and annual. From the business perspective these survey requests are often seen as an irritant that incur them costs and from which they receive no benefits - this can impact on how or if they respond. From the survey organisations perspective survey responses are vital to the quality of the final statistical outputs. A trade off therefore exists between the survey demands being placed on businesses and the need for quality statistics. This presentation will look at the causes, consequences and measurement of response burden, in relation to the design of business surveys.

3.3 Invited - Data collection challenges in Establishment Surveys

Tuesday 2 September 2.30pm – 4pm

On the relationship between response burden and response quality in official establishment surveys

Kari-Anne Lund, Øyvin Kleven, Gustav Haraldsen, Frode Berglund
Statistics Norway, Oslo, Norway

The European Statistics Code of Practise (CoP) for the National and Community Statistical Authorities states that we, as the national statistical office, are obliged to keep the response burden as low as possible. An underlying premise for this principle is the assumption that there is a correlation between response burden and response quality. This paper summarizes findings from 3 separate but inter-connected studies exploring the correlation between response burden and response quality.

Our first study explored the relationship between the response burden and the measurement error by comparing the level of perceived response burden and the rate of control violations. Conclusions from this study suggest no significant connection between perceived response burden and response quality. In our second study, we concluded that there was a low correlation between response burden and response quality using a slightly different approach. Here, the suggested relationship between the response burden and measurement error were explored by comparing both "perceived" and "actual" response burden with the level of manually edited key variables and the total number of reminders. In our third study, we tried to shed light on both previous studies by using new empirical data from Statistics Norway's annual survey on depreciation of operational resources.

3.3 Invited - Data collection challenges in Establishment Surveys

Tuesday 2 September 2.30pm – 4pm

Questionnaire Testing and Evaluation in Establishment Surveys

Deirdre Giesen
Statistics Netherlands, Heerlen, The Netherlands

Pre-testing and evaluating questionnaires helps researchers to improve the quality and efficiency of their data collection. Traditionally, general methodology of data collection and specific methods and techniques to pre-test and evaluate questionnaires were mainly focused on surveys of households and individuals. However, in the last ten years we have seen an increasing awareness of the importance of the data collection methodology for establishment surveys. Nowadays, in many National Statistical Institutes questionnaire testing and evaluation includes both household and establishment surveys. Of course in essence responding to a survey is the same for respondents in establishment surveys and household surveys: a person provides requested data and is more or less able and willing to do so. However in establishment surveys, the performance of this task occurs within an organisational setting and may be affected by characteristics of that setting.

This paper will give an overview of the methods commonly used in the testing and evaluation of establishment survey questionnaires. I will discuss in more detail how cognitive interviewing techniques can be used in this context and illustrate this with some examples of projects we have conducted at Statistics Netherlands.

3.3 Invited - Data collection challenges in Establishment Surveys

Tuesday 2 September 2.30pm – 4pm

Increasing motivation for better reporting in establishment surveys

Mojca Bavdaž[1], Johan Erikson[2], Boris Lorenc[3], Ger Snijkers[4]
[1]University of Ljubljana, Ljubljana, Slovenia, [2]Statistics Sweden, Örebro, Sweden, [3]Statistics Sweden, Stockholm, Sweden, [4]Statistics Netherlands, Heerlen, The Netherlands

The main challenges of conducting establishment surveys nowadays concern declining or low response rates, burden complaints and poor accuracy of reported data. They are present even in mandatory official establishment surveys. National Statistical Institutes have initiated many activities to reduce burden but persistent complaints suggest that a driving force, i.e. the motivation to participate and give accurate and timely response, is insufficient or lacking. The paper integrates findings from empirical studies in three countries: two inter-connected experiments in Slovenia with a follow-up evaluation, two experiments in Sweden with a follow-up evaluation and an evaluation study in the Netherlands. All studies were designed to improve response behaviour in establishment surveys by raising respondents' motivation through various kinds of communication efforts, especially alternative texts and images in written communication, telephone contacts and offering the web option. For instance, a revised text used a more inviting and personal tone, a reminder sounded less threatening, usefulness for society or organisation was emphasised in the text etc. The experiments showed very few significant results. The largest impact was observed in a long and complex survey. As the evaluations suggest, the main reason for unexpected results lies in the fact that changes went largely unnoticed due to routinised reporting activities. Future research should thus place more attention to reaching respondents while also distinguishing between newly and previously sampled establishments, and "professional" and novice respondents.

4.1 Contributed - Communication of Statistical Ideas

Tuesday 2 September 4.30pm – 5.30pm

Normal Enough? Lost Lessons of the Masters

Neil Spencer, Lindsey Kevan de Lopez, Margaret Lay
Statistical Services and Consultancy Unit, Hertfordshire Business School, University of Hertfordshire, Hatfield, UK

The issue of whether data are "Normal enough" for parametric hypothesis tests is one with which Statisticians seem to have little problem. They will happily look at a histogram and conclude that the data appear to be from a population "Normal enough" for tests that are known to be robust to "reasonable" departures from Normality.

However, quite understandably, researchers who would not call themselves Statisticians struggle to see how such conclusions are reached as explanations are largely based on arm-waving (sometimes literally in order to illustrate a Normal curve) and the use of nebulous words such as "sufficiently" and "reasonably". Not surprisingly, they seek more concrete guidance and "rules of thumb" from the literature within their own disciplines. This in itself is an issue for the statistical community - should it not be the case that they themselves are providing such advice?

This talk highlights the (often poor) guidance available to researchers from within their own subject areas and contrasts this with the advice offered by the Past Masters of Statistics such as Pearson, Tukey and Box; advice which today appears to have been largely forgotten by the statistical community or at least overtaken by the notion of simply glancing at a histogram.

A moulding of the Masters' advice into practical guidance which achieves a balance between statistical purity and pragmatic exposition is undertaken. It is hoped that this will lead to a reduction in the needless rejection of parametric analyses by researchers in favour of less powerful alternatives.

4.1 Contributed - Communication of Statistical Ideas

Tuesday 2 September 4.30pm – 5.30pm

A Statistical Analysis Assistant – the future or folly?

William Browne[1], Richard Parker[1], Danius Michaelides[2], Chris Charlton[1]
[1]University of Bristol, Bristol, UK, [2]University of Southampton, Southampton, UK

The practice of performing a statistical analysis for an applied researcher or statistician has changed dramatically over the last 50 years. The increasing speed and storage capacity of computers has meant that software is available for most standard statistical tasks that a researcher might require and so to some degree the computer itself is a statistical analysis assistant. Of course in reality the computer is a tool that can be very helpful in performing statistical analysis but dangerous in the wrong hands! Often statistical software (and books) consists of a series of procedures to perform particular operations that might be considered part of a statistical analysis, for example drawing a graph, fitting a regression etc. and the researcher then needs to piece these operations together in a logical fashion with their research question and data to perform an analysis with possible guidance from documentation.

At the Centre for Multilevel Modelling in Bristol we have spent many years developing software packages funded by the ESRC (MLwiN, MLPowSim, RealCom, Stat-JR) to allow researchers access to cutting edge statistics techniques. Our most recent package, Stat-JR, developed with University of Southampton, has an eBook interface that allows the incorporation of statistical software components within an electronic book. We have obtained funding to extend this work in the direction of constructing a statistical analysis assistant which will synthesise the work of experts and guide the researcher through their statistical analysis. In this talk we will discuss initial progress on the project.

4.1 Contributed - Communication of Statistical Ideas

Tuesday 2 September 4.30pm – 5.30pm

Visual Assessment of Cluster Structure via Dendrograms

Nema Dean[1], Rebecca Nugent[2]
[1]*University of Glasgow, Glasgow, UK,* [2]*Carnegie Mellon University, Pittsburgh, PA, USA*

Cluster analysis is a set of methods designed to explore and find unknown group structure in (multivariate) data. There are a wide variety of methods available, the application of which can result in many different proposed cluster structures. Particularly for more heuristic methods (e.g. hierarchical clustering), it can be difficult to make an objective decision on which method/number of clusters gives the "best" answer. One alternative that claims to address this flaw is the model-based clustering methodology - the application of, often Gaussian, mixture models usually with some likelihood based criterion for selection of the best model/number of mixture components. An issue with this approach can be its tendency to overestimate the number of groups when associating each mixture component with a cluster (estimated group). This talk seeks to present novel applications of an old fashioned clustering tool – the dendrogram - to visually assess either combination of mixture components into clusters or to assess the similarity/grouping of clustering solutions from a variety of methods (applied to the same data). The dendrogram's tree diagram presentation is a particularly useful graphic as it can be easily understood by laypeople and is a visually appealing tool for exploratory analysis into group structure. It is also a good summary for data of arbitrary dimension.

4.2 Contributed - Time Series

Tuesday 2 September 4.30pm – 5.30pm

Semi-parametric inference of time series using the Whittle Likelihood

Adam Sykulski[1,2], Sofia Olhede[2], Jonathan Lilly[1], Jeffrey Early[1]
[1]NorthWest Research Associates, Seattle, Washington, USA, [2]University College London, London, UK

Deriving estimators by maximising the Whittle Likelihood is a computationally efficient technique for estimating parameters of time series models, introduced by Whittle in 1954. The classical theory is asymptotic, and does not account for finite sample effects, nor for temporal variation. In this talk we present new flexible semi-parametric modifications to the Whittle Likelihood, by permitting models to be specified semi-parametrically in frequency, as well as in time. Such an approach is particularly useful, for example, if high-frequency data is heavily biased due to observational noise or pre-processing from interpolation, as would often be the case with financial, environmental and medical data. To account for this, we demonstrate how certain frequencies can be omitted from the Whittle likelihood to remove significant biases in parameter estimation. We also discuss new computationally efficient methods of dealing with finite sample modifications to remove estimation bias.

We then demonstrate extensions to bivariate data. Here the semi-parametric Whittle Likelihood can also be used to account for time series that are contaminated in particular multivariate directions. The methods can also be used to effectively capture elliptical or anisotropic structure in bivariate data - particularly if the degree of anisotropy is scale-dependent. Overall such methods are particularly useful with physical observations such as with oceanographic flow data, as we shall demonstrate with examples from real data.

4.2 Contributed - Time Series

Tuesday 2 September 4.30pm – 5.30pm

Dealing with Missing Data in Spectral Analysis

Audrey Kueh
Warwick University, Coventry, UK

There are many reasons why data could be missing. On one hand, data could be corrupted. For example, the Cosmic Microwave Background yields key information about the origins of our universe, yet the band containing the Milky Way is polluted with emissions from the Milky Way itself, thus we have incomplete data. On the other hand, data could come with a cost. For example, in computed tomography, X-ray attenuation allows us to discover the geometry of objects without physically cutting them open, but (even in the industrial case) data is limited by time and cost constraints. Standard Fourier methods have therefore to be adapted to deal with missing data. One of the main ways to do this is to adapt the Fourier series to a form which is localised in space as well as frequency.

I will first expand on the motivation above, and then I will focus on density estimation in the spherical case. Here, Narcowich, Petrushev and Ward built spherical wavelets on the spherical harmonics to form a tight frame for the space $L^2(S^{d-1})$. Inspired by work of Baldi, Kerkyacharian, Marinucci and Picard, I show how thresholding techniques can be used to find a density estimate which adapts to the local regularity of the density function, which I proved to be minimax-optimal within logarithmic factors. I will also describe construction of a confidence interval whose size is adaptive to the local regularity.

4.2 Contributed - Time Series

Tuesday 2 September 4.30pm – 5.30pm

Penalized Likelihood Estimation in High-Dimensional Time Series Models and Its Application

Yoshimasa Uematsu
Institute of Statistical Mathematics, Tachikawa, Tokyo, Japan

This paper presents a general theoretical framework and its application of penalized quasi-maximum likelihood (PQML) estimation for stationary time series models with the number of parameters diverging. The PQML estimator is defined as a maximizer of a penalized quasi-likelihood function. We show the existence, rate of convergence, sparsity and asymptotic normality of this estimator under general high-level assumptions in the first half of the paper. In the latter half, we propose a method of sparse estimation in large dimensional vector autoregressive models by utilizing the results obtained in the first half. Finally, the usability is conformed through a simulation study and empirical analysis on a yield curve prediction.

4.3 Contributed - Government surveys and evaluation

Tuesday 2 September 4.30pm – 5.30pm

Standard error estimation of regional estimates from the Annual Business Survey

Jennifer Davies, Matthew Greenaway, Gary Brown
Office for National Statistics, Newport, UK

The Annual Business Survey (ABS) is the UK's largest business survey, sampling over 70,000 businesses to provide essential information on the production, construction, distribution and services sectors of the UK economy. Estimates are published at UK and regional level - but historically, standard errors were only published for UK estimates. This paper reports on a project to estimate standard errors of ABS regional estimates.

Like the national estimates, the regional estimates are produced using design-based estimation, however there is an additional modelling stage involved, which takes the survey returns from businesses and assigns these to the regions in which they are based. Two approaches were used to estimate standard errors: an analytic variance formula and bootstrapping. The analytic formula (similar to that used for the national standard errors) assumes that the results of the modelling stage are fixed. The bootstrapping method uses as a basis the Rao-Wu rescaling bootstrap and refits the model parameters at each iteration to reflect that the fitted model depends on the survey data.

The standard errors produced using the two methods were compared to assess the impact of the modelling stage. The results varied between industries according to the structure of businesses. The final recommendations of the project took into account the assumptions underlying each method, the differences in results that they produced, and practical considerations around production of standard errors. The final report was published on the ONS website in February 2014, and the recommended methodology is expected to be implemented in July 2014.

4.3 Contributed - Government surveys and evaluation

Tuesday 2 September 4.30pm – 5.30pm

Evaluation in Government

Anna Athanasopoulou, Phil Bradburn, Anne Jennings, Thomas Williams, Michael Kell
National Audit Office, London, UK

Ex-post evaluations examine the implementation and impacts of policy interventions. Good-quality evaluations can provide evidence on attribution and causality - that is, whether the policy delivered the intended outcomes or impacts, and to what extent those were due to the policy. Evaluations should be a key source of information on the cost-effectiveness of government activities. Government should use them for accountability purposes, as a means to improve existing policies and to design future policies better. This report focuses on impact and cost-effectiveness evaluation relating to government spending, taxation and regulatory interventions, across the 17 main departments. The report aims to add to existing assessments of government evaluation by providing quantitative answers to four questions:

1. What proportion of government activities does existing evaluation evidence cover?
2. What is the quality of this evaluation evidence?
3. How well does this evaluation evidence support strategic resource allocation, policy development and policy implementation?
4. How much does the government spend in producing this evaluation evidence?

We found that the government spends significant resources evaluating the impact and cost-effectiveness of its spending programmes and other activities. However, the coverage of these evaluations is incomplete, and the rationale for what the government evaluates is unclear. Often evaluations are not robust enough to identify the impact of programmes reliably. The government fails to use the learning from these evaluations effectively.

4.3 Contributed - Government surveys and evaluation

Tuesday 2 September 4.30pm – 5.30pm

Exiting Unemployment: An Analysis Using the Australian Longitudinal Labour Force Survey File

Cristian Rotaru
Australian Bureau of Statistics, Canberra, Australia

What affects the probability that an individual who has just entered unemployment finds employment within a given timeframe? Does the probability of exiting unemployment depend on the length of the individual's unemployment spell?

This paper reflects on these questions and analyses the transitions from unemployment of Australians aged 20-65 years, over the 2008-2010 period. The analysis makes use of the new Australian Bureau of Statistics Longitudinal Labour Force Survey file – a dataset that combines labour force data for more than 150,000 households observed on a monthly basis for a period of up to eight months. This paper is the first longitudinal analysis conducted on the file.

Using the job-search theoretical framework, the paper builds a model aimed at analysing the factors that influence transitions from unemployment. A range of methodological techniques are implemented, including the creation of time intervals and the subsequent discrete duration analysis, the adoption of the competing-risks framework to account for the different types of exits from unemployment, as well as the inclusion of random effects to account for unobserved heterogeneity.

Amongst others, the findings indicate the following: single parents with children are the least likely to exit into full-time employment, there is evidence of Global Financial Crisis effects, and that the longer the unemployment spell, the less likely it is for people to exit into employment, and the less likely it is for them to leave the labour force (contrary to the discouraged job seeker effect).

4.4 Contributed – Medical

Tuesday 2 September 4.30pm – 5.30pm

Analysis for single patient studies

Paul Garthwaite
Open University, Buckinghamshire, UK

Statistics rarely tries to draw conclusions from a single individual. However, in neuropsychology an individual with brain injury may display unusual combinations of abilities that give insight into the architecture of the brain - perhaps showing that two tasks that appear similar are, in fact, performed by different parts of the brain as the brain injury has impaired performance on one task but not the other. Clinical neuropsychologists also need to evaluate whether a patient has a deficit on some specified task, asking the question "How extreme are the patient's scores compared with the scores that would be obtained in the general population?"

This talk describes effective statistical methods for comparing a single patient with a control population when scores are recorded on three or more measures. The fundamental questions are whether the patient could be part of the general population and the degree to which the patient is unusual. Approaches to obtaining point and interval estimates are discussed and compared. Also, a means of determining the contribution of individual scores to a Mahalanobis distance is described, which helps explore the features that make a person's profile unusual. This work continues a long-term collaboration with a neuropsychology professor that has produced well-used methods for analysing single patient data when scores are recorded on only one or two measures.

Methods are implemented in user-friendly free software that may be run over the web without downloading it. Data input is minimal - only sample means, standard deviations and correlations are required.

4.4 Contributed – Medical

Tuesday 2 September 4.30pm – 5.30pm

A Method for evaluating accumulating scientific evidence: Application to public health policy and deciding product liability cases.

Joseph Gastwirth
George Washington University, Washington, DC, USA

When case reports suggest that a product creates a health risk, epidemiologists often conduct case control studies to estimate the magnitude of the risk of exposed individuals so policy makers can take appropriate action.

The Bayesian framework enables one to update the evidence as the studies accumulate. Utilizing the first study to determine two priors that are "consistent" with the data: one centered on the estimated odds ratio and one centered on the low end of 90% confidence interval, which favors the manufacturer, one can estimate the posterior probability that the odds ratio is at least a pre-set threshold, e. g 2.0, that will suffice for warning the public or determining legal liability. The method is applied to the studies that linked aspirin use in children to Reye syndrome. It will be seen that the evidence was "strong" enough to support a warning after the third study, as the FDA recommended in 1982. Policy makers in the U.S. and U.K., however, required further studies, which confirmed the association, before warning the public in 1985 (US) or 1986 (UK). Had they acted earlier, over 50 lives would likely have been saved. A model describing the relationship between the time required in order to detect a relative risk (RR) and the magnitude of that risk will be presented; along with implications for public health policy and the monitoring adverse events of newly approved drugs.

4.4 Contributed – Medical

Tuesday 2 September 4.30pm – 5.30pm

Methods for Accounting for Measurement Error in Mediation Models as Applied to the MRC Pre-School Autism Communication Trial

Victoria Harris, Andrew Pickles, PACT Consortium
Institute of Psychiatry, London, UK

Objectives:

The Pre-School Autism Communication Trial focused on a parent targeted intervention with a main outcome of the ADOS-G score, which measures autism symptom severity. The treatment target was the proportion of parent acts that were determined to be synchronous with the child. In order to understand the mechanism by which changes in the targeted parent behaviour influenced child behaviour and subsequently ADOS-G mediation analysis was used. One problem that arises in mediation models is that the indirect effect can be underestimated in the presence measurement error.

Methods:

One approach that has been advocated for accounting for measurement error is instrumental variables, which are variables that influence the mediator but have no direct effect on the outcome. However unless these variables have been specifically incorporated in to the study design they can be difficult to identify. An alternative approach is to include additional sources of information about the reliability of the mediator in to the model, such as the intraclass correlation or values of the mediator at additional time points. We illustrate these methods

Conclusions:

Whilst the instrumental variable approach produced estimates of direct and indirect effects with confidence intervals so large as to render them uninterpretable, making use of repeated measures of the mediator provided more reliable estimates that accounted for measurement error. As predicted the proportion of the total effect that was indirect increased when measurement error was accounted for. These models provide evidence that supports the hypothesised mechanism by which the treatment influenced the main and intermediate outcomes.

4.5 Contributed - Data Science

Tuesday 2 September 4.30pm – 5.30pm

Investigating the Impact of Longitudinal Outliers: The use of Robust Joint Models

Lisa McCrink[1], Adele Marshall[1], Karen Cairns[1], Damian Fogarty[2]
[1]Queen's University Belfast, Belfast, UK, [2]Belfast Health and Social Care Trust, Belfast, UK

The use of joint models to simultaneously analyse longitudinal and survival data has rapidly grown in popularity over the past few decades. Despite this however little research has investigated the impact of the commonly held normality assumptions of both the random effects and random error terms in the presence of longitudinal outliers. This research presents novel methodology in the investigation of this issue, highlighting the need to utilise more robust joint models in the presence of either individuals who do not conform to population trends (b-outliers) or those who have outlying observations that do not follow their own individual-specific trends over time (e-outliers).

A robust joint model is presented in which the gamma-normal hierarchical form of the mixed model is linked with a Cox Proportional Hazards model through shared latent random effects. By altering the distributional assumptions of both the random effects and random error terms from normality to t-distributional assumptions enables the down weighing of both types of outliers thus allowing more precise estimates to be obtained.

A comparison of robust and standard joint models is provided through an illustrative example using ten years of Northern Irish renal data. With an ageing population, the analysis of chronic diseases is becoming more pressing. This research investigates a key issue that impacts the survival of haemodialysis patients, their haemoglobin levels. Due to the presence of longitudinal outliers, the robust joint model is shown to provide a significantly better fit to the data than a standard joint model with normality assumptions.

4.5 Contributed - Data Science

Tuesday 2 September 4.30pm – 5.30pm

Optimal design of experiments on social networks

Steven Gilmour[1], Ben Parker[1], John Schormans[2]
[1]University of Southampton, Southampton, UK, [2]Queen Mary, University of London, London, UK

We investigate how connections between subjects in a social network affect the design of experiments on those subjects. Specifically, where we have unstructured treatments, whose effect propagates according to a linear network effects model, we show that optimal designs are no longer necessarily balanced; we further demonstrate how experiments which do not take a network effect into account can lead to much higher variance than necessary and/or a large bias.

4.5 Contributed - Data Science

Tuesday 2 September 4.30pm – 5.30pm

Restricted b-spline models for longitudinal data analysis

Geoff Jones[1], Wes Johnson[2]
[1]Massey University, Palmerston North, New Zealand, [2]University of California, Irvine, Irvine, USA

The response profiles of individual subjects in a longitudinal study may not be well described by a parametric model. In such situations a nonparametric technique, such as b-splines, is usually adopted for modelling the profiles. If these profiles are believed to transition between stable initial and final levels, there may be advantages in constraining the fitted curves to have initial and final horizontal asymptotes. We show how this can be done by placing simple restrictions on the b-spline coefficients, or equivalently adapting the b-spline basis.

We illustrate the methodology using an analysis of the hormone levels of women as they transition from pre- to post-menopause, focusing in particular on the effects, if any, of age and ethnicity on the individual profiles. We also consider, in a Bayesian framework, the use of the Dirichlet Process Mixture (DPM) for uncovering hidden structure in the covariance.

4.6 Contributed - Challenges for the Statistician

Tuesday 2 September 4.30pm – 5.30pm

Small Island Statistics in a world of Big Data

Paula McLeod[1,2]
[1]St Helena Statistics Office, South Atlantic, Saint Helena, [2]Office for National Statistics, Titchfield, UK

St Helena is a small island heading for big change. One of the most isolated islands in the world St Helena has been used as a stopover for passing ships on the pre-Suez canal route from Europe to Asia and South Africa and as a place of exile (http://en.wikipedia.org/wiki/Saint_Helena). However, the days of physical isolation are numbered. St Helena is facing up to irrevocable change with the impending arrival of air access- construction is well underway for an airport scheduled for operational completion in February 2016.

As we head into this change St Helena has a higher than ever demand for reliable data on the people, the economy and the environment. The statistics need to be bang up-to-date (where possible) and accessible to all (always!). This means many changes in the way we collect data, process information, and then report and disseminate. An equal priority as we strive to improve the quality of statistics is the need to support users in how they user and interpret the information at hand.

We give an overview of how the production and dissemination of statistics in a National Institute with staff of four on an island of population 4,500. The challenges of working in such an environment are many, so too are the rewards. The immediacy of a small island is a keen reminder of just how vital good statistics are.

4.6 Contributed - Challenges for the Statistician

Tuesday 2 September 4.30pm – 5.30pm

The PLEASANT Trial: A Personal Perspective of a Statistician being a Chief Investigator

Steven Julious
University of Sheffield, Sheffield, UK

Introduction

In the UK there is a pronounced increase in the number of visits to the doctor by school-age children with asthma in September. It is thought that this might be caused by the return back to school, when children with asthma are suddenly mixing with other children again and picking up bugs which can affect their asthma.

We have shown that during the school summer holidays there is a drop in the number of prescriptions for asthma medications. It has been suggested that August is a good month to be an asthmatic therefore children with asthma might not take their medication as they should or allow their medication to run low.

Methods

The PLEASANT study- Preventing and Lessening Exacerbations of Asthma in School-age children Associated with a New Term –is a cluster randomised controlled trial investigating the effect of a postal intervention, sent to parents/carers of school age asthmatic children during the summer holidays, with the objective of reducing unscheduled medical contacts in September following the return to school.

The target recruitment was 140 primary care practices (we achieved 142) with the intervention sent out in

Conclusions

The grant to undertake the trial was a culmination of years of research and its successful award presented the challenge of being the chief investigator of a clinical trial. Although an experienced statistician being the chief investigator gave new insights in to the trial process being accountable for: the protocol; governance; external interaction and study reporting which will be discussed and highlighted

4.6 Contributed - Challenges for the Statistician

Tuesday 2 September 4.30pm – 5.30pm

African Institute for Mathematical Sciences: opportunities for statistical capacity buildling.

Jane Hutton
The University of Warwick, Coventry, UK

The African Institutes for Mathematical Sciences (AIMS) are centres for tertiary education and research, which promote mathematics and science in Africa. AIMS trains talented students and teachers in order to build capacity for African education, research, and technology. The first centre, by the sea and mountains at Muizenberg, Cape Town opened in 2003. AIMS Senegal opened September 2011 in MBour, within a seaside nature reserve. Courses are given either in French or in English. AIMS Ghana launched in 2012 and AIMS Cameroon in 2014. AIMS has already trained about 500 people, of whom a third are women, from more than 35 African countries.

The programme has introductory skills courses, and then a series of six three-week blocks in which students choose two out of three review courses. At present, the majority of courses offered are in theoretical physics, traditional applied mathematics and pure mathematics, which reflects the impressive work of the founders Neil Turok and Fritz Hahne. My vision is that a statistics course is always one of the skills courses, and that in each review block there is a statistics course. The main barrier to the provision of substantial statistics capacity building through AIMS is that there have not been enough volunteers offering courses and supervision of essays.

One advantage of teaching in Africa is reduced travel and living costs. Another is that more than half of students trained at AIMS remain in Africa to support Africa's developmental growth. I hope to inspire you to contribute.

PD4 - Best practice for running courses

Tuesday 2 September 4.30pm – 5.30pm

The new ways of delivering training, should you be using them?
Shirley Coleman
Newcastle University

The standard way of delivering statistical training is to bring all the delegates into a single room for a day or more and to deliver a mix of lectures, exercises and case studies. For some delegates this works but trainers should be aware of other ways of delivering training. In this session, an experienced trainer will talk about alternatives such as webinars, skype calls, 1-on-1 coaching plus other considerations that apply to all courses such as pre-reading, choice of location, pre-course questionnaires, post-course feedback, etc. This session will also review how the RSS PDC is looking to adopt these alternatives and the lessons learned so far.

What is best practice for delivering statistical training to non-statisticians and how to adapt your courses for differing learning styles and backgrounds?
Nigel Marriott
Marriott Statistical Consulting

Training for non-statisticians has to be different from that for statisticians since the former may not have the statistical thinking skills that we statisticians take for granted. The speaker has spent over 15 years delivering statistical training to a wide variety of non-statisticians in many different industries and roles and in this talk, he will share his experience of adapting courses to take into account differing learning styles and backgrounds. The speaker does not pretend to have all the answers and so the session will posit a number of scenarios and situations that a statistical trainer might face and will seek feedback from the audience as to how they might tackle these.

5.1 Invited – Young Statisticians Meeting 2014 Prize winners

Wednesday 3 September 9am - 10am

Experts Sharing Knowledge through Wikipedia - Experiences at the Office for National Statistics

Hannah Thomas
Office for National Statistics, Newport, South Wales, UK

In this talk I will outline why I think it is important for statisticians and other experts to share their knowledge online by updating and creating relevant Wikipedia pages.

I will also talk about what researchers and publishers of research can get out of Wikipedia.

I will discuss the process I have carried out at the ONS that has aimed to get more statisticians and economists updating and monitoring Wikipedia articles.

I will then describe how we continue to promote the use of Wikipedia across the ONS and what we intend for the future. I will share my experiences so far, together with any difficulties we have faced in the process.

5.1 Invited – Young Statisticians Meeting 2014 Prize winners

Wednesday 3 September 9am - 10am

Predicting the Results of the Scottish Referendum

Zhou Fang
Biomathematics and Statistics Scotland, Edinburgh, UK

With the coming referendum on Scottish independence, there is a lot of discussion over what the likely outcome will be. But thus far not much rigorous statistics have been done.

Now, we know that opinion polling provides a rich source of data for election forecasting, and statistical techniques have been very effective in the past - see for example the success of Nate Silver in the last US elections.

This highly informal and light hearted talk covers work done in my spare time, applying techniques from past election predictions, and perhaps some new ones. We shall attempt to estimate what the likely outcomes will be. We'll talk about the shortcomings of election prediction. And we shall discuss the possibility that this presentation will be very embarrassing in a few weeks' time.

5.2 Contributed - Statistical Methods & Theory

Wednesday 3 September 9am - 10am

Sufficient Dimension Reduction through Support Vector Machine variants

Andreas Artemiou
Cardiff University, Cardiff, UK

Sufficient Dimension Reduction (SDR) and Support Vector Machine (SVM) have been used separately for handling high dimensional datasets. Li, Artemiou and Li (2011) combined these two areas to achieve linear and nonlinear sufficient dimension reduction in a common framework. In this talk we will show a number of extentions using variants of the classic SVM algorithm.

First, we demonstrate that under specific circumstances, part of the solution the algorithm derives is not unique. Therefore we propose the use of LqSVM (q>1) which ensures the uniqueness of the solution through a modified objective function. Since asymptotic theory of the algorithms depends on the nonunique part of the solution the use of the new algorithm makes the asymptotic results easier to use to develop inferential procedures. We also demonstrate the benefits of the new algorithm in performance through simulation results and real data analysis.

Second, due to the construction of the original algorithm there are naturally unbalanced classes which make the solution of the algorithm biased towards larger class. A cost reweighted variant of SVM is used in the SDR framework and we demonstrate the improved estimation that is achieved through this algorithm.

Finally in an effort to eliminate the effect of outliers on the estimation procedure, we propose an adaptively reweighted algorithm which eliminates the effect of the outliers and a robust estimation is ensured.

Overall in this talk a number of advances for better feature extraction will be presented and a number of directions for future research will be discussed.

5.2 Contributed - Statistical Methods & Theory

Wednesday 3 September 9am - 10am

Fitting regression models with survey data

Alastair Scott, Thomas Lumley
University of Auckland, Auckland, New Zealand

Analysing survey data has become big business, driven by public access to results from large health and social surveys such as the British Household Panel Survey or NHANES. To give some indication of scale, Google Scholar lists almost 50,000 papers reporting regression analyses of data from the British Household Panel Survey. Typically, researchers analysing such data know what analysis they want to do, and would be able to implement it using a standard package, if the data had been collected by simple random sampling. There are problems with the technical details when data is collected via a complex survey. However, the underlying population is not changed by the method of data collection and researchers want to answer the same questions using similar techniques.

Much of this is now possible; all the main packages have survey versions of standard techniques such as linear or logistic regression. However, there are still some widely-used quantities (likelihood-ratio tests, AIC, BIC etc.), missing from these packages. In this talk we develop survey analogues of these quantities. In particular, we show that the theory underlying Rao-Scott tests for log-linear models applies almost unchanged to likelihood ratio tests in general. We also show that AIC can be extended to models fitted to survey data by replacing the usual penalty term with the trace of a "design effect" matrix. Software to implement all the methods has been developed and is available in the R package survey.

5.2 Contributed - Statistical Methods & Theory

Wednesday 3 September 9am - 10am

Misspecified Model Approach for Sensitivity Analysis

Nan Xuan Lin, William Henley, David Llewellyn
Institute of Health Research, University of Exeter Medical School, exeter, UK

Objective: Model uncertainty due to incomplete data or untestable assumptions is a common cause of bias in statistics. Complete control for these biases is unlikely but their impacts can be estimated by means of sensitivity analyses. Sensitivity analysis provides a flexible framework that allows one systematically to integrate uncertainty about different sources of bias into conventional analyses. In this talk, we will introduce a sensitivity analysis framework based on the misspecified model approach.

Method: Many bias problems, including unmeasured confounding, measurement error and missing not at random, can be considered as model misspecification. Based on the misspecified model approach, we formalize a general equation for estimating the true parameter of interest under a range of plausible assumptions about the underling bias problems. We show that the solution of the estimating equation is asymptotically normally distributed with mean equal to the true value and a sandwich-form variance.

Results: We applied the method to two problems: (a) measurement error in meta-analysis and (b) unmeasured covariates in the Cox model. For each problem, we derived the estimating equations and sandwich estimated variances. The results were used in sensitivity analyses for a meta-analysis of rehabilitation programmes for juvenile offending and a randomised controlled trial (RCT) of vitamin supplementation for children with Down syndrome. The first sensitivity analysis showed that the conclusions of the meta-analysis are sensitive to the presence of measurement error. The second suggested that the findings of the RCT are robust to the presence of realistic levels of unmeasured confounding.

5.3 Contributed - Medical/Survival Data

Wednesday 3 September 9am - 10am

A practical divergence measure for survival distributions

Trevor Cox, Gabriela Czanner
University of Liverpool, Liverpool, UK

A new simple divergence measure between two survival distributions is introduced based on the integral of the absolute difference in probabilities that a patient from one group dies at time t and a patient from the other group survives beyond time t and vice versa. In the case of non-crossing hazard functions, the divergence measure reduces to the absolute difference of the expected values of one survival function with respect to the distribution of the other. If the hazard functions cross then a small adaption is necessary. The new divergence is a useful measure as it exists for all pairs of survival distributions, whereas, for instance, the hazard ratio is only appropriate when proportional hazards pertain. The measure can be used in a dynamic way where the divergence between two survival distributions from time zero up to time t is calculated. The divergence can be found for theoretical survival distributions or can be estimated non-parametrically from survival data. For the latter, the new divergence measure is shown to be more accurate and reliable than a similar non-parametric estimate of the Kullback-Leibler divergence. For the case of proportional hazards, the constituent parts of the measure can be used to assess the proportional hazards assumption. The use of the divergence measure is illustrated on the survival of pancreatic cancer patients, looking at fifteen subgroups. The divergences for all pairs of subgroups are found and subjected to multidimensional scaling for insight into relationships between them.

5.3 Contributed - Medical/Survival Data

Wednesday 3 September 9am - 10am

Statistical modelling of biomarkers incorporating non-proportional effects for survival data

Jacqueline Stephen[1], Gordon Murray[1], John Bartlett[2,3], David Cameron[3]
[1]University of Edinburgh, Edinburgh, UK, [2]Ontario Institute for Cancer Research, Toronto, Canada, [3]Edinburgh Cancer Research Centre, Edinburgh, UK

Personalised medicine is replacing the one-drug-fits-all approach with many prognostic models incorporating biomarkers available for risk stratifying patients. Evidence has been emerging that the effects of biomarkers change over time and therefore violate the assumption of proportional hazards when performing Cox regression. Analysis using the Cox model when the assumptions are invalid can result in misleading conclusions.

We report the results of a review of existing approaches for the analysis of non-proportional effects with respect to survival data which identified a number of well-developed approaches but a lack of application of these approaches in practice. The review indicated there is a need for more widespread use of flexible modelling to move away from standard analysis using a Cox model when the assumption of proportional hazards is violated.

We further illustrate the use of two key approaches; the multivariable fractional polynomial time (MFPT) approach by Sauerbrei *et al.* and flexible parametric models proposed by Royston & Parmer, to develop a model for predicting survival of patients with early breast cancer. We illustrate their respective advantages and disadvantages in the development and evaluation of a prediction model.

5.4 Contributed – Roads & Transport

Wednesday 3 September 9am - 10am

Road worker safety trials: balancing statistical requirements with practical application

Caroline Reeves
TRL, Wokingham, UK

Roadworks are essential to maintaining the road network; however, the activity is not without its risks. One of the highest risk activities for road workers is exposure to live traffic whilst deploying and removing signs, particularly when placing signs in the central reservation as road workers have to carry equipment across the carriageway.

In 2010, the Highways Agency developed its Aiming for Zero Road Worker Safety Strategy, which included the aim of eliminating the requirement for road workers to be on foot on a live carriageway. Eliminating the need for signs in the central reservation would help achieve this. An on-road trial to evaluate driver behaviour on approach to roadworks with nearside-only signing was commissioned.

The statistical design and analysis of the trial was challenging: the requirements of a robust trial had to be balanced with the practicalities of road worker operations. Only a limited amount of data could be collected, but results needed to be as generalizable as possible across the network. Factors such as traffic flow and composition had to be considered, and the presence of junctions - and the effect these have on lane choice - needed to be controlled for.

The trial compared the distribution of vehicles by lane on approach to roadworks that had signs on both sides of the carriageway, to works that had nearside-only signing. The trial resulted in substantial change to the way road workers operate, reducing the risk associated with deploying and removing signs.

5.4 Contributed – Roads & Transport

Wednesday 3 September 9am - 10am

How DfT have embraced data visualisation to make statistics have an impact and reach new audiences.

Julie Brown, Paul McEvoy, Samuel Dickinson
Department for Transport, London, UK

Transport Statistics Great Britain (TSGB) is the Department's flagship statistical publication. We recently reviewed TSGB to ensure it was still meeting customer needs. In light of this feedback we re-designed and simplified, to present users with the information they sought at a quick glance, using infographics, simple charts and summary tables. The publication's launch was accompanied by a social media campaign using Youtube, Twitter and e-mails. The Permanent Secretary, quoted the revamped publication as one of DfT's best achievements over the last year.

We recently launched a new "Transport facts" product, using the latest data visualisations to tell a transport story. As well as having impact internally; ministers and policy colleagues have used the fact sheets at a variety of external meetings, generating significant interest. This was clearly demonstrated following a recent phone call to our Statistical Head of Profession. Our ministers had taken the "Transports facts" product along to a meeting with the Home office ministers, resulting in a clear request from the Home office minister to their statisticians - "I want one of these"!

Both documents won Awards (Star and Communications) at the DfT 2014 Analyst Awards.

The National Travel Survey also published a infographic page alongside the main Statistical Release and tables. The infographic was tweeted by external users and generated a Twitter debate about how far people travel on average. The infographic has also been well received by the survey interviewers who show it to respondents when encouraging them to participate in the survey.

5.4 Contributed – Roads & Transport

Wednesday 3 September 9am - 10am

The potential uptake of fully Electric Vehicles - does everybody want one?

Louise Lloyd[1], Neale Kinnear[1], Jenny Stannard[1], Steve Skippon[2]
[1]TRL, Berkshire, UK, [2]Shell Technology Centre, Chester, UK

Electric vehicles are a key part of reducing transport emissions, and their take up is of interest to many people. Two large studies have investigated the interest and likely take up of fully electric vehicles across the population.

In a randomised controlled trial, a sample of people was given a full battery electric vehicle (BEV) for a number of days, and a comparable group was given experience of an equivalent diesel car.

A statistical exploration of the participant responses from before and after their experience has provided insights into likely intentions to purchase a BEV in the near future and the reasons and attitudes leading to that intention.

A relatively small proportion of all participants claimed that they would be fairly or very likely to choose a BEV as their main car in the next five years, and this proportion went down after experiencing use of the vehicle. They were, however more likely to consider an electric car as a second car than as the main car in the household after the experience. General attitudes towards electric vehicles were more positive after the trial, but the practical issues with owning a BEV appear to explain reluctance to consider one.

5.5 Contributed – Statistics in Sport

Wednesday 3 September 9am - 10am

Time-varying ratings in association football: the all-time greatest team is...

Ian McHale, Rose Baker
University of Manchester, UK

The paper presents a new methodology to estimate time-varying team strengths of football teams. Our dynamic model allows for deterministic, rather than stochastic, evolution of team strengths. We use the model to identify the best team in England since the English Football Association was formed and match results were recorded in 1888. Our results suggest Chelsea in 2007 were stronger than any other team has been but that Manchester United have experienced the period of most dominance.

5.5 Contributed – Statistics in Sport

Wednesday 3 September 9am - 10am

An Investigation into Factors Affecting Who Wins Matches in Major Tennis and Volleyball Tournaments Using Official and "Page Ranking" Algorithms and Logistic Regression

Edyta Dziedzic, Gordon Hunter
Kingston University, Kingston upon Thames, Surrey, UK

Successfully predicting the winner of sporting events is important to many people: players, coaches, fans, gamblers, bookmakers, and legal authorities investigating betting irregularities. Here, we develop and test models to predict the outcomes of tennis singles matches and international volleyball matches, based on information known in advance. We investigate using both "official rankings" and "page rankings" (based on methods for ranking web pages by relevance to keyword searches), in combination with various other "profile features", of each player or team.

Using these features, and data from previous matches involving the relevant players/teams, we construct logistic regression models of the odds, and hence probability, of one specified player/team beating another in a particular tournament. The models trained on data from the last 20 (tennis) or 10 (volleyball) major tournaments, until the end of 2013, are applied to predicting match outcomes in the 2014 Australian Open Tennis, 2014 Volleyball World League and 2014 Volleyball Grand Prix tournaments. The success rate of each model is compared with a simple strategy of selecting the bookmaker's favourite.

In all cases, the "official" or "page" ranking of the players/teams are dominant factors in the models predicting the match outcome. Other notable factors include the home continent of the player or the team. For tennis, the model which proved the better predictor used page ranking for men, but official ranking for women. Conversely, for volleyball, the better model for predicting results of men's matches used official rankings, but that for women was based on the page ranking.

5.5 Contributed – Statistics in Sport

Wednesday 3 September 9am - 10am

A profitable adjustive rating system for NBA teams

Jonathan Sargent
RMIT University, Melbourne, Australia

This paper investigates an adjustive rating system for National Basketball Association (NBA) teams. Adjustive ratings in sport are derived from evaluations of the performance of a team or individual player—most often with prior performances in mind—where ratings increase, decrease or remain constant depending on above, below and met expectations respectively. With knowledge of NBA team and opponent pre-match ratings, and any home-court advantage, reasonable assumptions can be made regarding the outcome of a match between the two teams. This paper will demonstrate how fitting an Elo-influenced logistic curve to match data is profitable for betting on head-to-head markets in the NBA. Moreover, return on investment can be increased by optimising home-court advantage and margin of victory parameters with respect to the win likelihood in an impending match. Finally, a polynomial curve is fitted to the match prediction data to estimate a margin of victory, aiding in the determination of a team “covering the line”. It is anticipated that this methodology will translate simply to other invasion sports such as soccer and hockey.

6.2 Invited – What does the ubiquity of data mean for Society?

Wednesday 3 September 10.10am – 11.40am

Data as Culture: How will we live in a data-driven society?

Ulrich Atz
Head of Statistics at the Open Data Institute

We are experiencing a data revolution. Projections alone for the volume of data speak of 40ZB, or over 5TB per person, by 2020. The UK plays a leading role in the open data movement, in the discussion of how we deal with a data-driven society and how data is culture.

For example, how can open data help shape the way we analyse electoral behaviour and increase democratic engagement? New markets such as digital peer-to-peer lending platforms have a different relationship with data, where competition moves away from data access to whoever can build the most effective algorithms. Policy consultations such as the closure of 10 of London's fire stations, traditionally behind closed doors, may now happen in the open.

The open data era in health and social care introduces new questions on the trade-off between usefulness and protecting privacy. How can we find cures for cancer while creating trust with patients that data is shared in responsible ways? What are the challenges when dealing with data? The future of statistics lies on defining our limits no longer in terms of familiar tools and familiar problems, but on embracing the information age.

6.3 Invited - Quantum statistics

Wednesday 3 September 10.10am – 11.40am

Statistics, Causality and Bell's Theorem

Richard Gill
Leiden University, Leiden, The Netherlands

I will present a new yet completely elementary proof of Bell's theorem - the theorem stating that quantum mechanics contradicts the classical physical principles of locality and realism. My proof is built around a finite N, probabilistic or statistical version of the Bell inequality; the usual physicist's version is obtained by taking the large N limit. I will explain the many links to thinking about causality in modern statistics (graphical models, counterfactual variables) and other key statistical concepts (randomization in experimental design, for instance). The new proof also turns out to be extremely useful when engaging with quantum crackpots. Coming up with disproofs of Bell's theorem is the present day version of the perpetuum mobile or squaring the circle. The quantum Randi challenge or QRC (a notion recently invented by Sascha Vongehr) is the challenge of writing an impossible computer program (at least, as long as we do not have quantum computers). This keeps the crackpots busy; moreover, they cannot complain about establishment conspiracies, since if only they would be successful, their computer program would be rapidly disseminated over internet, the whole quantum entanglement science mafia would be totally discredited, and the successful programmer would get the Nobel prize for exhibiting quantum entanglement within classical physical systems. Our probability inequalities give us useful information about how to design a good QRC.

Reference:

Statistics, Causality and Bell's Theorem

Richard D. Gill

http://arxiv.org/abs/1207.5103

To appear in Statistical Science (special issue on causality).

6.3 Invited - Quantum statistics

Wednesday 3 September 10.10am – 11.40am

Nuclear-norm regularization for quantum and classical estimation problems

David Gross
University of Freiburg, Freiburg, Germany

The theory of compressed sensing provides rigorous methods for analyzing the performance of estimators that include a sparsity-enhancing 1-norm regularization term. Since around 2009, a "non-commutative" version of compressed sensing has been developed. Here, the aim is to efficiently recover matrices under a low-rank assumption, most commonly using nuclear-norm regularization. The program was initially motivated by purely classical estimation problems - e.g. the influential "Netflix problem" of predicting user preferences in online shops. However, early on, a fruitful interaction between classical and quantum theory ensued: In one direction, it has been realized that low-rank methods lead to rigorous and very tight performance guarantees for quantum state estimation procedures. In the other direction, mathematical methods originally developed in the context of quantum information theory allowed for a significant generalization and simplification of the rigorous results on low-rank recovery. I will give an introduction into the theory, as well as classical and quantum applications.

6.4 Invited – Statistics and Emergency Care

Wednesday 3 September 10.10am – 11.40am

Emergency system performance using Hospital Episode Statistics

Jon Nicholl, Pat Colman, John Jenkins, Emma Knowles, Alicia O'Cathain
University of Sheffield, Sheffield, UK

Background

We have developed a set of indicators of the performance of emergency care systems using routine data. These system indicators are important because they are patient focused in contrast to service indicators which are management focused. These indicators are population based and instead of answering the question 'which service is best', answer the question 'which population is best served'.

Methods

Using a Delphi exercise a set of 16 possible system indicators was identified by a group of experts. Six of these can easily be calculated from routine Hospital Episode Statistics – mortality and case-fatality rates for indicator conditions, emergency admission and readmission rates for indicator conditions, and unnecessary referrals and transfers to ED.. Using English HES data for the period 2006/7 – 20010/11 we have calculated the indicators for 152 English health administrative areas. We have examined a number of properties for these indicators such as face validity, stability, spatial organisation, and whether they can identify 'out-of-control' performance.

Results

The indicators were easy to calculate, had face validity, showed expected patterns of change, exhibited spatial patterning indicative of system wide effects, and were able to identify known poor performance.

Conclusions

We think that these indicators should be used in public health datasets to monitor the performance of the emergency and urgent care system. The Trauma Audit and Research Network (TARN)

6.4 Invited – Statistics and Emergency Care

Wednesday 3 September 10.10am – 11.40am

What is happening to the numbers of patients attending Emergency Departments?

Suzanne Mason[1,4], Ellen Weber[2], Joanne Coster[1], Jennifer Freeman[3], Thomas Locker[4], Jon Nicholl[1]
[1]University of Sheffield, Sheffield, UK, [2]University of San Francisco, California, USA, [3]University of Leeds, Leeds, UK, [4]Sheffield Teaching Hospitals Foundation NHS Trust, Sheffield, UK

Background

Attendances at English Emergency Departments (ED) have been rising on average by 1.5-2% a year since 1966. This, coupled with increasingly complex population health, the ability of healthcare to provide more sophisticated approaches to the management of illness and changes in the way the public access healthcare has created challenges for EDs. ED crowding is an international problem and has been shown by many studies to lead to poorer health outcomes.

Methods

Targets and performance measures are increasingly being used to assure quality and value for money. In 2005, 98% of patients treated in England's EDs had to be discharged or admitted within four hours of arrival. We undertook routine patient level data analysis to evaluate the impact of the implementation of this target on time spent in the ED, investigations and outcomes.

Results

735,588 ED visits were analysed. The proportion of patients treated within four hours improved from 83.9% to 96.3%. Adjusted total length of ED stay increased by 8 minutes; time to physician was unchanged. There was a 'spike' in the proportion of patients leaving the ED during the last 20 minutes before four hours, which increased from 4.7% of patients in 2003 to 8.4% in 2006 affecting the admitted and elderly patients the most.

Discussion

Introducing a time target reduced the proportion of patients staying >4 hours, total time in department increased and time to physician was unchanged. The growing "spike" in activity just before the deadline implies system difficulties that affect the most vulnerable patients.

6.4 Invited – Statistics and Emergency Care

Wednesday 3 September 10.10am – 11.40am

The Trauma Audit and Research Network (TARN)

Fiona Lecky[1,2]
[1]University of Sheffield, Sheffield, UK, [2]Salford Royal Hospital NHS Foundation Trust, Salford, UK

The Trauma Audit and Research Network (TARN) was established in 1989 as the UK Major Trauma Outcome Study - with a membership of 13 hospitals. Its aim has been to improve the quality of trauma care through benchmarking. One of the main tools for achieving this aim has been the risk prediction model (logistic regression) which was initially derived from that used to predict average survival probabilities for US Trauma patients.

This has then allowed observed - expected mortality rates to be derived for UK hospitals which are now publicly available on the TARN website (www.tarn.ac.uk), and multiple research publications.

The presentation will review what impact if any this endeavour has had on patient outcomes.

6.4 Invited – Statistics and Emergency Care

Wednesday 3 September 10.10am – 11.40am

Explaining variation in emergency admissions

Alicia O'Cathain, Emma Knowles, Ravi Maheswaran, Tim Pearson, Joanne Turner, Enid Hirst, Jon Nicholl
University of Sheffield, Sheffield, UK

Background: Some emergency admissions are potentially avoidable if the health services providing emergency and urgent care are available and accessible. A set of 14 conditions, likely to be rich in potentially avoidable emergency admissions, has been identified by expert consensus. We aimed to explain variation in avoidable emergency admissions between different areas of England.

Methods: National ecological study of 152 emergency and urgent care systems in England defined by 2006-2013 primary care trust populations. Hospital Episode Statistics data on emergency admissions were used to calculate an age sex directly Standardised Avoidable Admission Rate (SAAR) for each system in 2008-11. Explanatory factors were tested in a regression analysis, including population, geography, health, and health service availability, accessibility and quality. The analysis was repeated for systems defined by the catchment areas of acute trusts.

Results: There were 3,273,395 relevant admissions in 2008-11, accounting for 22% of all emergency admissions. The mean SAAR was 2258 per year per 100,000 population, with a 3.4 fold variation between systems (1268 to 4359). Factors beyond the control of health services explained the majority of variation: employment deprivation alone explained 72%, with urban/rural status explaining further variation (R^2=75%). Factors related to emergency departments, hospitals and emergency ambulance services explained further variation (R^2=87%). Factors related to the availability, acceptability and quality of care in general practice did not explain further variation.

Implications: Interventions to reduce avoidable admissions should be targeted at deprived communities. The use and operation of emergency departments and ambulance services affects emergency admissions

6.5 Invited – Statistics in Sport

Wednesday 3 September 10.10am – 11.40am

A Markov model for football possession and its outcomes

Javier López Peña[1,2]
[1]University College London, London, UK, [2]Kickdex Ltd, London, UK

We propose a bottom-up approach to the study of possession and its outcomes for association football, based on probabilistic finite state automata with transition probabilities described by a Markov process. We show how even a very simple model yields faithful approximations to the distribution of passing sequences and chances of taking shots for English Premier League teams, which we fit using a whole season of granular game data (380 games). We compare the resulting model with classical top-down distributions traditionally used to describe possessions, showing that the Markov models yield a more accurate asymptotic behaviour.

6.6 Contributed – Meta-analysis

Wednesday 3 September 10.10am – 11.40am

Bayesian meta-analysis without MCMC

Kirsty Rhodes[1], Rebecca Turner[1], Dan Jackson[1], Julian Higgins[2,3]
[1]MRC Biostatistics Unit, Cambridge, UK, [2]University of Bristol, Bristol, UK, [3]University of York, York, UK

BACKGROUND

Many meta-analyses combine results from only a small number of studies, a situation in which between-study variance is imprecisely estimated when standard methods are applied. Bayesian meta-analysis allows incorporation of external evidence on heterogeneity, providing the potential for more robust inference on the effect size of interest.

METHODS

We propose two methods for performing Bayesian meta-analysis, using data augmentation and importance sampling techniques. Both methods are implemented in standard statistical software and provide much less complex alternatives to Markov chain Monte Carlo (MCMC) approaches. In a simulation study, we compare the performance of the proposed methods.

RESULTS

An importance sampling approach produces almost identical results to standard MCMC approaches, and results obtained through data augmentation are very similar. We compare the methods formally and also apply them to real datasets. For example, a meta-analysis combining four studies evaluating the effectiveness of fluoride for lower limb pain is considered. In a conventional random-effects meta-analysis, the between-study variance τ^2 is high at 1.78, but very imprecisely estimated (95% CI 0.39 to 52.2). The estimated summary odds ratio is 4.14 (95% CI 0.92 to 18.4). When incorporating an informative inverse-gamma prior for τ^2 using importance sampling, the heterogeneity estimate reduces to 0.54, with 95% credible interval 0.04 to 5.33. The summary odds ratio changes to 3.46 (95% CI 1.17 to 14.3).

CONCLUSION

The proposed methods facilitate Bayesian meta-analysis in a way that is accessible to the applied researchers who are commonly carrying out meta-analyses.

6.6 Contributed – Meta-analysis

Wednesday 3 September 10.10am – 11.40am

Outcome reporting bias in randomised controlled trial of pharmacological treatment for Rheumatoid Arthritis: an empirical assessment using multivariate meta-analysis.

Giacomo Frosi[1], Richard David Riley[2], Paula Williamson[1], Jamie Kirkham[1]
[1]University of Liverpool, Liverpool, Merseyside, UK, [2]University of Birmingham, Birmingham, West Midlands, UK

Background Outcome reporting bias (ORB) in randomised trials leads to under-reporting of non-significant outcomes, and is thus a threat to the validity of systematic reviews. Multivariate meta-analysis can potentially reduce the impact of ORB by utilising the correlation between reported and unreported outcomes. The objective of this study was to assess ORB in Cochrane systematic reviews of the treatment of patients with rheumatoid arthritis, and then to demonstrate how multivariate meta-analysis may be applied to reduce its impact.

Methods Systematic reviews published by the Cochrane Musculoskeletal Group were identified. Reviews were assessed for ORB in relation to an established core set of eight outcomes for rheumatoid arthritis. A nine-point classification system previously developed was used. For a review of Auranofin versus placebo, the impact of ORB was assessed by comparing estimates from a univariate and multivariate meta-analysis for both fixed and random effects models.

Results All 21 reviews that were eligible to be included in the ORB assessment contained missing data on at least one of the eight core outcomes. For 515 (46%) of the 1118 evaluable outcomes in our study from the 155 assessable trials, a core outcome was either partially reported or not reported. It was clear from the trial publications that nearly a quarter (22%) of these evaluable outcomes were classified as high suspicion of ORB.

Conclusions In one such review, the summary treatment effect estimates and their statistical significance changed importantly when multivariate meta-analysis was used to reduce ORB through additional information from correlated outcomes.

6.6 Contributed – Meta-analysis

Wednesday 3 September 10.10am – 11.40am

Incorporating external information on between-study heterogeneity in network meta-analysis

Rebecca Turner, Kirsty Rhodes, Dan Jackson, Ian White
MRC Biostatistics Unit, Cambridge, UK

Objectives
In a network meta-analysis comparing multiple treatments, between-study heterogeneity variances are often very imprecisely estimated because data are sparse, and so standard errors can be highly unstable. External evidence obtained from modelling a large database of meta-analyses can provide informative prior distributions for heterogeneity, tailored to particular settings. Our objectives are to explore and compare approaches for specifying informative priors for multiple heterogeneity variances in network meta-analysis.

Methods
In the simplest model assuming equal heterogeneity variances, it is straightforward to construct an informative prior for the common variance. Models allowing heterogeneity variances to be unequal are more realistic; however, care must be taken to ensure that implied variance-covariance matrices remain positive semi-definite, as discussed by Lu and Ades (2009). We consider several strategies for specifying informative priors for multiple heterogeneity variances: proportional relationships among the variances; exchangeability across similar treatment comparisons; or separate priors.

Results
Appropriate prior distributions are obtained through modelling empirical data from the Cochrane Database of Systematic Reviews. The models are applied to a network meta-analysis comparing eight treatments for localised prostate cancer. For example, the odds ratio comparing surgery against standard care is estimated as 0.78 (95%CrI 0.35,1.64) when using vague priors. This changes to 0.79 (95%CrI 0.62, 1.01) when specifying informative priors assuming proportional heterogeneity variances, or to 0.78 (95%CrI 0.54, 1.10) when assuming exchangeability of variances across similar treatment comparisons.

Conclusions
It is possible to incorporate relevant prior information on heterogeneity into network meta-analyses, without making unrealistic assumptions. This may improve precision for estimating treatment differences.

6.6 Contributed – Meta-analysis

Wednesday 3 September 10.10am – 11.40am

Random effects meta-analysis model with Box-Cox transformation

Yusuke Yamaguchi[1], Richard Riley[2]
[1]Astellas Pharma Inc., Tokyo, Japan, [2]University of Birmingham, Birmingham, UK

In a random effects meta-analysis model, true treatment effects for each study are routinely assumed to follow a normal distribution. However, the normality is a strong assumption in practice and becomes problematic especially in the presence of skewness on the observed treatment effect estimates. The misspecification of the random effects distribution may lead to a biased estimate of overall mean for the treatment effects as well as an inappropriate quantification for the extent of heterogeneity across studies. We consider two types of random effects models with Box-Cox transformation, where either the observed treatment effect estimates or the true treatment effects are transformed in order to achieve approximate normality of the random effects distribution. Applying the transformation, we can estimate the mean (or the median) and the variance of the random effects distribution appropriately for the skewed treatment effects. These methods are so flexible that the observed data determine the shape of the random effects distribution. Thorough an application to a real example and simulation studies, we demonstrate the impact of skewed treatment effects on the conclusion of the meta-analysis. For the skewed treatment effect, a potential benefit of our methods using the Box-Cox transformation is also shown in terms of estimating the overall mean, the prediction interval and the extent of heterogeneity. We discuss how many studies will be needed to implement our methods.

Plenary 2 – Significance Lecture

Wednesday 3 September 12.10pm – 1.10pm

The Big Data Trap

Tim Harford

BBC Radio 4/Financial Times

'Big Data' is the catch-all term that is big news for business. Tim Harford looks at some of the most celebrated uses of 'big data' and asks whether we're at risk of forgetting some old and important statistical lessons.

7.2 Invited - Bayes meets Bellman

Wednesday 3 September 2.30pm – 4pm

Bayes, bounds and Bellman

Stephen Roberts, Jan-Peter Calliess, Antonis Papachristodoulou, Michael Osborne
University of Oxford, Oxford, UK

Although many advances in machine learning allow for better prediction, estimation and handling of uncertainty, few consider the key importance of control goals such as guarantees on stability and performance, or the fact that good replication of open-loop behaviour is neither necessary nor sufficient for good prediction of performance under feedback. This talk discusses several research strands related to these deficiencies; the development of learning paradigms in which unknown control systems may be inferred directly from observation with no forwards plant model, proving and using robust bounds on the models we develop, even in the presence of uncertainty and probabilistic approaches to optimisation and active sampling.

7.2 Invited - Bayes meets Bellman

Wednesday 3 September 2.30pm – 4pm

Learning Dynamics and Control

Carl Edward Rasmussen
University of Cambridge, Cambridge, UK

In many practical control applications, the dynamics of the system are partially unknown. Machine learning techniques can be used to capture more knowledge about the dynamics from measurements of states. This has typically been used for "parameter estimation", but non-parametric methods, such as Gaussian processes can also be used to find the structure of system. When learning dynamics from small sets of observations, residual uncertainty exists about the system. When optimizing a controller for the system it is crucial to keep careful account of these uncertainties. This can be achieved using probabilistic models. We show a number of examples, where rapid learning of non-linear dynamics is done using Gaussian process models, enabling the construction of controllers based on only very limited interaction with the system. We discuss the implications of these highly flexible and rapidly adapting controllers.

7.2 Invited - Bayes meets Bellman

Wednesday 3 September 2.30pm – 4pm

Bayesian techniques for black-box optimization

Matthew Hoffman
University of Cambridge, Cambridge, UK

Many real-world optimization tasks consist of objective functions that can only be evaluated point-wise, subject to noise, and absent gradient information. A common technique for solving such problems relies on building a surrogate model of the objective in question. However, it is equally important to model the uncertainty in this surrogate. In this work we present Bayesian approaches to this task which rely on probabilistic predictions of the unknown objective function. This predictive model can then be used to direct exploration either by using credible intervals or directly maximizing information gain. We will show that these methods are surprisingly sample-efficient, obtaining near-optimal function values in very few iterations, all without the use of gradients.

7.2 Invited - Bayes meets Bellman

Wednesday 3 September 2.30pm – 4pm

Bayesian Bandits, An Information Perspective

Nathaniel Korda, Thomas Gunter
Oxford University, Oxford, UK

Bandit problems model the simplest of experimental design problems, and ideas from this literature have begun to seep into many other important areas, such as Monte Carlo integration and Reinforcement Learning under unknown dynamics. Thompson sampling, or probability matching methods have proved particularly successful despite a lack of theoretical understanding of this performance. Recent work has suggested an information theoretic perspective can explain the good behaviour of these methods. We contribute to this understanding, through a new, finite time posterior concentration bound, based on ideas from information geometry. Further work will try to use this tool to analyse the performance of many other Bayesian machine learning algorithms.

7.4 Invited – Measuring segregation

Wednesday 3 September 2.30pm – 4pm

Bayesian inference for segregation indices in the presence of spatial autocorrelation

Duncan Lee, Jon Minton, Gwilym Pryce
University of Glasgow, Glasgow, UK

The degree of segregation between two or more sub-populations has been studied since the 1950s, and examples include segregation along racial and religious lines. The Dissimilarity index is commonly used to estimate segregation, using population level data for a set of areal units that comprise a city or country. However, the construction of this index ignores the spatial autocorrelation present in the data, and it is also typically presented without a measure of uncertainty. Therefore we propose a Bayesian hierarchical modelling approach for estimating the Dissimilarity index and quantifying its uncertainty, which utilises a conditional autoregressive model to account for the spatial autocorrelation in the data. Our proposed method could also be used to perform inference for a variety of other segregation indices. The modelling proposed here is motivated by a study of religious segregation in Northern Ireland, and quantifies the extent to which it has changed between 2001 and 2011.

7.4 Invited – Measuring segregation

Wednesday 3 September 2.30pm – 4pm

Measuring residential ethnic segregation through demographic change

Ludi Simpson and Nissa Finney

University of Manchester

Most measures of ethnic segregation provide an overall index of the degree to which two or more populations are not perfectly mixed. However, an ideal of perfect mixing is both misleading and likely to induce a permanent state of disappointment. It ignores the historical processes that bring people to be where they live, and does not illuminate the social and demographic processes that lead to the distribution of new people and the redistribution of existing population.

This paper therefore examines how population processes shape residential sorting, and how this is socio-economically and ethnically distinct. Births and deaths do make a significant difference to the perceived segregation of populations in the UK, but the major impact, and that on which policy should focus, is of international and internal migration.

We will review evidence from the 1990s and 2000s, and bring it up to date as far as is allowed by published 2011 Census data.

7.4 Invited – Measuring segregation

Wednesday 3 September 2.30pm – 4pm

Characterising Multi-Scale Segregation

Christopher Lloyd
University of Liverpool, Liverpool, UK

The measurement of segregation is usually focused on a single spatial scale - for example, neighbourhoods within cities, or regions within the UK. However, segregation is increasingly recognised as a phenomenon which operates and exists at different spatial scales, and its meaning is not the same across scales. Therefore, there is a need to understand how populations are structured at multiple scales and how this structuring has changed over time. Are population sub-groups within cities becoming more or less uneven? Are regions becoming more similar while cities internally are becoming more different? This paper uses a case study focusing on socio-economic and demographic segregation in England and Wales to consider the importance of spatial scale in measuring segregation, and the ways in which different dimensions of segregation change as the scale of analysis is altered. The results suggest that, for example, the dimension of exposure may make sense at some scales, but not at others. In common with previous research, measures of segregation for multiple spatial scales are shown to capture key aspects of segregation which are ignored in conventional measures of segregation, whether based on indices or typologies. The measurement of segregation is a crucial element in understanding the ways in which a society is becoming more or less geographical unequal and multi-scale measures are essential if we are to be able to properly characterise these inequalities and the impact they have both within and between regions. The paper concludes by making recommendations for more nuanced analyses of segregation.

7.5 Invited – Statistics in Sport

Wednesday 3 September 2.30pm – 4pm

Practising statistics in a fast moving environment: examples from the real world of statistics

Rob Mastrodomenico

Global Sports Statistics

To quote the famous line in Top Gun "I have a need...a need for speed" and during my talk I will show how this need has spread to my statistics work. This talk unlike many at this conference will not focus on the specifics of any statistical method but more so on the implementation in the real world. Through a variety of real world sports examples I will demonstrate issues that can occur from having to practice statistics in real time and address solutions to such problems. To do this I will get into the computational tools that allow us achieve optimal solutions. During the talk I will cover issues such as which computing language is best to use, what data storage solutions should is best to use, how can slow code be sped up, as well as going into my preferred statistical computing setup. This talk is aimed at all levels of statistician and no knowledge of any computing languages is required.

7.6 Contributed – Medical Statistics

Wednesday 3 September 2.30pm – 4pm

A Comparison of Multiple Imputation Methods for Handling Missing Data in Repeated Measurements Observational Studies

Rumana Omar, Oya Kalaycioglu, Andrew Copas
University College London, London, UK

Objective: To evaluate the performance of several multiple imputation (MI) methods for the analysis of missing at random data from repeated measurements observational studies. A comparison is also made with all available case (AC) analysis.

Methods: Standard and random effects imputation by chained equations (ICE), multivariate normal imputation (MVNI), Bayesian MI and AC analysis were compared regarding bias and efficiency of regression coefficient estimates using simulation studies based on real data. Flexibility in handling distributions and correlations of incomplete variables were also compared.

Results: MI maybe preferable to AC analysis when explanatory variables are missing to obtain greater precision. MVNI produced the least bias in most situations, is theoretically well justified and allows flexible correlation for the repeated measurements in the imputation model. It can be recommended for imputation of continuous variables. Bayesian MI is efficient and maybe preferable in presence of missing categorical variables with extreme prevalences. It offers flexibility on distributions for missing continuous variables, although careful consideration is necessary as an inappropriate choice may lead to bias. ICE approaches were sensitive to the choice of correlation by the imputation model. ICE moving time window approach maybe used for imputing normally distributed continuous variables with autoregressive correlation between their repeated measurements. Standard ICE maybe used for the same provided data are available to fit the imputation model. The missingness pattern and degree, distribution of continuous and prevalence for categorical incomplete variables and correlation between the repeated measurements of incomplete variables should be examined to choose an optimal method.

7.6 Contributed – Medical Statistics

Wednesday 3 September 2.30pm – 4pm

Meta-analysis of time-to-event data: what can we do about missing Individual Participant Data?

Sarah Nolan, Anthony Marson, Catrin Tudur Smith
University of Liverpool, Liverpool, UK

Background: Inadequate reporting of time-to-event outcomes and statistical analyses in individual randomised controlled trials is well documented, so meta-analyses of time-to-event outcomes frequently require the re-analysis of individual participant data (IPD). This approach has been undertaken in seven Cochrane reviews in epilepsy, however up to 50% of IPD for eligible trials is not available for re-analysis in these reviews, potentially introducing bias. Methodology has been developed for meta-analysing aggregate data, including indirect estimation methods. Such methods could be of use when an IPD approach is not practical or if a proportion of IPD is unavailable for inclusion in meta-analysis. The feasibility of indirect estimation methods also depend on the extent and quality of statistical reporting in individual trials.

Methods: Our main objective was to determine if we can make use of any published summary statistics via indirect estimation or otherwise and therefore avoid biases related to unavailability of IPD. We assessed the consistency and quality of reporting of outcomes and statistical analyses of time-to-event data in the context of published epilepsy monotherapy designed studies according to guidelines for the reporting of outcomes in monotherapy studies and according to statistical recommendations for the reporting of survival analyses.

Results: 108 eligible studies published from 1978-2012 were identified. 54 studies reported at least one time-to-event outcome. Quality and consistency of reporting of published outcomes and analyses were variable across studies and it is unlikely indirect estimation methods would be able to be applied to most studies. Detailed results and conclusions will be presented.

7.6 Contributed – Medical Statistics

Wednesday 3 September 2.30pm – 4pm

Real time monitoring of progression towards renal failure in primary care patients

Peter Diggle[1,3], Ines Sousa[2], Ozgur Asar[1]
[1]Lancaster University, Lancaster, UK, [2]University of Minho, Minho, Portugal, [3]University of Liverpool, Liverpool, UK

Renal disease can be asymptomatic for many years, but early detection and treatment can slow the rate of progression towards renal failure. Analysis of routinely collected biomarkers of kidney function can assist early detection. Current UK guidelines use the estimated glomerular filtration rate (eGFR) as an overall measure of kidney function and recommend that a patient who is losing kidney function at a rate of at least 5% per year, as measured by their eGFR, should be referred to a specialist treatment centre. In this study, we consider use of dynamic linear modelling to obtain the predictive distribution of the underlying rate of change in kidney function. Our model assumes that kidney function within any one patient evolves according to a continuous-time, non-stationary stochastic process, accommodates between-patient heterogeneity by a combination of baseline covariates and a random patient-specific intercept, and treats eGFR as a noisy measurement of a patient's underlying kidney function. Our overall aim is to incorporate model-based predictions into a real-time surveillance system that can alert general practitioners to the possible need for the referral of their patient to a specialist treatment centre.

7.6 Contributed – Medical Statistics

Wednesday 3 September 2.30pm – 4pm

Treating patients with Rheumatoid arthritis: a Mixed Treatment Comparison analysis of anti-TNF drug effects.

Ingunn Fride Tvete[1], Bent Natvig[2], Jørund Gåsemyr[2], Nils Meland[3], Marianne Roine[3], Tor Skomedal[4], Marianne Klemp[5]
[1]The Norwegian Computing Center, Oslo, Norway, [2]Department of Mathematics, University of Oslo, Oslo, Norway, [3]Smerud Medical Research International AS, Oslo, Norway, [4]Department of Pharmacology, University of Oslo, Oslo, Norway, [5]The Norwegian Knowledge Centre for the Health Services, Oslo, Norway

Objective:

Treatment of Rheumatoid arthritis includes drugs to help control this chronic disease and limit joint damage. The relatively inexpensive disease modifying anti-rheumatic drugs (DMARDs) are common, while the biological anti-TNF (Tumor necrosis factor) drugs are also used. These latter expensive drugs have been tested in clinical trials either alone or in combination with DMARDs, against another anti-TNF drug, just DMARDs, placebo or placebo and DMARDs combined.

Methods or models, and

With 48 eligible clinical trials containing various comparisons of 9 different anti-TNF drugs, DMARDs and placebo combinations we have constructed a complex structural network of comparisons. We have developed a Bayesian MTC model enabling a comparison and ranking of the anti-TNF drugs with respect to their ACR50 (American College of Rheumatology) effect. Our model takes into consideration relevant factors such as duration of RA disease prior to study start and drug doses given during the trials.

Conclusions

We find clear differences between the various anti-TNF (Tumor necrosis factor) drugs with respect to efficiency. The ranking of them can change when taken with additional treatment such as the DMARDs.

8.1 Invited – Research Students Conference 2014 Prize winners

Wednesday 3 September 4.20pm – 5.20pm

Central limit theorems and Markov chain Monte Carlo estimators

Samuel Livingstone[1], Mark Girolami[2]
[1]*University College London, London, UK,* [2]*University of Warwick, Coventry, UK*

Markov chain Monte Carlo (MCMC) is a well-used tool in many areas of Statistics. The objective is typically to construct an estimator for some expectation of interest, by averaging the values in a Markov chain with a specified limiting distribution. An ergodic theorem ensures that this average will converge to the desired quantity with probability one. Without further guarantees, however, we have little assurance on the rate of convergence. To establish central limit theorems (CLTs) for these estimators we must explore the ergodic properties of the Markov chain itself. In this talk we review the basic requirements for a CLT to exist, and then highlight ongoing work to establish these for some modern MCMC methods based on Langevin diffusions and Hamiltonian flow.

8.2 Contributed – Bayesian Methods

Wednesday 3 September 4.20pm – 5.20pm

Convergence rates for Approximate Bayesian Computation

Mark Webster, Jochen Voss, Stuart Barber
University of Leeds, Leeds, UK

Approximate Bayesian Computation (ABC) is a popular computational method for approximate Bayesian inference when the likelihood is intractable. It only requires that we can generate simulated data relatively easily from our chosen model. Each ABC iteration generates data given a candidate set of parameters drawn from a prior distribution. Parameters which generate simulated data "close" to the observed data are regarded as plausible, and a sample of such simulated parameters is used to estimate the posterior parameter distribution given the observed data.

This begs the question of what is meant by "close". Suitable choice of a distance measure and a tolerance parameter is vital for the ABC method to work.

We consider the second problem and show that under very weak conditions ABC estimates converge to the correct values. We also quantify the rate of convergence and use our results to suggest strategies for choosing the tolerance parameter - effectively balancing the computational effort against the desired accuracy of the resulting estimates.

8.2 Contributed – Bayesian Methods

Wednesday 3 September 4.20pm – 5.20pm

Sample size and classification error for Bayesian change-point models with unlabelled sub-groups and incomplete follow-up

Simon White[1], Graciela Muniz-Terrera[2], Fiona Matthews[1]
[1]MRC Biostatistics Unit, Cambridge, UK, [2]MRC Unit for Lifelong Health and Ageing, London, UK

Many medical (and ecological) processes involve the change of shape whereby one trajectory changes into another trajectory at a specific time point. There has been little investigation as to the study design needed to investigate these models.

We consider the class of fixed effect change-point models with an underlying shape comprising two joined linear segments, also known as broken-stick models. We extend the model to include two sub-groups that exhibit a different shift at the change-point, a change and no change class, and a missingness model leading to individuals with incomplete follow-up.

Through a simulation study we consider the relationship of sample size to the estimates of the underlying shape, the existence of a change-point, and the classification-error of sub-group labels.

In summary the estimation of a fixed change-point was acceptable with relatively small numbers of individuals (150) post the change-point, with initial sample sizes of two to five hundred.

8.2 Contributed – Bayesian Methods

Wednesday 3 September 4.20pm – 5.20pm

Bayesian Networks for sex-related homicides: structure learning and prediction

Stephan Stahlschmidt[1], Helmut Tausendteufel[2], Wolfgang Härdle[1]
[1]Humboldt-Universität zu Berlin, Berlin, Germany, [2]Berlin School of Economics and Law, Berlin, Germany

Sex-related homicides tend to arouse wide media coverage and thus raise the urgency to find the responsible offender. However, due to the low frequency of such crimes, domain knowledge lacks completeness. We have therefore accumulated a large data-set and apply several structural learning algorithms to the data in order to combine their results into a single general graphic model.

The graphical model broadly presents a distinction between an offender and a situation-driven crime. A situation-driven crime may be characterised by, amongst others, an offender lacking preparation and typically attacking a known victim in familiar surroundings. On the other hand, offender-driven crimes may be identified by the high level of forensic awareness demonstrated by the offender and the sophisticated measures applied to control the victim.

The prediction performance of the graphical model is evaluated via a model averaging approach on the outcome variable offender's age. The combined graph undercuts the error rate of the single algorithms and an appropriate threshold results in an error rate of less than 10%, which describes a promising level for an actual implementation by the police.

8.3 Contributed – Medical/Observational Data

Wednesday 3 September 4.20pm – 5.20pm

A comparison of multilevel multiple imputation techniques for Big Data in epidemiological research

Paul Baxter, Rakesh Loi, Mark Gilthorpe, W Robert Long
University of Leeds, Division of Epidemiology & Biostatistics, LIGHT, School of Medicine, Leeds, UK

Objectives

The rise of Big Data in epidemiological research routinely requires analysis of hundreds of thousands of patients and hundreds of variables. Missing data is generally unavoidable and leads to biased and inefficient inference if handled inappropriately. Multiple Imputation (MI) has emerged as a reference standard for dealing with missing data. Chained equation methods have become straightforward in software such as the MICE package for R. However, Big Data in epidemiology is frequently hierarchical in context: longitudinal measurements within patients, patients within hospitals and hospitals within trusts. Congeniality between analysis and imputation models is an established concept in missing data; as such several multilevel MI methods have appeared in the literature but have not yet been compared.

Methods

Multilevel MI is available in the MICE package for R. The approach is a Full Conditional Specification (FCS) framework for MI, limited initially to normally distributed multilevel data but readily extended to binary data. The REALCOM IMPUTE software is based on generalised latent multivariate multilevel models (an extension of Joint Modelling (JM) for MI). JM has better theoretical grounding than the FCS approach, though FCS has proved empirically effective in many contexts.

Results

We compare these two approaches to multilevel MI in terms of bias, variance and computation time. We artificially induce Missing At Random data into a previously complete Bangladesh fertility survey (predicting contraceptive use by urban residence, age and number of children, with a random intercept by district) of Huq and Cleland.

8.3 Contributed – Medical/Observational Data

Wednesday 3 September 4.20pm – 5.20pm

Modelling progression in Parkinson's disease.

Michael Schulzer, A Jon Stoessl, Lisa Kuramoto, Raul de la Fuente-Fernandez, Vesna Sossi, Ramachandiran Nandhagopal, Jacquelyn Cragg, Edwin Mak
UBC Pacific Parkinson's Research Centre, Vancouver, B.C., Canada

At the Pacific Parkinson's Research Centre in Vancouver we collected longitudinal Positron Emission Tomography (PET) measurements from 4 subregions on each side of the brain, using 3 different radiotracers, on 78 Parkinson's disease (PD) patients. Non-linear, multivariate, longitudinal random effects modelling was applied to analyze and interpret the data.

The results showed a non-linear decline in PET measurements, which was successfully modelled by an exponential function depending on two patient-related covariates: duration since symptom onset and age at symptom onset. Exploration of the model showed a significantly greater degree of damage in the posterior than in the anterior putamen throughout the disease, and a progressive decrease in the common initial difference between the less affected and the more affected sides with increasing duration. Younger patients had significantly poorer measurements than older patients at the time of symptom onset, suggesting that they had more effective compensatory mechanisms. Cautious backward extrapolation to the intersection with our normal control data showed that disease onset had occurred some 8 to 17 years prior to symptom onset.

Our model provides important new biological insights into the pathogenesis of PD, as well as into its pre-clinical aspects.

Clinical references:

Nandhagopal R et al (2009): Brain 132:2970-2979.

de la Fuente-Fernandez R et al (2011): Annals of Neurology 69:803-810.

Nandhagopal R et al (2011): Brain 134:3290-3298.

Overview of findings:

Kuramoto L et al (2013): PLOS One, October 18.

8.3 Contributed – Medical/Observational Data

Wednesday 3 September 4.20pm – 5.20pm

Introducing RAMMIE - a new hierarchical multi-level modelling approach to syndromic surveillance for public health.

Roger Morbey, Gillian Smith, Andre Charlett, Alex Elliot, Nick Andrews, Neville Verlander
Public Health England, England, UK

Public Health England's real time syndromic surveillance team (ReSST) monitors daily data feeds from a range of sources; including GP consultations, emergency department attendances and calls to the 111 telephone helpline. Surveillance involves tackling challenges of big data; assessing 680,000 health consultations daily to identify activity of public health concern. Detection capabilities require identification of a very wide range of events, from small local one day spikes, to national gradual rises spread over many months. Statistical methods need to be robust enough to cope with a wide range of count data, depending on location and rareness of syndrome; unpredictable changes in data coverage, coding practices or other data quality issues; and sizeable seasonal and day of week effects. The "Rising activity, multi-level modelling and indicator emphasis" (RAMMIE) method was developed as a standard approach, involving creating models for over 11,000 signals and using these to identify local, regional or national areas where activity is unusually high for the time of year or has recently risen significantly. Multi-level hierarchical modelling is being used for the first time in daily syndromic surveillance to enable local surveillance, where counts are low, by ‘borrowing strength' from other areas. Unusual activity is prioritised to ensure a manageable number of ‘alarms' which can be investigated prior to a clinical assessment of the likely impact on public health. This presentation describes the RAMMIE method and its effectiveness in terms of specificity, sensitivity and timeliness.

8.4 Contributed – Older people and victims of theft

Wednesday 3 September 4.20pm – 5.20pm

Who are the victims of theft from the person and robbery?

Rebecca Thompson
Senior Researcher, Institute for Public Safety, Crime and Justice, Northampton, UK

This paper explores the nature of theft from the person and robbery of personal property over a 17-year period (1994-2010/11) in England and Wales. Between 1995 and 2010/11, all crime recorded by the Crime Survey for England and Wales (CSEW) fell 50 per cent, with a 27 and 17 per cent fall in robbery and theft from the person respectively. Using CSEW data, a number of socio-demographic, lifestyle and area characteristics are analysed via negative binomial regression models. This is to investigate whether there are particular characteristics which influence the incidence of theft and robbery and if these have remained stable over time. Age, sex, marital status, frequency of activity outside the home and car ownership have a consistent effect over time. In addition, there are clear and important differences between the characteristics of victims of attempted victimisations as opposed to 'completed'. These findings may help improve both the practical allocation of crime prevention resources and our theoretical understanding of theft and robbery patterns since 1994.

Keywords: theft from the person, robbery, incidence, negative binomial regression models.

8.4 Contributed – Older people and victims of theft

Wednesday 3 September 4.20pm – 5.20pm

The impact of household and partner characteristics on the labour force participation of older women in England

Jennifer Prattley
University of Manchester, Manchester, UK

The economic wellbeing and physical and mental health of the ageing population in the United Kingdom is associated with continued participation in the labour force. Encouraging later life employment is therefore a key policy issue. Research into older person's employment trajectories is concentrated on men's working patterns, and often takes an individualistic approach that does not account for the domestic context. Understanding of older women's labour force participation has been informed by small scale qualitative studies that do consider the household domain but these findings cannot be generalized to the wider population. This research investigates the factors associated with continued employment of women aged 50 to 59 using data from the nationally representative English Longitudinal Study of Ageing (ELSA). Transition rates out of employment between 1998 and 2011 are modelled using multilevel discrete time event history methods that permit the inclusion of time varying covariates. Women and their partners are positioned within a household structure and asymmetric effects of factors on the transition rate of each couple member are considered. Particular focus is placed on the impact of family financial resources and spousal health. The total income of the partner, private pension wealth and the development of a long term limiting health condition are shown to have a differential impact on the employment chances of each partner. These findings emphasise the importance of conducting research into later life employment trajectories on a household, rather than individual, basis.

8.5 Contributed – Statistics in Sport (and Song)

Wednesday 3 September 4.20pm – 5.20pm

An Analysis of Home Advantage for Use in Predictive Models

George Foulds[1,2], Jonathan Tawn[1], Mike Wright[1], Roger Brooks[1]
[1]Lancaster University, Lancaster, UK, [2]ATASS Sports, Exeter, UK

Home advantage refers to the positive effect experienced by a player or team playing at home. Research in this field will initially focus on association football, most notably the four major English Leagues. Data analysis may reveal quantative relationships between home goal advantage and various aspects such as crowd size, travel and referee bias. Simulation models will be explored, allowing the evolution of an archetypal model capable of predicting match outcome whilst taking into account home advantage. Particle filter methods could be used to capture underlying parameters of such a model over time, allowing rapid online updating.

8.5 Contributed – Statistics in Sport (and Song)

Wednesday 3 September 4.20pm – 5.20pm

Accounting for Rink Effects in the National Hockey League's Real Time Scoring System

Michael Schuckers[1], Brian Macdonald[2]
[1]St. Lawrence University, Canton, NY, USA, [2]United States Military Academy, West Point, NY, USA

The foundation of good statistics is quality data. Unfortunately, it has long been known that recording tendencies for the Real Time Scoring System (RTSS) statistics in the National Hockey League tend to vary from rink to rink. Those issues have been covered by numerous authors including Boyle, Zona, Desjardins, McCurdy, etc. Since many advanced statistical methodologies for the NHL, such as CorsiRel, Schuckers and Curro's Total Hockey Ratings and Macdonald's Expected Goals Plus-Minus depend on the RTSS data, it would be useful to account for those rink differences. In this paper we apply a cross-validated elastic net approach to fit a log-linear regression models to estimate the differences in the recording rates of events by rink. We use 5-on-5 even strength non-empty net regular season event**s** from the 2007-08 season through the 2012-13 season adjusted for the total amount of ES time per game. In addition to the rink in which a given game occurred we include the teams involved in the game as well as other explanatory covariates which could affect the rates of each event. The events we considered are the following: shots, misses, blocks, hits, giveaways, and takeaways, The primary result of this model is a set of multipliers for reweighting each event to counteract these rink effects. Additionally, we are able to identify the rinks (Toronto and New Jersey) that have particularly poor recording of events relative to the rest of the league.

8.5 Contributed – Statistics in Sport (and Song)

Wednesday 3 September 4.20pm – 5.20pm

Evidence of bias in the Eurovision song contest: modelling the votes using Bayesian hierarchical models

Gianluca Baio[1], Marta Blangiardo[2]
[1]University College London, London, UK, [2]Imperial College London, London, UK

The Eurovision Song Contest is an annual musical competition held among active members of the European Broadcasting Union since 1956. The event is televised live across Europe. Each participating country presents a song and receive a vote based on a combination of tele-voting and jury. Over the years, this has led to speculations of tactical voting, discriminating against some participants and thus inducing bias in the final results. In this paper we investigate the presence of positive or negative bias (which may roughly indicate favouritisms or discrimination) in the votes based on geographical proximity, migration and cultural characteristics of the participating countries through a Bayesian hierarchical model. Our analysis found no evidence of negative bias, although mild positive bias does seem to emerge systematically, linking voters to performers.

PD7 - CStat revalidation - experience to date and plans for 2015

Wednesday 3 September 4.20pm – 5.20pm

Trevor Lewis

RSS Theme Director for Professional Affairs

This session will give an overview of the CStat revalidation process and provide feedback on experience to-date and plans for 2015.

CStat revalidation was undertaken for the first time earlier this year by a cohort of 190 chartered statisticians who had 1st January 2015 as their revalidation date. The outcome of this process and lessons learnt will be summarised. The process for revalidation in 2015 for those who have a revalidation date of 1st January 2016 will also be outlined. There will be opportunity to ask questions and gain clarity on the requirements for revalidation, the continuing professional development policy of the society and the recording of professional development activities.

Plenary 3

Wednesday 3 September 5.30 – 6.20pm

Removing unwanted variation

Terry Speed

Walter and Eliza Hall Institute, Melbourne

Ordinary least-squares is a venerable tool for the analysis of scientific data originating in the work of A-M. Legendre and C. F. Gauss around 1800. Gauss used the method extensively in astronomy and geodesy. In 1907, motivated by social science, G. U. Yule presented a new notation and derived some identities for linear regression and correlation. Generalized least squares is more recent, originating with A. C. Aitken in 1934, though weighted least squares was widely used long before that. At around the same time (1933) H. Hotelling introduced principal components analysis to psychology. Its modern form is the singular value decomposition, (re-)introduced by Eckhart and Young (1936). Random effects models date back to astronomical work in the mid-19th century, but it was through the work of C. R. Henderson and others in animal science in the 1950s that their connexion with generalized least squares was firmly made.

These are the diverse origins of our story, which concerns the removal of unwanted variation in high-dimensional genomic and other “omic” data using negative controls. We start with a linear model that Gauss would recognize, with ordinary least squares in mind, but we add unobserved terms to deal with unwanted variation. A singular value decomposition, one of Yule’s identities, and negative control measurements (here genes) permit the identification of our model. In a surprising twist, our initial solution turns out to be equivalent to a form of generalized least squares. This is the starting point for much of our recent work. In this talk I will try to explain how a rather eclectic mix of familiar statistical ideas can combine with equally familiar notions from biology (negative and positive controls) to give a useful new set of tools for omic data analysis. Other statisticians have come close to the same endpoint from a different perspectives, including Bayesian, sparse linear and random effects models.

Terry Speed, jointly with Johann Gagnon-Bartsch and Laurent Jacob (Bioinformatics Division, Walter & Eliza Hall Institute of Medical Research, and Department of Statistics, University of California at Berkeley)

9.1 Contributed – Statistics methods & theory/regression

Thursday 4 September 9.20pm – 10.40pm

Measuring Asymmetry and Testing Symmetry

Christopher Partlett[1], Prakash Patil[2], Richard Riley[1]
[1]University of Birmingham, Birmingham, UK, [2]Mississippi State University, Mississippi, USA

We show that some of the most commonly used tests of symmetry do not have power which is reflective of the size of asymmetry. That is, while the tests have good rejection levels for non-symmetric distributions, their power does not increase as asymmetry increases. This is because the primary rationale for the test statistics that are proposed in the literature to test for symmetry is to detect the departure from symmetry, rather than the quantification of the asymmetry. As a result, tests of symmetry based upon these statistics do not necessarily generate power that is representative of the departure from the null hypothesis of symmetry.

Recent research has produced a new measure of asymmetry, which has been shown to effectively quantify the amount of asymmetry in a variety of random variables. We derive the asymptotic distribution of an estimate of this measure, which provides the platform for developing a new statistical test for symmetry. Using simulated data we show that this new test has a number of desirable features compared to existing methods. Moreover, using the asymptotic distribution theory we are able to construct approximate confidence intervals for the measure of asymmetry. We also discuss a number of potential applications for the test and measure using real-life medical statistics data.

9.1 Contributed – Statistics methods & theory/regression

Thursday 4 September 9.20pm – 10.40pm

Inference for generalized linear mixed models with sparse structure

Helen Ogden
University of Warwick, Coventry, UK

Generalized linear mixed models are a natural and widely used class of models, but one in which the likelihood often involves an integral of very high dimension. Because of this intractability, it is common to conduct inference by using an approximation to the likelihood in place of the true likelihood. However, existing approximations to the likelihood, such as the Laplace approximation, often fail in models which have sparse structure, in that the data only provide a small amount of information on each random effect. The use of such approximations is routine in practice, but can give misleading inference for the model parameters.

Motivated by the poor performance of existing approximations, a new approximation method is introduced, which exploits the structure of the integrand of the likelihood to reduce the cost of finding a good approximation to the likelihood. The method gives a large improvement over existing approximation methods in many models with sparse structure. The new method is demonstrated for models for tournaments between pairs of players.

9.1 Contributed – Statistics methods & theory/regression

Thursday 4 September 9.20pm – 10.40pm

On the Computation of R-estimators

Yuankun Wang, Kanchan Mukherjee
Lancaster University, Lancaster, UK

A major branch of nonparametric statistics deals with the rank-estimation (R-estimation) of parameters by minimizing certain dispersions or equivalently, solving a system of equations based on the ranks of residual observations. These estimators never have much lower but sometimes infinitely higher efficiencies than the sample mean or the difference of means in the case of two-sample location. Although R-estimators are useful robust estimators, unfortunately their computation is a challenging and long-standing problem.

In this talk, we have proposed an iterative algorithm which can be applied routinely to compute R-estimates based on any score function. This algorithm depends on the form of a mean function that is linear in parameters. The algorithm yields convergent sequence of estimates for many well-known datasets. In fact, we applied R-estimators to identify some outliers which would not have been detected using least squares.

Because of its simplicity, the algorithm can be applied to compute R-estimators from bootstrap samples. In fact, it opens up various other possibilities and avenues to use R-estimators as one of the most competitive robust class of estimators in various fields of statistics.

9.1 Contributed – Statistics methods & theory/regression

Thursday 4 September 9.20pm – 10.40pm

Empirical likelihood confidence intervals for nonparametric nonstationary regression model

Ryota Yabe
Hitotsubashi University, Kunitachi/Tokyo, Japan

In this talk, we would like to introduce the empirical likelihood (EL) based point-wise confidence interval construction for a nonparametric nonstationary regression model.

In nonparametric time series literature, much attention has been paid to the case where the observed process is assumed to be stationary. However, the stationary assumption seems very restrictive to analyze economic and financial data that are often considered to have nonstationary properties such as a random walk.

To deal with nonstationary data in a nonparametric model, the nonparametric nonstationary model recently receives so much attention. For example, Wang and Phillips (2009, 2011) have developed the asymptotic theory of the local constant and linear estimators and constructed the point-wise confidence interval of the regression function by using the asymptotic distributions of the estimators. The confidence interval based on the asymptotic distribution of the estimators is always symmetry around the estimates. This limitation is crucial in some empirical studies such as a money demand function.

Since it is known that the EL based confidence interval is data-dependent and flexible, this procedure can avoid the symmetric restriction. We show that the Wilks theorem holds even if the covariate process is nonstationary. The asymptotic distribution is same as that of the case where the covariate process is stationary. For illustration of our procedure, we introduce the estimation result of the money demand function by using Japanese data.

9.2 Contributed – Loans, Prices, Inflation and the Income Gap

Thursday 4 September 9.20pm – 10.40pm

Forecasting student loan repayments

Helen Woodward
Department of Business, Innovation and Skills, London, UK

The Department of Business Innovation and Skills is responsible for Higher Education and the student loan reforms of 2012. Fee loans are now up to £9,000 and maintenance loans contribute up to £7,675 per student. The loans are income contingent and are only paid back above a threshold of £21k. A new model was needed to forecast the costs to the pubic purse of making student loans. This was outlined in the December 2013 NAO report on Student Loan repayments.

As the developer of the new model I would like to give a short talk on the modelling process and the model my team has delivered. However there are sensitivities around this and I would need to get permission from my Department. If RSS is interested in this sort of talk I would be happy to gain permission and submit a fuller abstract.

9.2 Contributed – Loans, Prices, Inflation and the Income Gap

Thursday 4 September 9.20pm – 10.40pm

Calculating retrospective consumer price indices using the Fisher and Tornqvist formulae

Ria Sanderson, Gareth Clews, Anselma Dobson-McKittrick, Jeff Ralph
Office for National Statistics, Newport, UK

Consumer price indices are typically calculated using Laspeyres-type index number formulae. These formulae generally rely on the availability of household expenditure weights that are derived in the base period or earlier. Index number formulae that rely on current period expenditure weights are not practical choices in the compilation of a timely consumer price index due to the practical difficulties in collecting household expenditure data for the current period in time. It is, however, possible to make use of index number formulae that require current weights, including those of Fisher and Tornqvist on a retrospective basis; this is an approach that is carried out by a number of other National Statistical Institutes. These index formulae are often considered to be desirable choices both due to their mathematical properties and their place within the economic approach to index number theory. In this presentation, we will present the results of calculating retrospective Fisher and Tornqvist indices using data from the UK Consumer Prices Index (CPI) over the period 2007 to 2009. We will highlight the practical difficulties of calculating these indices, the impact of the necessary assumptions on the resulting indices and the key limitations of the analysis. We will also consider whether it is possible to approximate these indices using formulae than can be calculated in a timely fashion.

9.2 Contributed – Loans, Prices, Inflation and the Income Gap

Thursday 4 September 9.20pm – 10.40pm

Will the real inflation rate please stand up - overlooked "quirks" of a favoured chain-linking technique

Jens Mehrhoff
Deutsche Bundesbank, Frankfurt am Main, Germany

After Eurostat published the December 2013 flash estimate of euro-area annual inflation (0.8%), Financial Times writer Claire Jones reported on 8 January that a "methodological quirk in the calculation of holiday costs in the currency bloc's largest economy helped push core inflation in the eurozone to an all-time low in December". One day later, the President of the European Central Bank, Mario Draghi, related to this statement in the press conference Q&A.

Yet, there was no underestimation of inflation due to the "quirk" in December 2013! Quite the contrary, annual inflation had been overestimated in Germany from January to November 2013. More precisely, accommodation services and package holidays (totalling to less than 5% weight) drove up measured inflation from 2012 to 2013 by 0.3 percentage point for Germany and just under 0.1 percentage point for the euro area.

This was the outcome of the legally binding linking procedures in the HICP (chaining the new price series to the month of December of the previous year); it, hence, seems likely that such twists happened not only in Germany but in other countries as well - however, they were not detected. This is rather unfortunate since marked statistical distortions in the measurement of annual rates of inflation seriously hamper the interpretability of the HICP figures.

In this talk, the impacts on price index levels, year-on-year rates and average annual inflation are quantified; and a possible way forward for the future - change to a more robust linking method - is discussed.

9.2 Contributed – Loans, Prices, Inflation and the Income Gap

Thursday 4 September 9.20pm – 10.40pm

Crossing the income gap: understanding and presenting the evolution of income disparity in the UK

Stefano Ceccon[1,2], Nicola Hughes[1], Callum Christie[1], Megan Lucero[1]
[1]The Times and The Sunday Times Data Team, News UK, London, UK, [2]School of Health Sciences, City University London, London, UK

Purpose: Evolution of income distribution reflects economic status within society. Nevertheless, these issues are often presented with complicated metrics and fragmented news, increasing the distance between the public and policy. The aim of this study was to analyse the evolution of sex and age income disparity in the UK and to present it in a compelling, simple and effective way.

Methods: Over 90,000 answers to the Living Costs and Food Survey [1] for years 1971 and 1981 to 2011 were studied retrospectively. Inflation-adjusted median income was estimated for each age and sex using bootstrapping. The income median figures were then analysed using a polynomial regression model. The predicted values of the model were presented through a digital tool in the form of interactive D3 animated graphs [2].

Results: From 1971 to 2011 the age at which the population reached the median income was found to have increased (24.9 to 29.3). Women income was found to have reached the median population income in 1998 and the income gap between men and women was found to be narrowing. The tool allowed the public to understand the results of the analysis and see how their income ranked in the 1971 and 2011 economy according to their age and gender.

Conclusions: Analysis of survey data allowed quantifying the evolution of income inequalities in the UK. The interactive tool was deemed successful and provided a simple yet effective communication of the results.

[1] Office for National Statistics, 2011. Living Costs and Food Survey series.
[2] http://www.thesundaytimes.co.uk/sto/public/gendergap

9.3 Contributed – Genetic sequencing

Thursday 4 September 9.20pm – 10.40pm

Using the Bayesian Normal Gamma prior to identify associated genetic sequence variants.

Elizabeth Boggis, Kevin Walters, Marta Milo
University of Sheffield, Sheffield, UK

The Normal Gamma prior, a Bayesian adaptive shrinkage method which is implemented using MCMC, is compared to other statistical methods as an eQTL approach to identifying causal or associated genetic mutations (SNPs). The methods are compared on simulated data, where the results show the Normal Gamma prior to be a far superior method. On human data it is more difficult to assess the results for accuracy, but we can conclude that the Normal Gamma prior highlights SNPs in concordance with other methods. We also note that the Normal Gamma prior, although enforcing very harsh shrinkage, reports many less false positive SNPs than other methods.

We develop the Normal Gamma prior to include functional information which we use to differentially penalise synonymous and non-synonymous SNPs, as well as other types of SNPs where necessary. In initial simulation studies, the prior distribution penalises synonymous SNPs on average more than non-synonymous SNPs due to a larger functional significance score for the non-synonymous SNPs. The effect of this on the differential shrinkage between the two sets of SNPs can be seen in the posterior rankings and effect size estimates. We believe that this differential shrinkage form of the Normal Gamma prior is a powerful tool for detecting causal or associated SNPs.

9.3 Contributed – Genetic sequencing

Thursday 4 September 9.20pm – 10.40pm

Statistical challenges from next-generation sequence data

Arief Gusnanto[1], Charles Taylor[1], Henry Wood[2], Ibrahim Nafisah[1], Stefano Berri[2], Pamela Rabbitts[2], David Westhead[1]
[1]University of Leeds, Leeds, UK, [2]Leeds institute of Cancer and Pathology, Leeds, UK

The next-generation sequencing (NGS) has transformed the way we investigate genomic features. The technology generates short genomic sequences from biological specimens. After they are mapped to a reference genome, they can be utilised to indicate whether, for example, copy number alterations or gene regulations have taken place in some genomic regions. However, statistical inference from such data is not straightforward. Some preparational analyses are needed before any inference can be performed, and even more crucially in the case where the sequence data were obtained from clinical tumour samples. Random variation, GC imbalance, difference in the total number of sequences, contamination by normal cells, 'binning' the mapped sequences, and dependencies between samples are some of the challenges that we face in the preparational analyses. Once this is dealt with, then some challenges in statistical modelling and inference are awaiting, including how to build some appropriate tests for gene regulation and copy number alterations, and to classify cancer patients based on their genomic profiles. In this talk, I will discuss and review the steps that we have taken in our collaborative work to deal with some of those challenges.

9.3 Contributed – Genetic sequencing

Thursday 4 September 9.20pm – 10.40pm

Accurate statistical tests to detect changes in genome regulation using next-generation sequence data

Ibrahim Nafisah[1,2], Arief Gusnanto[1], Charles Taylor[1], David Westhead[3], Vijaya Shanmugiah[3]
[1]Department of Statistics, University of Leeds, Leeds, UK, [2]Department of Statistics, Faculty of Science, King Saud University, Riyadh, Saudi Arabia, [3]Faculty of Biological Science, University of Leeds, Leeds, UK

Next-generation sequencing (NGS), or high-throughput sequencing, has greatly transformed the way we investigate the effect of a specific transcription factor on genome regulation via ChIP-seq. The main challenge is to test the differences of genome regulation patterns between two samples, for example between wild-type and knock-out samples. Current approaches to this problem consist mainly of two steps: firstly, identification of enriched regions ($peaks$) in each sample and, secondly, comparison of differences of peaks at corresponding locations in both samples. This approach suffers from some drawbacks. Firstly, some considerable subjective view is still needed to define a peak, and secondly, the p-values of comparison are rendered meaningless since no underlying model for the peak nor the comparison is assumed. As a consequence, we are no longer sure whether we have controlled false positives at the desired level in the experiment. We develop a new statistical method to deal with this problem using a simple Poisson assumption on the pattern of genome regulation. Based on this assumption, we construct novel statistical tests, which enable us to rigorously identify differences in the genome regulation between two samples. A simulation study demonstrates that the novel tests have better sensitivity and proper control of false positives compared to currently existing methods. We illustrate their applications using a real dataset on the effect of $RUNX1$ on genome regulation, and some simulated data.

9.3 Contributed – Genetic sequencing

Thursday 4 September 9.20pm – 10.40pm

Reconstructing population ancestry using wavelet methods

Jean Sanderson, Murray Cox
Massey University, Palmerston North, New Zealand

Modern Pacific peoples have a complex history of both Asian and Melanesian ancestry. This shared ancestry is characterised by the production of `mosaic genomes' with alternating blocks of DNA from each ancestral population. The aim of this work is to use localised genetic information to learn about the demographic processes that have produced the observed genetic patterns. For example the date at which the two ancestral populations merged, the duration over which it happened, and the relative proportions of each population. Firstly, a principal components based approach is applied to sliding windows along the genome. Each individual genome is thus represented as a signal that reflects ancestry in local genomic regions. Previous work has demonstrated the ability of this technique for dating time of admixture, however the full potential for recreating more complex ancestral processes is yet to be addressed. By modeling the signals as locally stationary wavelet processes (LSW), we present a statistically rigorous framework for describing the observed variation. The LSW model provides a location-scale decomposition that captures several important features for reconstructing population ancestry. The methods are demonstrated using real genetic data from the Pacific region, as well as data from the MaCS genome-wide coalesent simulator under a range of different admixture scenarios.

9.4 Contributed – Data linkage and privacy

Thursday 4 September 9.20pm – 10.40pm

Using data linkage to create a national database of birth and maternity data

Nirupa Dattani[1], Alison Macfarlane[1], Preeti Datta-Nemdharry[2,1]
[1]City University London, London, UK, [2]Medicines and Health Research Agency, London, UK

Background

In England and Wales, the data recorded at birth registration are mainly socio-demographic such as names, address of residence, place of birth, occupation of the parents and date of birth. Different systems with little demographic information are used to record care at delivery. A third dataset, recorded when babies NHS number are allocated contains key items not present in the other two systems. Linkage of all these different national datasets enables analysis of births by time of day, ethnicity, country of birth of mother, gestational age and other socio-demographic and clinical information.

Method

Using data for 2005, linkage of NN4B to birth registration was piloted using NHS number, date of birth, postcode and sex. This was subsequently linked to maternity records in the Hospital Episode Statistics (HES) and Patient Episode Database for Wales (PEDW), using mother's NHS Number. For records where mother's NHS number was missing, different combinations of other indirect patient identifiers such as mother's date of birth, postcode and baby's date of birth were used.

Results

Linkage rate for birth registration to NHS numbers for Babies dataset was 99.8 per cent. The linkage was mainstreamed from 2006 onwards.

Overall 91 per cent of delivery records in Maternity HES and PEDW were linked to the birth registration/NHS Numbers for Babies records.

Discussion

Good linkage rate was obtained in linking national datasets on births and maternities. But to gain maximum benefit from this linkage, improvements are urgently needed in the quality of the data contained in Maternity HES.

9.4 Contributed – Data linkage and privacy

Thursday 4 September 9.20pm – 10.40pm

Are Partially-Synthetic Data the Answer To Care.Data's Woes?

Emmanuel Lazaridis[1]
[1]University College London, London, UK, [2]University of Southampton, Southampton, UK

Our team have been developing non-parametric imputation software for use in statistical disclosure control for large datasets, and a new method for identifying sensitive records: i.e., records associated with an increased risk of inadvertent disclosure of the individual data subject's identities. We focus on identifying risky data segments within individual records and synthesising those records using tree models and the Bayesian bootstrap. The new approach results in partially synthetic data sets that are estimated in such a way as to enhance data subject non-identifiability against a potentially-wide array of attacker scenarios. We apply this methodology to health data from the Myocardial Infarction National Audit Project to create micro-data collections in which data subjects cannot be precisely identified through data linkage, while multivariate structures in released data are largely preserved. Our work indicates that record-level health data, including data collected through the "care.data" initiative of the National Health Service which is to gather and link data from all care settings in England and Wales, may be made more widely accessible to the general public.

9.4 Contributed – Data linkage and privacy

Thursday 4 September 9.20pm – 10.40pm

The Challenges of Consent in Linking Survey and Administrative Data: Evidence from The Millennium Cohort Study.

Tarek Mostafa
Institute of Education, London, UK

This paper explores consent in linking survey and administrative data. It examines the correlates of consent across several domains and the cross-domain correlations using multivariate probit models. It expands our knowledge in different directions. First, it examines the predictors of consent for the main respondent across multiple domains. Secondly, it examines the determinants of consent at the household level when consent is sought from the main respondent and the partner. Thirdly, it models consent for the same domain and same respondent over time. The findings show that non-consent leads to bias in sample composition because it is related to a number of socio-demographic characteristics such as ethnicity, respondent's age, and income item non-response. However, the magnitude of this bias is not concerning since consenters do not differ substantially from the entire sample in terms of their characteristics. The impact of the predictors of consent varies according to the domain of consent, to whom is answering the question, and to whom consent is sought for. The findings also show that a strong and significant latent propensity to consent exists across domains for the same respondent, for the household as a whole, and over time.

9.4 Contributed – Data linkage and privacy

Thursday 4 September 9.20pm – 10.40pm

Extending log-linear capture-recapture models to handle erroneous records in linked administrative data to estimate population counts

Dilek Yildiz[1], Peter W.F. Smith[1], Peter G.M. van der Heijden[1,2]
[1]Southampton Statistical Sciences Research Institute, Southampton, UK, [2]Utrecht University, Utrecht, The Netherlands

The Beyond 2011 Programme of the ONS is evaluating alternative methods for collecting census data and producing small-area socio-demographic statistics. In the absence of a traditional census, one alternative method to estimate population counts is to use individual linkage between administrative sources. However, there are two problems with this approach: under-coverage and over-coverage. The usual log-linear capture-recapture models only deal with the under-coverage problem. We will extend this approach so that it can also take over-coverage into account.

The Patient Register and the Customer Information System (CIS) are two administrative sources for England and Wales. However, they are not designed to estimate the usual resident population and they collect information from different but overlapping population groups. The aim of this paper is to estimate the population of England and Wales by using information from the linked 2011 Patient Register and 2011 CIS as well as the 2011 Census estimates data table. For this particular research, we use one-way and two-way marginal information from the Census estimates. However, in the future we assume that the marginal information will be available from other sources such as coverage surveys, annual surveys or other administrative data sources.

Different log-linear models are considered to estimate population with or without the information from the Census estimates. Preliminary results show that the models which combine marginal information from the Census estimates and the linked Patient Register and CIS provide estimates that are closer to the census than the models which only use information from the linked data.

9.5 Contributed – Communication of Statistical Ideas

Thursday 4 September 9.20pm – 10.40pm

Five Steps for Analytical Insight

Paul Askew
Chartered Society of Physiotherapy, London, UK

This paper presents a consistent, structured and widely applicable approach for the communication, assessment and comparison of a range of data analysis. This applies to a range of substantive application areas including performance and operations and is particularly suited to the public sector and official and national statistics application. Example application sectors have included health, education and criminal justice, and in the context of central and local government policy development, regulation and investigation, and social media.

This evolving meta approach includes five incremental steps or components, which collectively provide a balanced and multidimensional overview. In summary these are: measurement, trend, benchmark, target and trajectory. Each of these primary components is further comprised of multiple and specific sub-components which are relatively weighted and presented collectively to provide an aggregated overview. The approach is used as context and focus for hypothesis driven analysis as well as a framework into which data-mining led analysis can be ordered.

With the increasing open and big data developments, and the related volume and complexity, this methodology provides a practical way to organise and structure the potential insights. The approach enables a proactive structure to encourage balanced and effective insight. In addition it provides an environment to enable more lay and public understanding and contextual interpretation of presented data, statistics and analysis, and the opportunity to enable targeted questions and challenge.

9.5 Contributed – Communication of Statistical Ideas

Thursday 4 September 9.20pm – 10.40pm

Sequence analysis as a graphical tool for investigating call record data

Olga Maslovskaya, Gabriele Durrant, Peter Smith
University of Southampton, Southampton, UK

For interviewer administered surveys many survey agencies nowadays routinely collect call record data. Examples of such data may be recordings of the day, time and outcome of each call attempt. Researchers have increasingly become interested in how best to use and analyse such data which can be large and may exhibit complex hierarchical and time-dependent data structures. It is hoped that a better understanding of the calling patterns and the mechanisms leading to particular call sequences will help improve data collection through improved interviewer calling practices. It might help identify more difficult cases and unusual interviewer behaviour earlier on in the data collection procedure and may provide strategies for improved nonresponse adjustment methods. Although survey researchers have become increasingly interested in understanding and improving the process of data collection, to date analysis of whole call histories is still limited.

This paper introduces sequence analysis as a graphical tool for investigating call record data first in cross-sectional and then in longitudinal context to better understand and improve survey processes. Sequence analysis offers an elegant way of visualising, displaying and summarising the normally quite complex call record data. Here, the method is used to inform survey management for adaptive and responsive survey designs. In this paper sequence analysis is combined with clustering, optimal matching and multidimensional scaling. The sequence analysis method is applied here to call record data from the first three available waves of the UK Understanding Society survey. Implications of the findings for survey practice are discussed in the paper.

9.5 Contributed – Communication of Statistical Ideas

Thursday 4 September 9.20pm – 10.40pm

The right graph for the right audience

Gordon Blunt
Gordon Blunt Analytics Ltd, Leicestershire, UK

Statisticians are well aware of the importance of visualising data.

Visualisation is essential in exploratory data analysis, but is also an important part of (for example) checking model diagnostics or when seeking structures in data. The development of software and computing power ensures that a wide array of graphical types are now possible, using a few lines of code or with a few clicks of a mouse.

In a commercial environment, however, the most commonly used graphics - often the only ones - are bar, line and pie charts. Partly this is a consequence of the fact that the default tools for data analysis and reporting are spreadsheets and presentation software. As a result, people may not have access to the best tools for the job. In addition, they may not have much training in a numerate discipline, where they could have learned some of the principles of good data analysis.

This talk will suggest ways that the statistician - whose clients are in the commercial sector - can help people derive greater value from their data. It will also describe methods the analyst can use to convince business people to move beyond the 'standard' chart types.

Examples will be given, based on 20 years working in this area, of helping clients understand their data by use of appropriate statistical graphics. It will also describe how we should feel confident to use complex plots, and will suggest suitable ways of introducing them to clients and the analysis of commercial data.

9.5 Contributed – Communication of Statistical Ideas

Thursday 4 September 9.20pm – 10.40pm

Crocheting the Normal Distribution: Exploring Creative Public Engagement with Statistics

Victoria Gorton
Lancaster University, Lancaster, Lancashire, UK

Most people do not have a sufficient level of statistical literacy to allow them to engage critically with the statistics present in their daily lives (von Roten, 2006). Improving statistical literacy is often framed as an educational issue, with emphasis placed on the transfer of knowledge from experts to the public (Pullinger, 2013; Gal, 2002). However, this deficit model has come under criticism, with some arguing that, rather than being ill-informed, the public simply hold a different position in relation to statistics (Miller, 2001).

Given negative public attitudes, it is argued that adopting a more creative, dialectic approach, to public engagement could help challenge these attitudes and facilitate improvements in statistical literacy, through acknowledging these different positions.

Here, I present a small-scale Masters project that aimed to adopt a more creative approach to public engagement with statistics. As part of this, a series of objects - Norm and his friends, crocheted probability distributions – were displayed with 138 visitors' reactions gathered. The results suggested that adopting a creative approach can help lower anxiety and challenge understandings of statistics. Several limitations and considerations for future projects are offered, including the difficulty of managing the wide attitudes and prior knowledge of the public.

Overall, it is argued that further research into public engagement with statistics specifically is essential if public attitudes and understandings are to be improved.

9.6 Contributed – Trial Design

Thursday 4 September 9.20pm – 10.40pm

On a simple quickest detection rule for health-care technology assessment

Daniele Bregantini, Jacco Thijssen
University of York, York, UK

In this paper we propose a solution to the Bayesian problem of a decision maker who chooses, while observing trial evidence, an optimal stopping time at which either to invest in a newly developed health care technology or abandon research.

We show how optimal stopping boundaries can be computed as a function of the observed cumulative net benefit derived from the new health care technology. At the optimal stopping time, the decision taken is optimal and the decision maker either invest or abandon the technology with consequent health benefits to patients. The model takes into account the cost of decision errors and explicitly models these in the payoff to the heath care system. Mathematically, the model is a sequential hypothesis test on the trend of an arithmetic Brownian motion. The stopping boundaries of this problem are characterized and analysed.

The implications in terms of opportunity costs of decisions taken at sub-optimal time is discussed and put in the value of information framework. In a case study it is shown that the proposed method, when compared with traditional ones, gives substantial economic gains both in terms of QALYs and reduced trial costs.

9.6 Contributed – Trial Design

Thursday 4 September 9.20pm – 10.40pm

Sample size calculations for comparative pharmacoepidemiology surveillance studies in healthcare databases

Nicholas Galwey, Michael Irizarry, Ashley Wivel, Jennifer Christian, Marilyn Metcalf
GlaxoSmithKline, Stevenage, UK

Ongoing surveillance of medicine safety after product launch poses well-known challenges with regard to availability of data and its interpretation; however, the use of claims data and electronic medical record (EMR) databases offers the prospect of new approaches. One of the key questions when planning such studies will be how long it will take to obtain enough data to address a particular safety question at a specified level of confidence.

We present statistical methods for making such predictions, adapting efficacy power calculations to address non-inferiority. A specified degree of risk elevation due to exposure is the null hypothesis to be rejected, and equivalence of risk in exposed and unexposed individuals is the alternative hypothesis. If the duration of exposure in exposed individuals and the follow-up time in unexposed individuals are nearly equal and constant, the calculation can be expressed in terms of a risk ratio. If they are highly variable, it should be expressed as an incidence rate ratio. To determine the required sample size, assumptions are made concerning the background incidence rate and the ratio of unexposed to exposed individuals. Additional assumptions concerning the number of eligible patients and the rate of uptake of the new medicine then permit prediction of study duration. However, assumptions should be checked over time and predictions updated accordingly. Considerations regarding the clinical context (e.g. severity of event) and potential confounders (e.g. early adopter effects, channeling bias etc.) can be taken into account and will be reflected in the predicted duration.

9.6 Contributed – Trial Design

Thursday 4 September 9.20pm – 10.40pm

How big should a pilot cluster randomised controlled trial be?

Stephen Walters
University of Sheffield, Sheffield, UK

Health technology assessment often requires the evaluation of interventions which are implemented at the level of the health service organisation unit (e.g. GP practice) for clusters of individuals. In a cluster randomised controlled trial (cRCT) clusters of patients are randomised. When estimating sample sizes for cRCT designs additional information, over that required for individually randomised controlled trial designs, is required: namely the expected cluster size and the intra cluster correlation (ICC).

Feasibility or pilot studies are usually pieces of research done before a main study in order to estimate important parameters (e.g. standard deviation of the outcome; number of eligible patients; follow-up rates; adherence rates; ICC for cRCTs) that are needed to design the main study. If a feasibility study is a small randomised controlled trial the usual sort of power calculation is not normally undertaken. Instead the sample size should be adequate to estimate the critical parameters (e.g. ICC) to the necessary degree of precision.

This project will investigate how many clusters we need in a pilot/feasibility cRCT to estimate the ICC with a reasonable degree of precision for continuous outcomes. We estimate the precision or standard error of the ICC for a variable number of clusters, subjects per cluster, and cluster effects.

There is diminishing marginal gain in the precision, of the ICC, after randomising 8 to 10 clusters; and 50 subjects per cluster. For situations where there are a limited number of clusters (<= 10) available investigators should consider an alternative internal pilot study design.

9.6 Contributed – Trial Design

Thursday 4 September 9.20am – 10.40am

Sequential Clinical Trial Design when Outcomes Arrive with Delay: a Bayes Decision-Theoretic Approach

Stephen Chick[2], Paolo Pertile[3], Martin Forster[1]
[1]University of York, York, UK, [2]INSEAD, Fontainebleau, France, [3]University of Verona, Verona, Italy

A sequential clinical trial permits researchers to monitor outcomes as they accumulate, offering the potential to stop sampling early, resulting in benefits for patients and resource savings. However, the problem of how to design trials when observations on the primary endpoint arrive with delay continues to challenge researchers. Hampson and Jennison's recent work (Hampson and Jennison, 2013) marks an important step forward in designing group sequential trials with delay according to prespecified Type I and Type II error rates.

We present a Bayesian, decision-theoretic, model of sequential clinical trial design when there exists delay in observing the primary endpoint. Our work builds on two recent contributions to the literature: Pertile et al. (2014), who developed optimal stopping rules in a sequential clinical trial in the absence of delay and Chick and Gans (2009), who solved models of optimal simulation selection using similar methodology. The model differs from that of Hampson and Jennison in a number of respects: rewards are made a continuous function of the primary endpoint, costs of sampling and monitoring are incorporated, values accruing to the trial participants are explicitly measured and prior information may be used to establish whether or not it is worth conducting a sequential trial or simply carrying out a trial of fixed sample size. An application illustrates the ideas and potential resource savings.

S. Chick and N. Gans (2009). Management Science. 55(3):421-437.

L. Hampson and C. Jennison (2013). JRSSB. 75(1):3-54.

P. Pertile, M. Forster and D. La Torre (2014). JRSSA. 177(2):419-438.

PD9 – Career development pathways for professional statisticians

Thursday 4 September 9.20am – 10.40am

Rob Mastrodomenico

Owner: Global Sports Statistics

Uli Burger

Novartis Pharmaceuticals R&D

Despite what many have predicted over three decades or more, the days of a 'job for life' are not truly gone. Even so, the concept of a linear career pathway based on promotion within a discipline is today only one of many options for professional statisticians in what has become a very pluralistic and diverse work environment. In this session two professional statisticians from quite different business sectors share a little of their own career histories as well as some of the insights into professional development they gleaned along the way.

10.1 Invited – Checking and Cleaning in Big Data

Thursday 4 September 11.10am – 12.40pm

Efficient Algorithms for Statistical Data Editing and Imputation

Ton de Waal[1,2]
[1]Statistics Netherlands, The Hague, The Netherlands, [2]Tilburg University, Tilburg, The Netherlands

National Statistical Institutes (NSIs) and other official statistical institutes have the task to provide high quality statistical information on many aspects of society, as up-to-date and as accurately as possible. This task has to be carried out as efficiently as possible, in terms of budget and response burden. In the past NSIs used to rely mainly on survey data. Over the last few decades these survey data have been supplemented by administrative data. A recent development is the availability of Big Data. One of the difficulties in performing the above-mentioned task arises from the fact that the data sources that are used for the production of statistical output - survey data, administrative data and Big Data - inevitably contain errors and missing values. These errors and missing values may affect the estimates of publication figures. In order to prevent substantial bias and inconsistencies in publication figures, NSIs therefore carry out an extensive process of checking the data and correcting them if necessary. This process encompasses a variety of procedures that together are referred to as statistical data editing. Replacing erroneous fields and, in particular, filling in missing values with (better) values is called imputation. In this presentation we first give an overview of efficient statistical data editing and imputation procedures that have originally been developed for traditional survey data. After a brief characterization of Big Data, we then examine to what extent these editing and imputation procedures can be applied to Big Data.

10.1 Invited – Checking and Cleaning in Big Data

Thursday 4 September 11.10am – 12.40pm

Beyond the codes: Making use of free text in e-health datasets

Elizabeth Ford
Brighton and Sussex Medical School, Brighton, UK

Research using GP databases, and other sources of electronic patient records (EPRs) is growing apace, with studies describing disease prevalence and aetiology, health service needs, prescribing practices and side effects of medications. Results of these studies are used by health service commissioners, pharmaceutical companies, and drug regulatory authorities. Historically, GP database studies have used only the medical codes in the records and not the free text. Text is made up of consultation notes and letters to and from the GP practice. There are many influences on whether GPs will add information as a code or in narrative text. Free text may contain a wealth of additional information, particularly regarding conditions with a stigma attached, patient reported outcomes, social information and culturally unsanctioned behaviours such as substance misuse, which may be recorded by GPs "under the radar".

Our work has shown that in patients with rheumatoid arthritis (RA), 12% of entries in their GP patient record contained RA-relevant disease information in the free text, such as symptoms, less specific diagnoses, and test results.

This talk will review and quantify the contribution that free text can make to our understanding of diseases when using EPRs for research. It will describe the basic process of, and barriers to, incorporating free text into studies, from keyword searching to full text annotation. It will give a flavour of the natural language processing technology which can be employed to automate extraction of the relevant information from the free text.

10.2 Invited – Mega trends in statistics

Thursday 4 September 11.10am – 12.40pm

Introduction and why megatrends matter

John Bibby

This session is an opportunity to take part in a high level discussion about the state of play of statistical thinking in today's society where the opportunities to make data driven decisions appear to be greater than ever. The session was inspired by personal observations made at Quality & Improvement Section meetings and so it will start with the personal thoughts and observations of two speakers. Both talks are intended to be provocative and to encourage debate in the open session!

10.2 Invited – Mega trends in statistics

Thursday 4 September 11.10am – 12.40pm

The Ghosts of Statistics Past, Statistics Present & Statistics Future.

Tony Bendell
the Anti-Fragility Academy & Services Ltd, Nottingham

Where we are today is not entirely the logical consequences of where we started from. The early history of statistics and of the RSS concerned the gathering of information about society, and it was many decades before mathematical approaches were regarded as part of the statistical approach.

So, where are we today, why are we here, is it the best place for us, and where do we want to be in the future? Mega trends in statistics are driven by external drivers, whether or not the statistical profession fully participates. We, the profession, may not want to change, but we may have to in order to survive. Or we may chose not to. Like Scrooge, confronting our origins, the reality of the present, and the inevitable consequences of our attitudes, actions and behaviours, should help us to face the future and be more worthy citizens.

The content of this talk, and the possible scenarios presented, is very much a personal view from someone who, many years ago was disciplined by the RSS Professional Affairs Committee for daring to suggest publically that Statistics should not just be done by statisticians…

10.2 Invited – Mega trends in statistics

Thursday 4 September 11.10am – 12.40pm

Are Statisticians shaping the Future of Statistical Thinking?

Nigel Marriott
Marriott Statistical Consulting Ltd

Analysing data to understand society and shape decision making is more prominent in the news these days. Buzz words such as “Data Science”, “Analytics”, “Open Data”, “Big Data” appear regularly and may appear threatening to traditional statistical thinking but equally they could offer an opportunity for statistical thinking to move beyond its traditional areas. The question that needs to be answered is whether these trends are being shaped by statisticians or not.

Using his personal experience as a consultant who has worked in many industries and more importantly as a trainer running courses in statistical thinking for non-statisticians, the speaker will compare three different scenarios for introducing statistical thinking to non-statisticians; How it is done today, how it might be done in 2050 if statisticians shape the future, how it might be done in 2050 if statisticians do not shape the future of statistical thinking. Again these thoughts will be personal views to provide food for thought for the discussion session.

10.3 Invited – Anonymisation Practices for Sharing Data

Thursday 4 September 11.10am – 12.40pm

EUL or OGD: a penetration test on two survey datasets.

Mark Elliot[1], Elaine Mackey[1], Susan O'Shea[1], Keith Spicer[1], Caroline Tudor[1]
[1]University of Manchester, Manchester, UK, [2]Office for National Statistics, Titchfield, UK

The transparency agenda is forcing data stewardship organizations to review their dissemination policies and to consider whether to release data which are currently only available under end user license as open data. Here we describe the results of a study providing evidence about the risks of such an approach via a simulated attack on two social survey datasets. This is the first systematic attempt to simulate a jigsaw identification attack on an anonymised dataset. The information that we draw on is scraped from multiple online data sources and augmented with purchasable commercial data. The results indicate that such an attack against open anonymised datasets is possible and that any move to open data should proceed with caution.

10.3 Invited – Anonymisation Practices for Sharing Data

Thursday 4 September 11.10am – 12.40pm

Data pseudonymisation and (privacy-preserving) record linkage

Duncan Smith
University of Manchester, Manchester, UK

Data on individuals are often contained in distinct databases at distinct locations. Record linkage techniques can be used to create combined datasets for research and planning purposes. In the absence of reliable, unique identifiers probabilistic methods are often used. Unfortunately the variables that are most useful for probabilistic record linkage (e.g. name, age, address) are those that are most likely to be withheld due to privacy concerns. Pseudonymised (hashed) values can be used in place of the underlying values. Pseudonymisation can be carried out in such a way that a pair of pseudonyms can be used to estimate similarity between the underlying values. This information can be exploited by probabilistic record linkage approaches. However, the most commonly used pseudonymisation approach is fairly insecure against certain froms of attack, and information regarding similarities tends to be incorporated into record linkage in relatively ad hoc ways.

An alternative pseudonymisation approach will be presented. It offers a number of advantages over the existing 'standard' approach. It is easily implemented and is far more robust against certain forms of attack. A grounded approach to record linkage incorporating similarity scores will be presented. It uses maximum likelihood estimation via the Expectation-Maximization approach. Empirical results will show that it offers improved classification compared to the more ad hoc approaches.

10.3 Invited – Anonymisation Practices for Sharing Data

Thursday 4 September 11.10am – 12.40pm

New developments in data sharing and the application of data linkage in Scotland from a health perspective

Anthea Springbett, Carol Morris
NHS National Services Scotland, Edinburgh, City of Edinburgh, UK

Scotland has some of the best health data in the world coupled with a population of an ideal size to facilitate national linkage projects. There is a very strong track record of health related research in which NHS Scotland has played a significant role, primarily as a data custodian and provider in collaboration with university researchers.

The Scottish Health Informatics Project (SHIP) brought together a number of academic institutions and NHS Scotland to provide a platform for enabling research and data linkage. This programme of work also produced innovative developments in information governance to help facilitate the implementation of cross sector data linkage projects.

The success of SHIP and the fostering ofcollaboration between Scotland and the rest of the UK linkage community have led to a new generation of linkage projects involving both health and administrative data. This talk will provide an overview of these new developments (including the Farr Institute and the Administrative Data Research Network) together with a discussion of the technical and governance challenges associated with the rapidly expanding fields of data sharing and data linkage.

10.4 Invited – Floods: risk assessment, management and attribution to climate change

Thursday 4 September 11.10am – 12.40pm

Towards realistic models of flood risk: a multivariate conditional exceedance approach to assess the probability of flood events

Rob Lamb[1,3], Ye Liu[2], Jonathan Tawn[3]
[1]JBA Trust, Skipton, North Yorkshire, UK, [2]JBA Risk Management, Skipton, North Yorkshire, UK, [3]Lancaster University, Lancaster, UK

There is a well-established tradition of statistical analysis in flood engineering, largely based on univariate approaches to estimate probability distributions for river flows or water levels. Typically each location on a river or coast has been treated separately and flood risk analysts have been trained to think in terms of “design events”. But the design event is an abstraction of reality. In nature, flood events have widely varying spatial patterns and evolve dynamically. In addition, flooding often involves a combination of physical processes such as the combined effects of high river flows and extreme sea levels.

Recently there has been significant progress made in the multivariate statistical analysis of flood risk to provide more realistic models. One such approach is based on the conditional probabilities of joint threshold exceedances. We will describe how this conditional exceedance approach has been applied to model realistic flood events at national scale. The focus will be on applications for flood risk management in the UK, which include estimation of loss distributions for the (re)insurance industry, improvements in understanding of flood risk for national strategic planning and assessment of risk to infrastructure networks. The talk will also touch on some of the important needs for further research and development.

10.4 Invited – Floods: risk assessment, management and attribution to climate change

Thursday 4 September 11.10am – 12.40pm

Bayesian Uncertainty Analysis for Hydrological Computer Models

Nathan Huntley, Michael Goldstein
Durham University, Durham, UK

When using computer models for physical processes, such as hydrological models, we must take care to account for all relevant uncertainties, such as those in the model inputs and those from the inherent limitations of the model. Some of these uncertainties can only be specified by model experts, but others can be estimated using perturbations of the forcing functions and by running other experiments on the model. We consider a collection of simple experiments that can be performed on practically any computer model to estimate the internal model uncertainty at a given set of parameters.

This approach is quite computationally intensive, so it is infeasible to provide such estimates at very many parameter choices. We can however use a standard strategy of emulation to create an estimate (and associated measurement of uncertainty) for the internal uncertainty at any parameter choice we have not sampled. This estimate can then be used in the following analysis of the model.

We present an example of such a strategy for the hydrological model FUSE, showing how to arrive at estimates for internal uncertainty for a given parameter choice, how to emulate this uncertainty for any parameter choice, and how to incorporate this analysis into a standard approach to determine plausible parameter choices to use in the future.

11.1 Invited – Who's afraid of data science

Thursday 4 September 2pm – 3.20pm

Peter Diggle

Lancaster University

Duncan Ross

Data Science, Teradata and Society of Data Miners

Sylvia Richardson

MRC Biostatistics Unit Cambridge Institute of Public Health

In recent years terms such as 'big data' and 'data science' have become relatively sexy. The statistical community has had a mixed relationship with these terms. Some have embraced them, others see them as representing something genuinely new and different, and many have felt they do not add much to what the discipline of statistics already covers. This panel session will bring a range of perspectives to the question of 'who is afraid of data science'? In other words how should the statistical community understand the rise of 'data science' and what might it mean for the future?

11.2 Invited – Advances in Astrostatistics

Thursday 4 September 2pm – 3.20pm

Gaussian processes regression for exoplanet detection and characterisation

Suzanne Aigrain[1], Stephen Roberts[1], Neale Gibson[2], Thomas Evans[1], Vinesh Rajpaul[1], Steven Reece[1]
[1]University of Oxford, Oxford, UK, [2]European Southern Observatory, Garching, Germany

The discovery and study of exoplanets (planets orbiting other stars than the Sun) has become one of the most active and exciting fields in Astrophysics, but it presents astronomers with new statistical challenges. Direct detection of exoplanets is extremely challenging, so most studies use indirect methods, which involve detecting the impact of the planet on the host star's position, velocity, apparent brightness or spectrum. All of these share some common characteristics: they rely on time-series data, the form of the planetary signal of interest is typically well known in advance, but its amplitude is very small compared to astrophysical and / or instrumental noise sources, which are rarely well-understood or predictable, and have a complex power spectrum. Gaussian process (GP) regression, introduced to the field relatively recently, offers an attractive solution to the problem of modelling the planetary signal and the noise simultaneously and marginalising over the latter to extract the former. I will present a number of applications of GP regression to exoplanet detection (using transits and radial velocities) and the characterisation of exoplanet atmospheres (using transit and eclipse spectroscopy). I will finish by highlighting some of the limitations of GPs for exoplanet applications, and outlining some alternatives which might enable us to overcome these issues.

11.2 Invited – Advances in Astrostatistics

Thursday 4 September 2pm – 3.20pm

MultiNest Algorithm for Bayesian Inference

Farhan Feroz
University of Cambridge, Cambridge, UK

Astrophysics and cosmology have increasingly become data driven with the availability of large amount of high quality data from missions like WMAP, Planck and LHC. This has resulted in the development of many innovative methods for performing robust statistical analyses. MultiNest is a Bayesian inference algorithm, based on nested sampling, which has been applied successfully to numerous challenging problems in cosmology and astroparticle physics due to its capability of efficiently exploring multi-modal parameter spaces. MultiNest can also calculate the Bayesian evidence and therefore provides means to carry out Bayesian model selection. I will give a brief description of this algorithm and review its applications in astrophysics and cosmology.

11.2 Invited – Advances in Astrostatistics

Thursday 4 September 2pm – 3.20pm

Sparsity in astrophysics: astrostatistics meets astroinformatics

Jason McEwen
University College London (UCL), London, UK

Astrostatistics has become a well-established sub-field, where powerful statistical methods are developed and applied to extract scientific information from astrophysical observations. In particular, Bayesian methods have now found wide-spread application in astrophysics. Astroinformatics, on the other hand, is an evolving but less mature sub-field, where informatics techniques provide a powerful alternative approach for extracting scientific information from observational data. Informatics techniques have close links with information theory, signal processing and computational harmonic analysis, and have been demonstrated to be very effective. Wavelet methods, for example, allow one to probe both spatial- and scale-dependent signal characteristics simultaneously. Such techniques are very effective in studying physical phenomena since many physical processes are manifest on particular physical scales, while also spatially localised. Recent developments in this domain have led to the theory of compressive sensing, a revolutionary breakthrough in the field of sampling theory, which exploits the sparsity of natural signals. I will discuss compressive sensing techniques from both the synthesis and analysis perspectives, including statistical connections, and the application of such techniques in astrophysics. In particular, I will discuss how sparsity can be exploited to study the cosmic microwave background (CMB), the relic radiation of the Big Bang.

11.3 Invited – Exploiting large genetic data sets

Thursday 4 September 2pm – 3.20pm

Learning gene knockout effects from wild type gene expression data

Marloes Maathuis
ETH Zurich, Zurich, Switzerland

We present recent progress on estimating bounds on causal effects from observational data, when assuming that these data are generated from an unknown directed acyclic graph. In particular, we present the IDA algorithm for this purpose. IDA is computationally feasible and consistent for high-dimensional sparse systems with many more variables than observations. We validated IDA in biological systems, and will present results on a yeast gene expression data set. Finally, we discuss possible instability issues in high-dimensional settings, as well as extensions towards allowing for hidden variables and predicting the effect of multiple simultaneous interventions.

11.4 Invited – Vic Barnett and his contributions to statistics

Thursday 4 September 2pm – 3.20pm

Statistics and Regulation, Inside and Outside

Tony O'Hagan
University of Sheffield, Sheffield, UK

Vic Barnett was highly influential in promoting statistics in the environmental sciences. In 1996, he was invited by the Royal Commission on Environmental Pollution to write a report on the statistical approach to handling uncertainty and variation in the setting of environmental standards, and he asked me to join him in preparing the report. This talk focuses on issues around statistics in the context of regulation.

An industry regulator aims to place such restrictions on the industry's activity as will reduce or minimise the risk of undesirable consequences for society at large. Where there is risk there must be statisticians! Sometimes the statisticians are inside regulation, working for the regulator, and sometimes they are outside, working for companies needing to satisfy the regulator. Sometimes both.

In this talk, I will sketch some applications that I have been involved with, both inside and outside regulation, including environmental standards, nuclear power, food, water, pharmaceuticals and railways.

11.4 Invited – Vic Barnett and his contributions to statistics

Thursday 4 September 2pm – 3.20pm

Combining simulated and observed wind speed data: Generating risk maps for electric grid disruptions in Portugal

K. Feridun Turkman
DEIO-CEAUL, Lisbon, Portugal

Extreme winds are one of the major causes of costly disruptions in electric grid. Risk maps that indicate likely places of such disruptions are important decision support tools for administering the power grid. Generating such risk maps depend on reliable wind data at fairly high spatial and temporal resolutions. Wind data are typically available at a limited number of monitoring sites, often with a signicant number of missing observations and outliers. On the other hand, data from a numerical-physical model are available at regular grid cells level, obtained at high spatial and temporal resolutions. However, these numerical models do not seem to match well the observed data, particularly on the tails. More accurate exposure assessment of wind on power grid disruptions can be made by combining observed and simulated data by bringing the simulated data into line with observations, particulary at the tails. We look at several alternative methods that can be used for this problem and in particular adopt the Tawn-Haernan(2004 J.R.Statist Soc.B,66 497-546) semi-parametric regression model and extend it to spatial context.

11.4 Invited – Vic Barnett and his contributions to statistics

Thursday 4 September 2pm – 3.20pm

Vic Barnett: some of his contributions to school statistics teaching.

Peter Holmes
RSS Centre for Statistics Education, Plymouth, UK

From the 1960s to the end of the century Vic was at the centre of most of the major developments in the school teaching of statistics in the UK and was also involved with international initiatives. I shall describe his role in many of these, including some low profile ones, and consider his approaches to the teaching of statistics.

Plenary 4

Thursday 4 September 3.40pm – 4.30pm

Recent advances in High Dimensional Covariance Matrix Estimation

Ming Yuan
University of Wisconsin-Madison

In this talk I will survey some of the recent progresses in large covariance matrix estimation and related problems.

POSTER PRESENTATIONS

Optimum Allocation of Multi-Items in Stratified Random Sampling using Principal Component Analysis

OLANIYI MATHEW OLAYIWOLA[1], F. S. APANTAKU[2], O.H Bisira[2]
[1]FEDERAL UNIVERSITY OF AGRICULTURE, ABEOKUTA, NIGERIA, ABEOKUTA, Nigeria, [2]FEDERAL UNIVERSITY OF AGRICULTURE, ABEOKUTA, NIGERIA, ABEOKUTA, Nigeria, [3]Lagos State Polytechnic. Lagos State, Nigeria., Lagos, Nigeria

The problem of allocation with more than one characteristic in stratified sampling is conflicting in nature, as the best allocation for one characteristic will not in general be best for others. Some compromise must be reached to obtain an allocation that is efficient for all characteristics. In this study, the allocation of a sample to strata which minimizes cost of investigation, subject to a given condition about the sampling error was considered.

The data on four socioeconomic characteristics of 400 heads of households in Abeokuta South and Ijebu North Local Government Areas (LGAs) of Ogun State, Nigeria were investigated. These comprised of 200 households from each LGA. The characteristics were occupation, income, household size and educational level. Optimal allocation in multi-item was developed as a multivariate optimization problem by finding the principal components. This was done by determining the overall linear combinations that concentrates the variability into few variables. From the principal component analysis, it was seen that for both Abeokuta and Ijebu data sets, the variance based on the four characteristics as multivariate is less than that of the variables when considered as a univariate.

From the results, it was seen that there was no difference in the percentage of the total variance accounted for by the different components from the merged sample when compared with the individual sample. Optimum allocation was achieved when there was stratification

Keywords: Stratified Sampling, Optimum allocation, Stratification, Optimization

Causal Inference through IV estimation - The effect of domestic violence on neonatal & infant mortality.

Seetha Menon
University of Essex, Colchester, UK

The number of children dying, worldwide, under the age of 5 is estimated at 6.9million (2011). India accounts for 1.7million of these deaths. This paper investigates domestic violence as a potential contributor to this situation. Specifically, is there a significant causal relationship between domestic violence and child mortality in India? In India, where the subjugation of women is the social norm, domestic violence has been estimated at approximately 40%. In addition, violence against women during pregnancy has been estimated at almost 13%. The current literature has succeeded in establishing an association between domestic violence and child mortality, but has to yet present evidence of a causal relationship. Data from the National Family Health Survey (NFHS 3, India) is used for this study. This dataset contains information on child mortality, maternal health, socioeconomic status and anthropometric data. In addition a domestic violence module was also executed. In order to overcome the inherent endogeneity of domestic violence in such analysis, an Instrumental Variable estimation strategy is used. The instrument used is the real price of gold at the time of marriage. Two models are estimated for neonatal and infant mortality. Results lend evidence to a bias in OLS estimates and show a positive and significant relationship between domestic violence and mortality in both models. The results remain robust to both two stage least square estimation and maximum likelihood estimation.

A PROPOSED NEW INFORMATION CRITERION FOR ORDER DETERMINATION IN TIME SERIES MODELING

Olanrewaju Shittu
University of Ibadan, Ibadan, Nigeria

Conventional model order selection criteria have been used by many researchers with considerable amount of success over the years. However, each of them has been identified with one limitation or the other with respect to their ability to correctly identify the correct order of an Autoregressive model. In this paper, a new information criterion (PNIC) capable of selecting the correct order of an Autoregressive model is proposed. The proposed model was derived as a convolution of the well-known order selection criteria with their respective weight of evidence will be compared with the existing order determination criteria.

The considered criteria are the Akaike information criterion (AIC); Bayesian information criterion (BIC) and the Hannan − Quinn criterion (HQ). Eight series of sizes 20, 30,100, 200, 500, 1000, were generated and two real data on money in circulation of size 30 and income data with size 467 were analysed. The comparison of the four model selection criteria was in terms of the number of times that they identify the “true” model.

The results show that for real and simulated data indicate that the proposed new information criterion (NIC) outperformed the conventional model order information criteria. It is therefore recommended as reliable criteria to identify the “true” model of any structure in time series modelling.

Keywords: Monte Carlo, information Criterion; Performance; Simulation.

Age Effect on Players' Performance in Twenty20 Cricket

Atanu Bhattacharjee[1], Hemanta Saikia[2], Dibyojyoti Bhattacharjee[3]
[1]Malabar Cancer Centre, Thalassery, Kerala, India, [2]Kaziranga University, Jorhat (Assam), India, [3]Assam University, Assam, India

Though most of the cricketing fraternity opines that Twenty20 cricket is a game of young cricketers, yet the performance of senior cricketers in the different seasons of IPL seems to be comparable to that of the youngsters. Thus, this study tries to examine the effect of age on on-field performance of the cricketers in Indian Premier League. To measure the performance of cricketers, a model is developed utilizing the four prime skills of the game viz., batting, fielding, bowling and wicket keeping. Various cricketing factors are considered related to the performance of cricketers under the above-mentioned skills. All these factors are normalized and accordingly adjusted by using appropriate weights on the basis of their relative importance. The performance measures are obtained for each cricketer separately in all the four seasons of IPL played so far and the regression model with random regression coefficient has been applied to determine the effect of age on cricketers' performance. The results obtained from the regression model confirm that the on-field performances of the cricketers are positively associated with the age of the players.

Keywords: age; performance measure; regression, statistics in sports, Twenty20 cricket

Multilevel Modeling of Two Level Survey Data

Shafquat Rozi[1], Sadia Mahmud[1], Gillian Lancaster[2]
[1]*Aga Khan University, Karachi, Pakistan,* [2]*Lancaster University, Lancashire, UK*

Introduction

Adolescents spend a considerable amount of their time in school, the school environment is therefore important for child outcomes. Random effect logistic regression model has been proposed to model correlation among subjects in the same cluster by including a cluster-specific random effect in the logit. Among young teens, about one in five smokes worldwide.

Objectives

- To develop two level random effects logistic regression model and GEE model to identify predictors of smoking on teenage children attending school.
- To develop random coefficient model to assess if the variability between schools is different for the public and private schools.
- To compare the results from the above mentioned models with a conventional logistic regression model.

Methods

A two-stage cluster sampling with stratification was employed. We interviewed 772 male secondary school students. The outcome variable is smoking status of the students. We have two level data with single level of clustering.

Results

Between cluster variance is significantly different from zero *(p-value of likelihood ratio test = 0.01)*, which indicates that there is variability between schools. The ICC quantifies consistencies among observations within each cluster and it is greater than zero *(ICC =0.15).*

A random coefficient model showed that there is variability among schools but it is not different for public & private *(p-value> 0.99)*.

Conclusion

Random effect model and GEE take correlation in to account in the inferential process and indicating that there is variability between schools and we need to take cluster variation in to account using multilevel modeling.

Cricket Captain's Dilemma: Selecting the Remaining Players

Dibyojyoti Bhattacharjee[1], Hemanta Saikia[2]
[1]Assam University, Silchar, Assam, India,[2]The Assam Kaziranga University, Jorhat, Assam, India

Selecting a balanced playing XI in cricket with the right mix of players by different specialization is a difficult decision making problem for the team management. To make the process more objective, optimization techniques can be applied to the process of selection of players from a given squad. Such an exercise has two dimensions. First, a suitable tool for performance measurement of cricketers needs to be defined. Secondly, for selecting a balanced team of XI players, an appropriate objective function and some constraints need to be formulated. Since the captain gets an obvious inclusion in the team, the area specialization of the captain influences the selection of other ten positions in the playing XI. The present study attempts to select the optimum balanced playing XI from a squad of players given the specialization of the captain using binary integer programming. To validate the exercise, data from the fifth season of the Indian Premier League has been used.

Statistical analysis and modelling of climatic variables

T. O. OLATAYO[1], O.I SHITTU[2], I.A. TAIWO[3]
[1]Olabisi Onabanjo University,Mathematical Sciences Department, P.M.B.2002., Ago-Iwoye, Ogun State, Nigeria, [2]University of Ibadan, Ibadan,Nigeria. Department of Statistics., Oyo State, Nigeria, [3]Olabisi Onabanjo University,Mathematical Sciences Department, P.M.B.2002., Ago-Iwoye, Ogun State, Nigeria

Climate change is an alteration in the state of the climate that can be identified by changes in the mean or the variability of its properties and that persists for an extended period, typically decades or longer. This study is then used to analyse the simultaneous relationship between annual mean series of some climatic variables, from south west Nigeria in order to have a more clear understanding about changes in climatic conditions.

The methods used are cointegration, two stage least square, vector autoregressive and granger causality test.

The results from the time plots showed that each year there is a simultaneously increase in the values of rainfall except a sharp shift in 2009, while temperature, humidity, evaporation and radiation fluctuate from year to year, cyclical in movement and a cycle is completed every three years. The Johansen co-integration test shows at least three climatic series are co-integrated at 5% level of significance. The two stage least square analysis result shows existence of linear and positive relationship between rainfall, independent and instrumental variables. The vector autoregressive models establish structural and dynamic interrelationship between the five climatic variables considered. The Granger Causality result shows unidirectional and bidirectional casual relationship exist between the climatic variables.

It is evident that there exist linear, casual, dynamic and equilibrium relationship between Nigerian climatic variables and the in-sample forecast values obtained shows a very similar pattern to the original fitted climatic variables.

Nonparametric Predictive Inference for Bivariate Copulas

Frank Coolen[1], Tahani Coolen-Maturi[2], Noryanti Muhammad[1]
[1]Durham University, Durham, UK, [2]Durham University Business School, Durham, UK

Many real-world problems of statistical inference involve bivariate data. This study presents a new semiparametric method for prediction of a future bivariate observation, by combining nonparametric predictive inference (NPI) for the marginals with a parametric or nonparametric copula to model and estimate dependence between two variables. We investigate the performance of this method via simulations, with particular attention to robustness with regard to the assumed copula.

NPI is a frequentist statistical framework for inference on a future observation based on past data observations. NPI uses lower and upper probabilities to quantify uncertainty based on only few modelling assumptions. In this research, we consider some classical bivariate one-parameter copulas, including Gaussian (Normal) copula, Clayton copula, Frank copula and Gumbel copula, and we also explore a nonparametric kernel smoothing copula. In the presented method, the imprecision in the marginals leads to robustness for predictions with regard which can counter the effect of a wrongly specified copula.

We discuss results from simulation studies to show how our method performs for different sample sizes. We also apply the method to data sets from the literature and briefly outline related challenges and opportunities for future research.

Advancing the methodology behind multilevel small area synthetic estimates: incorporating orthogonal polynomials into the estimation process

Graham Moon[1], Liz Twigg[2], Joanna Taylor[1]
[1]University of Southampton, Southampton, UK, [2]University of Portsmouth, Portsmouth, UK

Synthetic estimation has been used in recent years both within academia as well as the Office for National Statistics to provide small area data where direct estimates from social surveys are unavailable due to insufficient sample sizes. The multilevel synthetic estimation process takes the coefficients and residuals from a multilevel model to generate estimates for every locality, including those not covered by the base survey. In 'individual plus area' models, individual age is generally included as one of the independent variables. To date age has been commonly grouped into a finite number of categories. Previous studies have highlighted how the relationship between an individual's health status and their age is far from straightforward. Modelling age as a categorical variable may have oversimplified such relationships. The objective of this research is to extend the multilevel small area synthetic estimation methodology to test for quadratic and/or cubic relationships – in this instance with respect to age and health status – by employing orthogonal polynomials terms within a multilevel small area synthetic estimation framework. We demonstrate how this more complicated estimation process can be achieved using the multilevel modelling package MLwiN's customised predictions facility. Model diagnostic statistics confirm that employing quadratic and cubic relationships between age and health status represents a better fit of the data. The resulting synthetic estimates were compared against health data from the 2011 UK Census, with the Spearman's rank correlations being statistically significant at the 0.01 level.

MODELLING ROAD TRAFFIC CRASHES USING VARIANTS OF SPATIAL AUTOREGRESSIVE (SAR) MODELS

OLUSANYA OLUBUSOYE[1], GRACE KORTER[1,4], AFEES SALISU[1]
[1]University of Ibadan, Ibadan,Oyo State, Nigeria, [2]University of Ibadan, Ibadan,Oyo State, Nigeria, [3]University of Ibadan, Ibadan,Oyo State, Nigeria, [4]Federal Polytechnic, Offa,Kwara State, Nigeria

Road Traffic Crashes (RTC) are a global scourge characteristic of our technological era, whose list of victims insidiously grows longer day by day. The objective is to propose models that account for spatial effects (SE) to explain the dynamics of RTC. Thus, the spatial autoregressive (SAR) model, SAR model with SAR disturbances (SARAR), and SARAR model with additional endogenous variable (SARARIV) were estimated. To allow for comparability, the traditional classical linear regression model (TCLRM) that do not account for SE was estimated. The parameter estimates for the exogenous variables, that is, population; travel density; land area were positive, while, that of major road length was negative. The estimated SAR spatial dependence was 0.37; 1.37; 1.20 with p values equal 0.06; 0.00; 0.00 for the SAR, SARAR and SARARIV models respectively. This indicates RTC were clustered around spatial units rather than the expected random distribution. The estimated SAR spatial error dependence was -1.43; -1.18 with p values equal 0.07; 0.19 for the SARAR and SARARIV models respectively. This suggests, an exogenous shock to one spatial unit will cause moderate changes in the neighbourhood. The log - likelihood, Akaike Information Criterion and Schwarz Criterion indicated the proposed SAR model was a better fit in comparison to the TCLRM. The spectrum of analysis provides a means for linking RTC with neighbourhood characteristics. The framework should enable the orientation of safety and injury prevention policies.

Keywords: Generalized Spatial Two Stage Least Squares, Maximum Likelihood Estimate, Road Traffic Crashes, Spatial Effects, Spatial Modelling

Bayesian Analysis of Seemingly Unrelated Regression Model with Measurement Errors

Anoop Chaturvedi[1], Ravi Kant Saini[2], Shalabh Shalabh[2], Tripti Chitranshi[1]
[1]University of Allahabad, Allahabad, UP, India, [2]Indian Institute of Technology, Kanpur, India

The present paper considers Bayesian analysis of a Seemingly Unrelated Regression (SUR) model involving explanatory variable with measurement error. The conditional posterior densities for the parameters have been derived and Markov Chain Monte Carlo algorithm using Gibbs sampler has been implemented in a multiparametric setup to evaluate the marginal posterior densities and the Bayes predictors. In a uniparametric setup, the burn in sample length can be decided visually by observing the Markov Chain of the scalar parameter, which is not possible in multi parametric setup. We provide a solution to this problem of deciding a uniform burn in sample length in multi parametric setup length by plotting the Markov Chain for each scalar parameter involved in the vector and then choosing the maximum value of the burn-in sample length among the burn-in sample lengths of all the parameters. A simulation study has been carried out and its results have been presented. It is clearly revealed that as the measurement errors become high, the Bayesian methods absorbs the involved intricacies and the efficiency of estimators decreases as the measurement errors increase.

Interpreting p-values in the light of prior evidence: deepening understanding of statistical significance

Hilary Watt
Imperial College, London, UK

This article guides readers through skilful interpretation of p-values, making explicit two different approaches that can be used and the circumstances in which it is helpful to rely more strongly on each approach. Inspecting plots that represent significant and non-significant results can deepen understanding of the formal p-value definition. Significance tests and confidence intervals are defined based on what would happen if experiments and corresponding statistical analyses are repeated many times (i.e. they are based on the frequentist approach), even though few repetitions are undertaken in practice. The p-value is defined as the probability that our observed results, or more extreme, would occur in our study sample under the null hypothesis (i.e. assuming no association in the population from which samples are drawn). What we really want is probabilities for different strengths of association in the underlying population, given our observed data; it is tempting to (mis)interpret p-values in these terms. In order to find the latter probabilities for different strengths of association, results from our study data need to be combined with prior evidence for these different strengths of association (the Bayesian approach to statistics). Whilst we rarely do this in a formal mathematical way (because prior evidence is hard to quantify), such prior evidence is often informally incorporated into our interpretations of p-values in our choice of language. The implications for the consequences of undertaking (multiple) significance testing are assessed, with an encouragement to make explicit and sensible compromises, appropriate to the situation.

Determining the safety impact of changing street lighting

Paul Marchant[1,2]
[1]*Leeds Metropolitan University, Leeds, UK,* [2]*University of Leeds, Leeds, UK*

Examining the effect of changing street lighting, e.g. brightening & whitening or alternatively reducing it, on public safety in regard to road traffic accidents. The author's earlier work suggests that new Private Finance Initiative (PFI) lighting has not reduced crime by the 20% that was claimed it would. This work will examine the situation for road accidents, when thousands of lights have been changed in a rolled-out programme.

A multi-level analysis of road traffic accident data will be presented. The analysis, done at the small area level, utilises the times of lighting change & the times of accidents, whilst recognising that the accident risk can change over time in the absence of changing the lighting level.

This is work in progress. The author's previous results for road accidents using smaller data sets have not yet established a definite signal for either reduction or increase in the accident risk. The work is on-going with new, more extensive data.

The Disagreeable Behaviour of the Kappa Statistic

Laura Flight, Steven Julious
University of Sheffield, Sheffield, UK

Introduction: It is often of interest to measure the agreement between two raters when an outcome is ordinal. The Kappa statistic is a frequently used tool that measures agreement, however there are a number of limitations.

Methods: The work of Feinstein and Cicchetti is used to explain two ‘paradoxes' of the Kappa statistic [1]:
1. For high values of concordance low values of Kappa can be recorded.
2. Asymmetric, imperfectly imbalanced tables have a higher Kappa than perfectly imbalanced and symmetric tables.
Motivational and hypothetical examples are used to demonstrate how the Kappa can produce unreliable results.

Results: The Kappa is sensitive to the distribution of the marginal totals. In the motivational example the proportion of concordance is 0.682 suggesting good agreement, however the Kappa is very low (0.038). The two statistics result in different interpretations of the agreement in the data. The poor performance of the Kappa is a consequence of one category dominating in the 2x2 table.

Conclusion: Other statistics such as the proportion of concordance should be used to indicate how well the Kappa statistic represents agreement in the data and each Kappa should be considered and interpreted based on the context in hand.

Reference:
1. Feinstein AR and Cicchetti DV. High agreement but low Kappa: I. The Problems of Two Paradoxes. Journal of Clinical Epidemiology 1990; 43: 543-548.

What motivation drives Aikido training in Kangeiko?

Lu Zou
Queen Mary University of London, London, UK

The study was designed to evaluate the effect of motivation and/or amotivation on success within Aikido. All participants attended the annually one-week intensive training, Kangeiko, were invited to take part in a questionnaire based upon the sport motivation scale (SMS-II) about their reasons for participating in Aikido. SMS-II covers both extrinsic (external, introjected, integrated and identified regulations) and intrinsic aspects of motivation, as well as amotivation.

The response rate was 53% (17 out of 32). Each individual was asked to set themselves three goals for the week. Overall, their intrinsic motivation was high (Mean=5.92, SD=0.88) but was not correlated to Aikido experience (r=-0.05, p=0.850). Identified (r=0.32, p=0.205), introjected (r=0.25, p=0.205) and integrating (r=.41, p=0.101) were the next highest ranked motivations and showed positive correlation with aikido experience, while external (r=-0.27, p=0.292) and amotivation (r=-0.26, p=0.311) showed a negative correlation. Introjected motivation was shown to affect both completion (OR=1.95, CIs: 0.78-4.85, p=0.152) and training hours (B=3.94, SE=1.48, p=0.019). Identified motivation was selected by Mixed Effects Model for task performance (B=0.36, SE=0.14, p=0.025) and confidence (B=0.47, SE=0.16, p=0.011). Introjected motivation was chosen for task performance (B=-0.33, SE=0.16, p=0.063).

In general, people training Aikido are self-motivated. The guilt of avoidance was key to drive people training during Kangeiko. If an individual believed Aikido as a way to develop oneself, both their confidence and task performance improves. However, if motivation is secondary to guilt in not participating then task performance falters.

Forecasting seasonal data and nonparametric unit-root tests

Robert Kunst[1,2]
[1]*University of Vienna, Vienna, Austria,* [2]*Institute for Advanced Studies, Vienna, Austria*

Nonparametric unit-root tests are a useful addendum to the toolbox of time-series analysis. They tend to trade off power for enhanced robustness features. We consider a variant of the RURS (seasonal range unit roots) test statistic, a variant of the level-crossings count adapted to classes of seasonal patterns, and a combined test. These tests exploit two main characteristics of seasonal unit-root models, the range expansion typical of integrated processes and the low frequency of changes among main seasonal shapes.

In this contribution, the consequences of test-based decisions for predictions of seasonal time series are explored. It is of particular interest whether power gains relative to parametric tests are reflected in improved forecast accuracy. Apart from generating processes with seasonal unit roots and with deterministic seasonality, also processes with seasonal time deformation are considered. Our simulations demonstrate for which type of seasonal processes test-based decisions imply more accurate predictions.

The methods are applied to exemplary seasonal time series.

Don Bradman: sporting genius and statistical outlier?

Stephen Walters
University of Sheffield, Sheffield, UK

The test career batting average of, Australian Test cricketer, Sir Donald Bradman of 99.94 runs per innings is one of the most famous and iconic sporting performance statistics. Bradman only played 52 test matches, with 80 innings, over a 20 year career, from 1928 to 1948.

Although many more test matches are now played Bradman's performance still stands out. In the 137 years of test cricket no one has a batting average remotely close to Bradman's. His career batting average is almost 39 runs higher than the next batsman (who have batted more than 20 times); and Bradman scored a century every 2.75 innings.

But just how good was Bradman's sporting performance? How does his performance compare to his peers of other test cricket batsmen and other legendary figures in other sports?

This talk will investigate Bradman's performance and compare and contrast it with other cricketers and other sports using a variety of summary statistics including the number of standard deviations that they stand above the mean for their sport (the Z score statistic).

Bradman's batting average is between 6 and 7 standard deviations (or Z scores) away from the mean average for other test batsmen; no over batsmen has a Z score of 3 or more. Bradman could thus be regarded as is a statistical outlier.

This presentation will demonstrate that statistics show that no other sportsperson towers over an international sport to the extent that Bradman does test cricket.

Recruitment and retention of participants in publicly funded randomised controlled trials in the UK

Christopher Knox, Danny Hind, Steven Julious, Stephen Walters
University of Sheffield, Sheffield, UK

In 2012/13, the National Institute for Health Research (NIHR) funded £209 million of research grants. A substantial proportion of which were for Randomised Controlled Trials (RCTs). The NIHR has funded studies that, at the design and application stage, have been overly optimistic about the number of eligible patients, consent rates, recruitment rates and retention/attrition rates. A problem with publicly funded RCTs is that the recruitment of participants is often slower than expected, with many trials falling to reach their planned sample size within the originally envisaged trial timescale and trial funding envelope.

We describe an audit of the consent, recruitment and retention rates for single and multi-centre RCTs funded by the NIHR Health Technology Assessment (HTA) Programme.

Our audit of the reports of 99 RCTs, published between 2004 and 2013, found that: 51% (50/99) recruited to their target sample size; the median consent rate was 71% (Range 20 to 100%); the median recruitment rate was 0.8 participants per centre month (range 0.04 to 81); and the median retention rate was 90% (58 to 100%). The 80^{th} and 90^{th} percentiles for recruitment were 4.1 and 7.7 participants per centre per month respectively. So for most trial populations the recruitment rate is likely to be between 1 and 2 participants per centre per week. There is considerable variation in the consent, recruitment and retention rates in publicly funded RCTs. Investigators should bear this in mind at the planning stage of their study and not be overly optimistic about their recruitment projections.

The statistical interpretation of pilot trials: estimation instead of significance testing

Ellen Lee, Amy Whitehead, Richard Jacques, Steven Julious
School of Health and Related Research, University of Sheffield, Sheffield, UK

Objectives

In an evaluation of a new health technology, a pilot trial may be undertaken prior to a trial that makes a definitive assessment of benefit. The objective of a pilot trial is to provide sufficient evidence that a larger definitive trial can be undertaken and, at times, to provide a preliminary assessment of benefit. Pilot trials are usually underpowered to achieve statistical significance at the commonly used 5% level. However, despite recommendations that formal significance levels are not provided for pilot trials, many still quote and interpret P-values.

Methods

Using a worked example, we show how significance levels other than the traditional 5% should be considered to provide preliminary evidence for efficacy and how estimation and confidence intervals should be the focus to provide an estimated range of possible treatment effects. We also show how displaying confidence intervals of different widths alongside each other can help inform decision making and be used to illustrate the strength of preliminary evidence for efficacy.

Conclusions

We recommend that the analysis of pilot trials should focus on descriptive statistics and estimation, using confidence intervals rather than formal hypothesis testing. We suggest setting minimum prior requirements; that the mean treatment difference is above zero, and that a confidence interval of pre-defined length includes (or is above) the minimum clinically important difference.

Direct Risk Standardisation: A New Method for Comparing Casemix Adjusted Event Rates

Richard Jacques, Jon Nicholl, Michael Campbell
School of Health and Related Research, University of Sheffield, Sheffield, UK

Objectives

In all branches of the health and social sciences we need to be able to compare outcomes of groups of patients or people managed in different ways to understand the impact of different interventions, services and policies. Fair comparison of outcomes can be difficult to achieve because of differences in the characteristics of the patients and populations served. Distribution of these characteristics is known as casemix, and when casemix is associated with the outcomes, comparisons of outcomes are confounded with any differences in casemix. In theory this can problem can be solved by adjusting the comparison for casemix using standardisation. However, when the casemix adjustment models are complex, direct standardisation has been described as "practically impossible", and indirect standardisation may lead to unfair comparisons.

Methods

We propose an alternative method of standardisation where event rates are calculated in risk groups rather than casemix groups. The complex multidimensional casemix is converted into a simple one-dimensional risk distribution using a logistic regression model and then the events are directly standardised across the risk distribution. We illustrate this method using two existing models.

Conclusions

Direct risk standardisation is as straightforward as using conventional direct or indirect standardisation, enables fair comparisons of performance to be made, can use both continuous and categorical casemix covariates, and was found in our examples to have similar standard errors to the standardised mortality ratio. It should be preferred when there is a possibility that conventional direct or indirect standardisation will lead to unfair comparisons.

Multiple imputation in clinical trials: Should treatment group be included in the imputation equation?

Ellen Lee, Stephen Walters, Mike Bradburn
The University of Sheffield, Sheffield, UK

Background & Objectives: Missing data exists in almost every clinical trial and are almost unavoidable in research. Missing data presents statistical issues when estimating treatment effects. Multiple imputation (MI) is a common approach for estimating missing data in clinical trials. In multiple imputation, a variety of covariates are included in the imputation equation in order to predict the values of the missing data. It is recommended that all variables that will be in the subsequent analysis are included in the imputation equation. However, when testing for a treatment difference in clinical trials, is a (potentially) more conservative choice of excluding treatment group more appropriate?

Methods: We simulate clinical trial data to investigate the inclusion and exclusion of treatment group in multiple imputation equations. We compare the results against complete case analysis and the underlying treatment difference. A variety of scenarios are investigated, including when the data is missing completely at random, and when the response to treatment is different in the missing data group.

Results & Conclusions: We find that excluding treatment group from multiple imputation equations affects the point estimate and confidence interval of the treatment effect. The extent depends upon the pattern and amount of missing data. We recommend that the variables used in the MI equation are clearly defined and that results are interpreted in light of this. If treatment group is included in the imputation equation, an additional, more conservative, approach should be employed so that different deviations from the complete case assumptions are investigated.

Measuring Group Diversity with Incomplete Data

Jeremy Dawson
University of Sheffield, Sheffield, UK

Background
Organisational researchers are often interested in how the composition of work teams contributes to effectiveness. Frequently this involves studying the variability (diversity) of an attribute (e.g. age) as a predictor of performance. Such data are often gathered using surveys, with non-response a problem.

Objectives
To examine the accuracy of five commonly-used diversity indices (SD, variance, average absolute deviation (AD), coefficient of variation (CV), and rWG) with varying degrees of incomplete data, and determine how much missing data can be tolerated while maintaining sufficient accuracy.

Method
Simulations generated 4000 artificial teams with different distributions for each team size (N) between 3 and 12. For each team, 500 random samples of each size from 2 to N-1 were taken, and the standard error of indices across samples calculated to assess accuracy of individual estimates.

Results
The relative accuracy of sample diversity is approximately proportional to the adjusted finite population correction, (N-n)/Nn. The accuracy depends on the index being used: variance is the least accurate, followed by CV, SD, AD and rWG. For the SD, a value of (N-n)/Nn of 0.1 or lower is sufficient to generate a reasonably accurate estimate. Results differed little by the distribution of the original data.

Conclusions
The results suggest researchers should choose whether to include teams with incomplete data on the basis of the statistic (N-n)/Nn. However, these results only apply if response is random: if there is reason to believe probability of response might vary by the attribute being studied, complete data is important.

The Office for National Statistics Longitudinal Study

Nicky Rogers
Office for National Statistics, Fareham, UK

The ONS Longitudinal Study (LS) contains linked census and life event data for 1 per cent of the population of England and Wales. 2011 Census data have now been added to the study meaning that the LS now holds information on intention to stay, passports held, visitors, second addresses, main language and civil partnerships for the first time. It also means that for the second consecutive decade the LS will have information on general health, caring and religion, as well as linked data from five successive censuses that will provide supporting evidence for social and economic policy.

This poster will give an overview of the LS and key variables it contains and its potential use in research. Results will be presented from an exemplar research project that follows a cohort aged 15-24 in 1971 and examines their social and economic outcomes over the next four decades. The 2014 RSS conference presents a timely opportunity to highlight new data that are now available as the result of inclusion of 2011 Census data and will help researchers decide whether the LS is appropriate for their research.

Bayesian extreme value prediction using reference priors

Paul Northrop, Nicolas Attalides
University College London, London, UK

The Generalized Pareto (GP) and Generalized extreme value (GEV) distributions play important roles in extreme value analyses, as models for threshold excesses and block maxima respectively. We consider Bayesian inference for these distributions using ``reference'' prior distributions, specifically a Jeffreys' prior, the maximal data information (MDI) prior and independent uniform priors on separate model parameters.

Objectives

To investigate two important issues: whether these improper priors lead to proper posterior distributions, and, in the GP case, which of these priors should be used in order best to perform predictive inference about future extreme values.

Methods/models

To establish propriety (non-propriety) we use inequalities to bound above (below) the kernel of the posterior distribution by an integral that is finite (infinite). We carry out simulation studies to assess how well the largest value to be encountered over long future time periods is predicted under different priors.

Results and Conclusions

In the GP and GEV cases, the MDI prior, unless it is truncated suitably, never yields a proper posterior and that in the GEV case this also applies to the Jeffreys' prior. A sample size of three (four) is sufficient for independent uniform priors to yield a proper posterior distribution in the GP (GEV) case. In the GP case the uniform prior (and the Jeffreys' prior) lead to systematic over-predictions of future extreme values, and that the MDI prior leads to systematic under-predictions. We propose a simple family of priors that yields better predictions.

Random effects models for binary data: generation and estimation

Ilyas Bakbergenuly, Elena Kulinskaya
University of East Anglia, Norwich, UK

In meta-analysis of binary data arranged in K 2 x 2 contingency tables, the heterogeneity of the odds ratios across the studies is usually incorporated by standard (additive) random effects model (REM). An alternative, multiplicative random effects model is based on the concept of overdispersion. The multiplicative factor in this overdispersion random effects model(ODM) can be interpreted as an intra-class correlation parameter. In particular, this model arises when one or both binomial distributions in the 2 by 2 tables are changed to beta-binomial distributions, i.e. the probabilities of success are themselves beta-distributed. The Mantel-Haenzsel approach is conveniently extended to this setting. The estimation of the random effect parameter is based on profiling the modified Breslow-Day test. The coverage from the profiled Breslow-Day method is compared to standard methods through simulations. The results of the simulations show that the standard confidence intervals are biased for multiplicative model, and profiled Breslow-Day based confidence intervals work well in multiplicative model but are biased for the standard additive REM. Thus, misspecification of the REM in respect to the mechanism of its generation is an important issue and how to safeguard against such a misspecification is an open question.

Keywords: Intra-cluster correlation, odds ratio, random effects model, beta-binomial distribution, overdispersion, heterogeneity

Exact A- and D-Optimality Criteria for Partial Replication of the Central Composite Designs

Eugene Ukaegbu, Polycarp Chigbu
University of Nigeria, Nsukka, Nigera, Nsukka, Enugu State, Nigeria

Objectives

• To derive exact A- and D-optimality criteria for the central composite designs (CCD) when either the cube, star or both portions of the CCD are replicated for any number of centre points.
• To show that the A- and D-optimality criteria are functions of the information matrices of the various partially replicated central composite designs for any choice of , the axial distance in the design region of interest.

Method/Model
The second-order response surface model is used to study the central composite designs under the A- and D-optimality criteria. Analytical approaches are adopted to show that these optimality criteria are functions of the information matrices of the partially replicated designs. The MATLAB software is used extensively to conduct the tedious and rigorous matrix algebra.

Results and Conclusions
Analytical forms of the information matrices and inverse of the information matrices are developed for the three basic partial replications of the central composite designs considered in the study, namely, replicating only the cube, replicating only the star and replicating both the cube and the star. Exact D-efficiencies corresponding to the replicated options (cube, star or both) are developed from the analytical forms of the information matrices. Exact A-efficiencies are also developed from the analytical forms of the inverse of the information matrices for the partial replication of the cube, star and both cube and star. Numerical illustrations of the results are presented for the CCD in spherical region when the axial distance is , the practical alpha.

Prediction of equine fatal injuries in Thoroughbred flat racing in North America

Stamatis P. Georgopoulos, Matthew J. Denwood, Timothy D. H. Parkin
Boyd Orr Centre for Population and Ecosystem Health, School of Veterinary Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK

The aim of this paper is to develop and validate the performance of predictive models for fatal injuries in flat horse racing of Thoroughbreds in North America. Our analysis is based on data provided by The US Jockey Club, for the years 2010 to 2013.

Three different algorithms were used to develop predictive models for each year. Firstly, we used Multi-variable Logistic Regression, commonly used in risk factor analysis; secondly, Improved Balanced Random Forests, a machine learning algorithm based on a modification of the random forests algorithm and lastly, the Easy Ensemble algorithm, an ensemble of Classification And Regression Trees, optimized using Adaptive Boosting, trained on balanced random samples. Because fatalities are extremely rare events, less than 2 instances per 1000 starts on average, we dealt with the class-imbalance problem by using balanced samples.

The performance of each classifier was evaluated by calculating the Area Under the Receiver Operating Characteristic Curve, using the data available from the following year for validation. The Easy Ensemble algorithm produced marginally better classification results than the other algorithms. Results for each year range from 0.62 to 0.65.

This is the first study, to the best of our knowledge, to develop models that predict the probability of a horse sustaining a fatal injury prior to a race. The results could help identify horses at high risk on entering a race and inform the design and implementation of preventive measures aimed at minimising the number of Thoroughbreds sustaining fatal injuries during racing in North America.

On the Presentation of Results of Bayesian Inference on Statistical Models Containing Categorical Explanatory Variables.

Apostolos Gkatzionis
University of Warwick, Coventry, UK

We consider statistical models containing one or more categorical explanatory variables. Interest in such models lies in estimating contrasts among the category coefficients. However, in a Bayesian setting, the posterior density of contrasts is often unavailable in closed form: it is only possible to sample from it. And even in cases where the posterior can be calculated analytically, it may be too complicated to allow for reporting it fully, for example, in journal publications.

We introduce quasi-densities as a device for summarizing and presenting the posterior. Intuitively, the quasi-densities can be thought of as a set of univariate densities q_j that allow for approximating the posterior of any set of contrasts by considering the category coefficients as independent random variables, with the j-th coefficient having density q_j. We present ways of calculating accurate quasi-densities based on specific parametric assumptions and discuss ways of evaluating the error of approximation induced by using them.

Surrogate endpoint evaluation for ordinal outcomes: an information theory approach

Hannah Ensor, Cathie Sudlow, Martin Dennis, Catriona Graham, Christopher J. Weir
University of Edinburgh, Edinburgh, UK

Background

Using surrogates in place of primary endpoints can reduce the length, size and cost of clinical trials. Surrogates are biological measures taken at an early time point that can be used to predict treatment effect on an unmeasured later primary endpoint. The current foremost approach to surrogacy evaluation utilises information theory, assessing surrogacy through the measure R^2h.

Objectives

We aim (1) to extend the information theoretic methodology to incorporate a R^2h for assessment of a binary surrogate for an ordinal primary endpoint, (2) to evaluate its performance through simulation study and (3) to apply it to a case-study with binary and ordinal outcomes.

Methods

We use a multi-trial likelihood reduction factor configured on proportional odds models to estimate R^2h for a binary surrogate and ordinal true endpoint. We then investigate a case-study, the stroke trial CLOTs3, with 2876 patients enrolled in 94 centres. We consider whether a binary variable measuring the occurrence of deep vein thrombosis within 30 days is a surrogate for disability measured on the modified Rankin Scale at six months. In a simulation study we assess the impact of different numbers of patients and trials, the level of surrogacy and adherence to the proportional odds assumption.

Results and conclusions

The methodological development of this approach is an intuitive extension of existing theory; we foresee no issue in completion and implementation. The simulations and case-study will indicate whether this extension performs adequately under various real life scenarios.

Linear mixed effects models for the comparison of gastric emptying patterns in preterm infants

Anna Hepworth[1], Sharon Perrella[1], Karen Simmer[2,3], Peter Hartmann[1], Donna Geddes[1]
[1]School of Chemistry and Biochemistry, The University of Western Australia, Crawley, Western Australia, Australia, [2]Centre for Neonatal Research and Education, School of Paediatrics and Child Health, The University of Western Australia, Crawley, Western Australia, Australia, [3]Neonatology Clinical Care Unit, King Edward Memorial Hospital for Women & Princess Margaret Hospitals, Subiaco, Western Australia, Australia

The provision of adequate nutrition to hospitalised preterm infants is an ongoing challenge, often complicated by gastrointestinal disease and feeding intolerance. Gastric emptying studies provide insight into the appropriate management of feeding in this population.

While use of linear and non-linear mixed effects modelling is common in gastric emptying studies in other populations, the common approach in the preterm population has been to analyse summary statistics, such as the gastric half-emptying time. Among other obvious drawbacks, this approach increases the requisite number of participants, inappropriate and unethical in such a vulnerable population.

This study use linear mixed effects models to assess the response of the stable preterm infant to a) feeds of the same composition and volume, b) fortification of human milk for adequate infant nutrition and c) source of milk (mother's own milk or pasteurised donor human milk)

Application of linear mixed effects models to this data provided clear answers to the research questions using fewer participants than a power calculation for the 'standard' approach had determined. As for gastric emptying studies in the adult population, analysis that takes advantage of the full data set should be the standard.

Intention to treat verses completed treatment – modelling the effect of improved compliance

Anna Hepworth[1], Karen Simmer[2,3], ChooiHeen Kok[2,3], Kathryn Nancarrow[2,3], Donna Geddes[1]
[1]School of Chemistry and Biochemistry, The University of Western Australia, Crawley, Western Australia, Australia, [2]Centre for Neonatal Research and Education, School of Paediatrics and Child Health, The University of Western Australia, Crawley, Western Australia, Australia, [3]Neonatology Clinical Care Unit, King Edward Memorial Hospital for Women & Princess Margaret Hospital, Subiaco, Western Australia, Australia

While intention to treat analysis is the gold standard in randomised controlled trials, analysis of subsets such as those completing the treatment protocol may be important in order to answer questions relating to the benefit of improved compliance.

A recent study compared the maturation of preterm infants when expressed breastmilk and/or formula was fed using either a vacuum triggered teat (treatment group) or a standard constant flow teat (control). Three simultaneous analyses were run on the dataset from this project, being the intention to treat (n=97), a partial protocol subset (n=78, infants fed until discharge from the tertiary centre) and a complete protocol subset (n=67, infants fed according to protocol until discharge home). Due to the design constraint that twins were allocated to the same treatment arm, analysis used linear mixed effects modelling.

The smallest treatment effect sizes were seen in the intention to treat group, and the largest in the complete protocol group, with the difference being clinically relevant. Marginal significances of covariates was consistent across all three analyses for all outcome variables, increasing confidence in the findings.

Subgoup analysis allowed for the comparison of the ‘real world’ findings of the intention to treat analysis with the potential gains from improved compliance.

Comparing outcomes following kidney transplantation from donation after circulatory death and donation after brain stem death: A propensity score approach

Sally Rushton[1], Dave Collett[1], Paul White[2]
[1]NHS Blood and Transplant, Bristol, UK, [2]University of the West of England, Bristol, UK

Over the last decade, the UK kidney transplant programme has experienced changes in the population of individuals who donate their kidneys after death to patients with end-stage renal failure. Most strikingly, there has been a 20% decline in the traditionally preferred type of donor, those confirmed dead through brain stem death testing (DBD), and a 600% increase in the use of organs from circulatory death donors (DCD).

Potential differences between the outcomes of patients who receive DBD and DCD transplants have been a fixation of the international transplant community for the past few decades. A randomised controlled trial (RCT) to compare transplant outcomes across these two treatment groups is logistically and ethically impractical and so observational data from the UK Transplant Registry have been used to determine the extent of differences. Such data are subject to severe selection bias due to different practices across transplant teams. Propensity score (PS) methods, which seek to impose features of a RCT in observational studies of treatment effects, were applied in this context.

PS matching was used to impose homogeneity across numerous factors that affect outcome in an attempt to accurately compare the risk of transplant failure or patient death within three years following DBD and DCD kidney transplantation. The results of this analysis, compared with traditional regression adjustment, highlighted that if tissue type matching was as good as for DBD kidney transplants then the outcomes of DBD and DCD transplants would be more similar.

Estimating the prevalence of female genital mutilation in England and Wales

Alison Macfarlane[1], Efua Dorkenoo[2,1]
[1]*City University London, London, UK,* [2]*Equality Now, London, UK*

Background

It has been estimated that 66,000 mainly African women aged 15-49 resident in England and Wales in 2001 had undergone female genital mutilation (FGM) and 24,000 girls under the age of 15 largely from African communities were at risk of WHO Type III FGM, the severest form. These estimates have been widely quoted to demonstrate the need for appropriate care for affected women and to protect their daughters from FGM. Updated estimates are now being produced.

Objectives

To produce estimates of:

1. Numbers of women with FGM living in England and Wales as a whole and in each local authority area in 2011.
2. Numbers of women with FGM living in England and Wales and in each local authority area giving birth each year from 2000 to 2012
3. Numbers of daughters born to women born in FGM-practising countries resident in England and Wales and in each local authority area.

Methods

Data from surveys undertaken in the FGM practising countries are being used to derive proxy estimates of prevalence rates and applied to the numbers of women born in those countries who were enumerated in the 2011 census or who have registered births in England and Wales. In addition to country of birth and age group, the new estimates will draw other factors recorded in the census, including ethnicity, religion, first language and age on arrival in the UK.

Results and conclusions

The results will be presented and the limitations of the methods will be discussed.

Practical issues in ordinal response modelling

Altea Lorenzo-Arribas[1], Mark Brewer[1], Anke Fischer[2], Tony Craig[2], Anna Conniff[2]
[1]Biomathematics and Statistics Scotland, Scotland, UK, [2]The James Hutton Institute, Scotland, UK

When modelling ordinal response variables, a generic continuous approach is commonly used in some areas of research. There are good, published methodological reasons showing the inadequacy of this technique, but we want to take a different approach by providing positive advantages of an ordinal versus a continuous approach by way of an extensive simulation study. Assuming a latent variable approach for our ordered categories response variables, we have found that there is a greater consistency in the results between the ordinal and the underlying latent variable modelling compared to when the continuous model is considered. A variety of threshold patterns are considered to account for different types of response behaviours.

Next, we highlight some of the implications of the proportional odds assumption and more flexible approaches such as the partial proportional odds assumption.

Finally, we focus on the implementation of mixed models in this context and apply them to two case studies concerning perceptions of environmental matters. In both models the advantages of an ordered categories approach are shown.

This talk addresses both implementation and communication challenges in the development of these models, covering both our experiences and those of our non-statistical collaborators.

How big is big? Identifying which numbers in a series are significant, and applying them to migration.

Timothy Martyn Hill[1,2]
[1]*Liverpool Victoria, Bournemouth, UK,* [2]*ONS, Fareham, UK*

The ONS Data Visualisation department needed to display internal migration movements. There were too many movements to present easily on a map. So a way had to be found to decide which were significant. Using a paper from the seventies we built an algorithm to pick out the significant entries in a stream of numbers, and used that algorithm to state which internal migrations were significant. That same algorithm was later used outside the ONS for external migrations, and the results were used for an article in Significance (http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2014.00721.x/full). We present the derivation of the algorithm and its results for internal and external migration.

Calibration Approaches for Disease Models

Cathal Walsh, Arthur White, Susanne Schmitz
Trinity College Dublin, Dublin, Co. Dublin, Ireland

Mathematical models for disease are used to make predictions about the likely outcomes, such as morbidity and mortality, in a population, given different interventions, such as vaccination, screening and treatment. These are necessarily simplifications of the underlying mechanism of disease progression but have become essential tools in areas such as cost-effectiveness research. While many types of model can be used, in all cases it is of crucial importance to correctly calibrate the model parameters of the population, which are used for example to represent age of onset of illness, duration of particular health states, or disease specific mortality. In this work we examine and compare two approaches to model calibration for the modeling of infection with human papillomavirus (HPV) through to development of cervical cancer in affected women. Disease progression for an individual patient is characterized by a sequence of transitions between health states. The first approach which we examine is a likelihood based approach which uses a large, predefined collection of possible parameter values, and then calculates the goodness of fit of each of the sets to targets, selecting from these subsets of parameters that fit the data well. These plausible sets may then be used in later modeling. By contrast, the second approach employs a Metropolis-Hastings algorithm which searches through the joint parameter space until it converges in distribution around sets of values which may then be used for further modeling. The methods are described in detail and the results are compared.

An Application of Regression Modelling to the Incidence and Mortality Rates of Children from Tuberculosis in London and Glasgow at the Turn of the 20th Century

Stenford Ruvinga, Rosie O'Neil, Gordon Hunter
Kingston University, Kingston upon Thames, Surrey, UK

Tuberculosis (TB) is a significant cause of poor health and death across the World, but was a more serious disease before the development of antibiotics. In the 19th and early 20th centuries, it was endemic in underprivileged areas of cities such as London and Glasgow, where overcrowded living conditions and poor diet contributed to the disease spreading, with young children being particularly susceptible.

Recently, resources have become available regarding patients attending various hospitals prior to 1904. The Historical Hospital Admissions Records Project (HHARP) dataset provides detailed information on patients admitted to children's hospitals in London and Glasgow between 1890 and 1901.

Using this data, we have investigated child admissions for, and death from, TB over that period. Logistic regression indicates that, whilst TB was more prevalent in Glasgow than in London, children treated for the disease in Glasgow were more likely to survive. Younger children in both cities were more likely than older ones to be admitted to hospital with TB and to die from it. Although boys were more likely to be admitted with TB, in the early years of the period studied, they were more likely than girls to survive. However, the survival rate for both boys and girls rose over the period, before stabilising around 78% for both boys and girls by 1901.

This work illustrates the value of historical datasets for studying of development of public health in the U.K.. The methodology should be appropriate for the studies of other diseases, demographic groups and locations.

Evaluation of the Prediction Capabilities of Partially Replicated Orthogonal Central Composite Designs

Polycarp Chigbu, Eugene Ukaegbu
University of Nigeria, Nsukka, Nigeria

The stability and prediction capabilities of various replicated options of the orthogonal central composite designs (CCD) for 3 to 6 factors are evaluated. The bases of the evaluation are the fraction of design space (FDS) graphs with the D- and G-efficiencies. Also, the performances of the CCDs for the scaled and unscaled prediction variances for each of the replicated options and for each factor under consideration are compared.

Seven partially replicated options of the designs are simulated and compared using the D- and G-efficiencies as single value optimality criteria. The fraction of design space and their corresponding prediction variances are generated using Design Expert version 8. The MATLAB program is then used to plot the FDS graphs for the scaled and unscaled prediction variances (SPV and UPV).

Close examination of the FDS for the SPV and UPV show that the replicated-star options consistently lead to small prediction variance for k = 3, 4, 5 and 6 factors. The outstanding performances of the replicated star CCDs are made clearer by plotting the UPV where the design size is not constrained by the cost of experimentation and the replicated-star CCDs with the highest design runs among the other replicated-star options persistently yield minimum prediction variances. G-efficiency of the designs is improved by replicating the star with additional centre points, while replicating the cube does not improve the G-efficiency of the CCDs with orthogonal α in the spherical region.

Comparison of methods for imputing limited-range variables

Laura Rodwell[1,2], Katherine Lee[1,2], Helena Romaniuk[1,2], John Carlin[1,2]
[1]Murdoch Childrens Research Institute, Melbourne, Australia, [2]The University of Melbourne, Melbourne, Australia

Multiple imputation (MI) was developed to enable valid inferences in the presence of missing data rather than to re-create the missing values. Within the applied setting, it remains unclear how important it is that imputed values should be plausible. One variable for which MI may lead to implausible values is a limited-range variable, where imputed values may fall outside the observable range. The aim of this work was to compare methods for imputing limited-range variables.

We consider three variables, based on different scoring methods of the General Health Questionnaire (GHQ). These variables resulted in three continuous distributions with mild, moderate and severe positive skewness. In an otherwise complete dataset, we set 33% of the GHQ observations to missing at random; creating 1000 datasets with incomplete data.

We imputed values on the raw scale and following transformation using: regression with no rounding; post-imputation rounding; truncated normal regression; and predictive mean matching. We estimated the marginal mean of the GHQ and the association between the GHQ and a fully observed binary outcome, comparing the results with complete data statistics.

Imputation with no rounding performed well when applied to data on the raw scale. Post-imputation rounding and truncated normal regression produced higher marginal means for data with a moderate or severe skew. Predictive mean matching produced under-coverage of the complete data estimate. For the association, all methods produced similar estimates.

For highly skewed limited-range data, MI techniques that restrict the range of imputed values can result in biased estimates for the marginal mean.

Blogging in Biostatistics: impact metrics and social network analysis

Altea Lorenzo-Arribas[1], Pilar Cacheiro[2], Silvia Lladosa[3], Hèctor Perpiñán[4], Urko Aguirre[5]
[1]Biomathematics and Statistics Scotland, Scotland, UK, [2]Universidad de Santiago de Compostela, Santiago de Compostela, Spain, [3]Bayestats, S.L., Valencia, Spain, [4]Department of Statistics and OR, Universitat de València, Valencia, Spain, [5]Research Unit, Hospital Galdakao-Usansolo, Bilbao, Spain

The utility of blogs as information repositories has been boosted in recent times as they have been recognised and recorded in citation metrics tools. While the life sciences lead te way according to aggregators such as ResearchBlogging, statistics and more specifically biostatistics are also well represented in a wide variety of well-established online resources.

This poster reflects on the relevance of these online tools to the statistics community by means of an analytical exercise. The authors present temporal statistics and social media data, in order to create a holistic model representative of an active blog created by and for young researchers in Biostatistics.

Potential impact metrics are discussed and tested via FreshBiostats blog data and an exercise of social network analysis is implemented to demonstrate its reach over the course of 2 years of activity.

Beyond the personal benefits, in terms of networking opportunities, the blog has proven to be a learning tool for the authors themselves. The analysis also seems to indicate that it could also be playing a small role in the promotion of biostatistics, which is a great motivation to continue.

Multivariate and Multilevel Logistic Regression Analysis of Infant Mortality Rates across the Six Geopolitical Zones of Nigeria

Phillips Obasohan
College of Administrative and Business Studies, Niger State Polytechnic, Bida, Nigeria

At the turn of this millennium, World Health Organization and United Nations resolved as part of their commitments to tackle some developmental burdens of sub-Saharan African in what became known as Millennium Development Goals (MDGs) to be achieved before the end of 2015. Of these goals, 5 are either directly or indirectly health related. Specifically of note is the reduction of childhood mortality by 2/3 at the end of the time period. Researchers have in the past established the impact of individual-level characteristics on child survival; however, in this paper our focus will be to develop a predictive model and identify individual-level characteristics as well as the contextual impacts associated with infant mortality rates over the six geopolitical zones of Nigeria using multivariate and multilevel logistic regression analysis from the latest Nigerian Demographic and Health Survey (NDHS) with estimates of odds ratio and 95% confidence interval for the children nested in mothers and nested in six geopolitical zones. In view of the above aims, more specifically (i) we shall establish if Infant Mortality Rates (IMR) varies across the six geopolitical zones; (ii) assess the individual-level relationships between the six geopolitical zones and IMR and (iii) establish if contextual explanatory characteristics: (a) proportion of children immunized and (b) proportion of mothers who received any form of prenatal care, account for variations in IMR across the six geopolitical zone of Nigeria. The results will be discussed and appropriate suggestions made that will assist in policy formulation to improve infant and child survival.

Developing, validating and recalibrating a competing risks prognostic model for kidney failure patients.

Lucy Riley, M. John Bankart, Mark Lambie, Simon Davies
Keele University, Stoke on Trent, Staffordshire, UK

The objective of this study was to create a prognostic model that would predict the risk of developing a rare but serious complication (encapsulating peritoneal sclerosis or EPS) for peritoneal dialysis (PD) patients, accounting for the competing risk of death. The model would be used to inform patients and physicians of the risks of the complication in order to aid decisions regarding treatment switching.

This study utilises the Australian and New Zealand Dialysis and Transplant Registry of 16446 kidney failure patients. The Fine and Gray method was applied to model the 5 year risk of EPS, accounting for the competing risk of death, for patients after 3, 4 and 5 years of PD. A hierarchical backwards stepwise selection method was applied to determine which of the clinically relevant prognostic factors would be incorporated into the final prognostic model. AIC, BIC, discrimination and calibration measures were used to select the best model. The baseline cumulative sub-distribution hazard (SDH) was estimated using the Breslow method and a smoothing function constructed of restricted cubic splines was used to evaluate the hazard at 5 years.

Age^2 (SDHR 0.999, CI 0.998, 1.000) and primary renal diagnosis (SDHR 0.36, CI 0.19, 0.69) were found to be strong predictors of EPS.

The prognostic model was externally validated using the Scottish Renal Registry data. The prevalence of EPS was significantly different in this dataset, so the prognostic model was updated using recalibration-in-the-large and overall recalibration techniques.

Benzodiazepine drug dispensing and risk of abuse

Ingunn Fride Tvete[1], Trine Bjørner[2], Tor Skomedal[3]
[1]The Norwegian Computing Center, Oslo, Norway, [2]Department of General Practice/Family Medicine, Institute of Health and Society, University of Oslo, Oslo, Norway, [3]Department of Pharmacology, University of Oslo, Oslo, Norway

Objectives:
Benzodiazepines (BZDs) are tranquilizers that act on the central nervous system and produce sedation and muscle relaxation and lower anxiety levels. It is widely prescribed for a variety of conditions, particularly anxiety and insomnia. Guidelines state that these drugs should only be used for a short period of time, and concurrent use can give drug dependence.

Methods:
We consider BZD-naive individuals and their risk to become excessive BZD redeemers through a cox proportional hazard regression model where some of the explanatory variables varied over time. Data was obtained from the Norwegian drug prescription database, The Norwegian Labour and Welfare Service and Statistics Norway. We considered risk factors such as gender, age, which BZD previously dispensed, prescriber's specialty, previous medication use as indication of other relevant diseases, county, education, working situation and social benefits possibly received.

Results:
We find that only a small fraction of the previously BZD-naive individuals become excessive BZD redeemers. That is, most individuals redeem BZD according to the guidelines. The group becoming excessive redeemers differs to a large extent from the others with respect many of the risk factors mentioned above. Certain BZDs previously redeemed and certain specialties of prescribers gave a higher risk become excessive redeemers, together with having lower income and education and having previously received social benefits.

Modelling neonatal electroencephalogram signals in response to noxious and tactile stimuli

Simon Wallace[1], Sofia Olhede[1], Lorenzo Fabrizi[2], Maria Fitzgerald[2]
[1]University College London - Department of Statistical Science, London, UK, [2]University College London - Department of Neuroscience, Physiology and Pharmacology, London, UK

The number of premature births in the UK is increasing. A premature birth increases the probability of an infant requiring intensive care. The consequences of invasive intensive care procedures upon the development of the central nervous system is poorly understood but is correlated with long-term neurological problems. Here we focus upon the development of the response to touch and pain in prematurely born infants, measured by electroencephalography (EEG). We aim to construct a time series model that can describe most temporal variation observed in neonatal brain activity and from this infer characteristics and test scientific hypotheses.

Data was collected from inpatients at University College London Hospital (UCLH) at various stages of prematurity who underwent noxious skin lance as part of their clinical care and a recorded response to tactile stimuli. We present a stochastic model for the neonatal electroencephalogram (EEG) signals recorded from this sample. The model describes first and second order characteristics of these highly complex non-stationary time series; thus allowing us to describe brain development in response to such stimuli. Taking into account the evident long and short term dependence of the neonatal EEG signals we have constructed a relatively simple time homogeneous model for the baseline level of brain activity, to which we can include activity synonymous with spontaneous bursting and directly evoked by noxious and tactile stimuli.

ONS business survey returns: Winsorisation vs trimming for treating outliers

Laura Mulcahy
Office for National Statistics, Newport, UK

ONS business surveys provide essential information for monitoring the UK economy. The accuracy of this information depends on good quality responses to the surveys. If the responses are unusual, they can skew results and provide misinformation. Unusual responses can be errors, or just outliers in a distributional sense. After errors are corrected, there are several ways to deal with outliers, which include post-stratification (re-weighting them), trimming (deleting them) and Winsorisation (ameliorating them). Post-stratification and trimming can introduce a large amount of unseen bias into the estimates; therefore Winsorisation is often used as an alternative. Implementation of Winsorisation, in preference to trimming, depends on whether a small increase in bias is an acceptable trade-off over increasing the variance, resulting in a smaller mean squared error.

This poster will explain the methodology underlying Winsorisation (Winsor, et al, 1947) and compare with trimming in the context of an ONS business survey.

Normalization of plant gene expression data using support vector machine for identification of differentially expressed genes

Sandip Shil[1,2], Kishore K Das[2], Ananta Sarkar[3]
[1]Research Centre, Central Plantation Crops Research Institute, Guwahati, Assam, India, [2]Department of Statistics, Gauhati University, Guwahati, Assam, India, [3]Directorate of Research on Women in Agriculture, Bhubaneswa, Orissa, India

Background: Normalization of gene expression data refers the process of minimizing non biological variation in measured probe intensity levels so that biological differences in gene expression can be appropriately detected. Several normalization approaches have already been proposed, and most of which were formulated based on two channel gene expression datasets. Recently, use of non-linear methods has been proposed, which are assumed to be superior to linear methods. Furthermore, use of support vector machine (SVM) has been proposed to deal with non-linear problems in several studies. Therefore, we attempted to normalize our expression data using SVM based normalization methods, namely support vector regression (SVR) and support vector machine quantile regression (SVMQR). This paper deals with normalization of gene expression data using SVM based approaches for further application in plant gene expression datasets.

Methodology: SVR and SVMQR normalization methods have been implemented on simulated as well as bench mark datasets and their performance have been compared with respect to other standard normalization methods namely, locally weighted scatter plot smoothing and Kernel regression. Computation of variance within replicate groups, identification of differential genes based on the bench mark gene-lists were used as performance measures. Further, these methods were applied in coffee dataset to identify significantly expressed genes.

Results and Conclusion: It has been found that the normalized datasets based on proposed methods are capable of producing minimum variances within replicate groups and also able to detect truly expressible significant genes with respect to above mentioned other normalized datasets.

Estimating population characteristics of psychometric function using adaptive rules

Mark Edmondson-Jones[1,2], Abby McCormack[1,2], Heather Fortnum[1,2]
[1]NIHR Nottingham Hearing Biomedical Research Unit, Nottingham, UK, [2]Otology and Hearing group, Division of Clinical Neuroscience, School of Medicine, University of Nottingham, Nottingham, UK

Objectives

In psychoacoustics detection performance is typically modelled using a psychometric function, relating probability of correct identification to stimulus intensity. Such a model may have lower and upper asymptotes that differ from zero and one respectively due to e.g. guessing and attention lapses. Various methods of stimulus placement are used. Adaptive rules base stimulus level on prior performance and have been shown to produce biased estimates. Other methods, such as the method of constant stimuli (MOCS) require many trials and can induce non-stationarity in participant behaviour due to learning or fatigue. We explore the use of population-based estimates of psychometric characteristics from adaptive trials using hierarchical methods, presenting results from a motivating example using data from the UK Biobank study.

Method/Models

Simulations are used to illustrate the characteristics of psychometric parameter estimates with maximum likelihood and Bayesian methods using adaptive placement. We then illustrate the use of a Bayesian hierarchical approach to model population psychometric characteristics. Finally, we illustrate this method on Speech-in-Noise data from the UK Biobank study.

UK Biobank is a large cohort study including a SiN hearing test which involving listening to spoken digit triplets played against a noisy background and keying them via a touchscreen. The spoken digits' volume was varied adaptively according to the participants' performance. Fifteen triplets were presented to each ear. The test was repeated in a subset of participants after 3-4 years.

Results/Conclusions

We illustrate the method using simulations and results from UK Biobank.

Modelling Chronic Disease Progression in Longitudinal (General Practice) records.

Rosemary O'Neil, Peter Soan, Sybel Rames
Kingston University London, London, UK

The utilisation of routinely collected health data is an area of increasing interest and research activity but poses some complexities in the management and analysis of such datasets. A challenge for the statistical community lies in adapting existing and developing new techniques for analysing such large scale longitudinal data sources. Our aim is to develop and apply emerging methodologies to explore the scope and potential for investigating multiple aspects of progressive chronic diseases treated within the primary care setting. We use patient records from a large sample of General Practices in England and Wales, to study the natural history of Chronic Kidney Disease (CKD) but the methods will be applicable to any progressive chronic condition.

The research presented here focuses on investigating 'rapid decline' within patients diagnosed with CKD; and more specifically on determining if early morbidity indicators can predict which patients are at risk of rapid decline. Prevalence of CKD within our dataset reflects that seen nationally, as does the proportion of these experiencing rapid decline, approximately 10% based on regular eGFR readings. Early results indicate that there are strong links between rapid decline and some co-morbidities and other epidemiological factors.

Our analyses aim to evaluate the impact of contributing factors such as co-morbidities, treatment regimes and epidemiological factors on the progression of CKD within fixed time frames. Statistical methodologies include structural equation modelling (SEM) and causal pathway analysis both of which are being routinely used in the growing field of statistical analysis of very large datasets.

How much confidence should we have in the confidence interval for a proportion?

Sam Ellis[1], Mike Hicks[1], Robert Langford[2]
[1]Defence Ordnance Safety Group, Defence Equipment & Support, Bristol, UK, [2]Engineering Analysis, Munitions, BAE Systems, Glascoed, UK

There's the Wald, the Wilson score, the modified Wilson, the Sterne, the Crow, the Blyth and Still, the Agresti-Coull, the Bayesian with Jeffery's prior, ... the longest, the shortest, the highest density, the likelihood ratio, ... the exact Clopper Pearson and at least three other's calling themselves exact, though Neyman noted that "exact probability statements are impossible in the case of the Binomial Distribution". There's the problem of coverage, oscillations and Edgeworth expansions, those that guarantee an interval of 1-alpha There are refinements for each method, there are the rules that npq = 5 or 10 or n is large. Should we use a continuity correction or a finite population correction? What do we do when p is close to the boundaries 0 and 1? What confidence level should we choose?
If increasing the sample size leads to a better confidence interval does the increasing number of publications on the subject of binomial confidence intervals lead to a practitioner's decreasing confidence that they have chosen the right method to calculate an interval? Is the publication rate a sign that the current methods are invalid or is it an indication that there is a desperate need to get more publications? Should the RSS provide some closure and set a standard for the practitioner in regards to the confidence interval for a single proportion? This presentation/poster will describe the background, the history including publication rates, and use a comparison of some of the methods to explore these issues.

Optimising options and strategies for living donor kidney transplantation for incompatible donor-patient pairs

Matthew Robb, Rachel Johnson
NHS Blood and Transplant, Bristol, UK

Patients in need of a kidney transplant typically have to wait for a deceased donor and face an average waiting time of 3 years. If the patient has a willing living donor this dramatically reduces waiting time. However, due to biological incompatibility, having such a donor does not guarantee a transplant. Paired exchange allows patients needing a kidney transplant, but with an incompatible donor, another route to transplantation. Here, one incompatible donor-patient pair exchanges kidneys with another donor-patient pair, enabling both patients to have a living donor transplant. Unfortunately, some pairs find little improvement in waiting time due to their blood group and tissue type characteristics.

Using simulations that imitate national ‘matching runs’ with data generated from previous entrants into the scheme, we attempt to quantify the waiting time of a recipient given their blood group, level of antibodies and their donor’s characteristics. This time can be directly compared to time on the deceased donor waiting list and chances of other potential routes to transplant such as antibody incompatible transplants. Transplant survival times are compared using Cox models.

Combining this information, we can build a complete picture of a patient’s chance of transplant given the characteristics of both the patient and the donor. Early results indicate that waiting time is highly dependent on patient and donor characteristics, while transplant type also affects transplant survival. Factors such as these must be taken into account for the patient to receive suitable guidance on the options available.

Improving the Segmentation of Consumer Databases

Steve Wisher, Kate Marshall
Information by Design, Hull, UK

Objectives

Consumer goods companies commonly create large databases of consumers which they use for marketing their brands. These databases usually include a small number of variables which are used to create segmentation models which can then be applied to the database. Different segments of consumers are then targeted with different marketing communications and offers with the aim of increasing brand purchase and loyalty. This presentation examines a number of the issues faced in developing these segmentation models and suggests approaches that can be used to improve their reliability.

Method/Models

The suitability of some of the measures used to segment databases and the accuracy of the database data itself are examined using data from a number of UK and US studies. Database data is compared with data collected using consumer panels.

Results and Conclusions

Database data on brand purchase or consumption which is collected directly from consumers is generally affected by levels of under or over-claim. Under and over-claim are time dependent, age related also correlated with the level of purchase or consumption by the consumer. Appropriate measures of brand loyalty or brand affinity are generally more accurately recorded on consumer databases and effective in developing segmentation models.

The impact of including incorrectly analysed time to event data in meta-analysis

Kerry Dwan, Marty Richardson, Richard Jackson, Catrin Tudur-Smith
The University of Liverpool, Liverpool, UK

Background: Time to event data are often analysed using the Cox Proportional Hazards (Cox PH) model in order to obtain an estimate of the Hazard Ratio. Methods to check the suitability of the underlying assumption of proportional hazards include visual inspection of the Kaplan-Meier curves, Schoenfeld residual plots and incorporation of time-dependent covariate effects. Although the Cox PH model is fairly robust to deviations from this assumption, the cumulative impact of combining multiple studies that have incorrectly analysed the data in this way is unclear.

Objectives: To assess the impact of combining trials that analyse time to event data using a Cox PH model when underlying hazards are in fact non-proportional, on meta-analysis results.

Methods: Cochrane reviews and Health Technology Assessments published in 2013 were accessed. Reviews with time to event outcomes were included. The methodology of the review was checked to see whether the assumption of non-proportional hazards had been considered and, if so, how it was dealt with in the review. The included trial reports were obtained and the data analysis section was checked to see whether Cox PH was used and which, if any, methods were used to check the underlying assumptions of this approach. Any Kaplan Meir plots included in the trial report were also checked to see whether the curves crossed.

Survival time data were simulated for trials with varying degrees of non-proportional hazards. The impact of combining these studies within a meta-analysis was considered across a variety of scenarios.

Results: Results will be presented at The Royal Statistical Society Conference

A geometric treatment of overdispersion in binomial regression models

Karim Anaya-Izquierdo[1], Paul Marriott[2]
[1]*University of Bath, Bath, UK,* [2]*University of Waterloo, Waterloo, Canada*

There are different ways of dealing with overdispersion relative to the binomial distribution. First of all we will show how these apparently unrelated different approaches can be encompassed in a geometric framework. We will discuss the benefits of bringing in a geometric framework to the the overdispersion problem and the corresponding data analysis. We will illustrate these ideas with a couple of examples that make use of real data sets.

On development of Hyperbolic nonlinear growth models for application in Life Sciences and Environmental Studies

Oluwafemi Oyamakin[1], Angela Chukwu[0]
[1]*University of Ibadan, Ibadan, Oyo, Nigeria,* [2]*University of Ibadan, Ibadan, Oyo, Nigeria*

This paper proposed a new set of nonlinear growth models which is a modification of existing growth equation by introducing a scaling parameter theta conditioned on a hyperbolic sine function. The performance of the proposed models was compared with the classical growth models like exponential, monomolecular, gompertz, Richards and Von Bertallanffy The results indicated that the performance of the proposed growth models are best in terms of goodness of fit criteria (R^2, Adj R^2, MSE, AIC, & BIC) and hence, described the data most appropriately.

What is Normal? A Meta-Analysis of Phase 1 Placebo Data

Simon Kirby[1], William Denney[2]
[1]*Pfizer, Cambridge, UK,* [2]*Pfizer, Cambridge, USA*

Objectives: To summarise all adverse events (AE), vital signs, electrocardiograms (ECG), and lab measurements for healthy subjects receiving placebo in First in Human (FIH) and Multiple Ascending Dose (MAD) studies in Pfizer's Phase 1 Management System (PIMS) to aid in the interpretation of 'What is Normal?' and to provide informative prior distributions for Bayesian analyses.

Methods: All AE, vital sign, ECG and lab measurement data for healthy subjects receiving placebo in FIH and MAD studies were selected from PIMS. AEs were summarised by numbers and percentages of events, subjects, and studies with events. For vital signs, ECGs, and lab measurements, baseline, raw values, and change from baseline were summarised using distribution quantiles, histograms, and empirical distribution functions. Any numerical measurements with at least 100 subjects were modelled with a linear mixed effect model testing demographic parameters as fixed effects and random effects on intercept by study and subject within study using the lme4 function in R.

Results: The final data summarised were for 1204 subjects from 82 FIH and MAD studies. Updated ranges for extreme values of labs, vitals, and ECG measurements have been generated, and the importance of demographic parameters on measurements (or lack thereof) has been estimated with many subjects and dense measurements. The results were summarized and posted to an internal website allowing rapid queries without requiring specialized tools.

Conclusion: The analysis has allowed classification of potentially abnormal measurements incorporating the large data set of placebo subjects in similar populations.

Statistical methods to manage treatment non-compliance in RCTs with time-to-event outcomes (VenUS IV trial)

Caroline Fairhurst
York Trials Unit, University of York, York, UK

Objectives. To investigate analytical methods for dealing with treatment non-compliance in randomised controlled trials with time-to-event outcomes. Various techniques will be applied to data from the VenUS IV trial, which investigated the effectiveness of the four layer bandage and two layer compression hosiery in the treatment of venous leg ulcers. The primary outcome was time to healing of the ulcer.

Methods. The primary trial analysis of VenUS IV was conducted using the principle of intention-to-treat. The effect size derived from this analysis will be compared with those produced by appropriate application of other methods including: per protocol, complier average causal effect (CACE), a cox proportional hazards (CPH) model with treatment as a time-varying covariate and adjusted hazard ratio methods. A small simulation study will be conducted in addition.

Results. In an ITT analysis, using a CPH model adjusting for baseline ulcer area, ulcer duration and participant mobility with centre as a random effect, no evidence of a difference between the two treatment arms in terms of time to ulcer healing was found (HR 0.99, 95% CI 0.79 to 1.25, p=0.96). This work will form my dissertation for the MSc in Statistics with Medical Applications at the University of Sheffield, due to be completed this year. The dissertation will be submitted in mid-September and the bulk of the analysis will be conducted over Summer 2014. Therefore, results are not currently available but will be presented at the conference.

Dietary patterns amongst the United Kingdom adult population. A cross-sectional analysis of data from the National Diet and Nutrition Survey.

Benjamin Kearns[1], Katharine Roberts[1], Janet Cade[2], Michelle Holdsworth[1]
[1]The University of Sheffield, Sheffield, South Yorkshire, UK, [2]The University of Leeds, Leeds, West Yorkshire, UK

Objectives
Measuring and monitoring dietary intake is challenging. The primary objective of this work was to identify food groups that can help to explain the variance between different dietary patterns amongst the general population. The secondary objective was to examine inequalities in diet by different sub-groups (such as by gender and socio-economic status).

Methods
Data from the National Diet and Nutrition Survey 2008 to 2011 were used. This is a continuous cross-sectional survey about food consumption amongst the general population. Principal components analysis was performed to identify any potential dietary patterns. Associations between these patterns and sample characteristics (age, gender, ethnicity, body mass index, smoking status, socio-economic classification) were examined using linear regression.

Results
Data were available for 1,491 adults (aged 19 or over). Five dietary patterns were identified, and subjectively labelled as ‘healthier choices', ‘unhealthier choices', ‘mixed snacks and fast foods', ‘high-alcohol', and ‘meat and two vegetables'. There were differences in dietary patterns for each of the sample characteristics. For example, compared to people of White ethnicity, respondents of non-White ethnicity were more likely to report a ‘healthy' diet, with high scores on the ‘healthy choices' group and low scores on the ‘unhealthy choices' group.

Conclusions
Distinct dietary patterns, and inequalities within these patterns, were identified. This suggests that there are foods or food groups that may be suitable for use as indicators of healthy dietary patterns for population level surveillance in the UK. Further work is required to clarify whether these patterns are generalisable to other settings.

Patient characteristics associated with survival of 60+ year olds with a history of myocardial infarction using the UK THIN Data

L.A. Gitsels, E. Kulinskaya, N. Steel
University of East Anglia, Norwich, UK

Objective
Our aim was to identify patient characteristics associated with survival of individuals aged 60+ who have suffered a myocardial infarction (MI). Identification of higher-risk individuals is relevant to further medical management and calibration of enhanced annuities.

Methods
Data on four cohorts of individuals aged 60, 65, 70, and 75 (N=9,752; 30,500; 56,096; and 59,360) were obtained from THIN Primary Care database. Patients with a history of MI were matched with controls on sex, year of birth, and GP practice (matching ratio 1:3). We fitted a Cox's proportional hazards regression with a shared frailty term on GP practice. Predictors included sex, year of birth, socio-economic status, MI, heart and kidney disease (CAD, CVS, CKD), diabetes, hypertension, hypercholesterolemia, heart surgery, cardiovascular and lipid-regulating drugs, BMI, alcohol and smoking status, and their second-order interactions. The final models were obtained through backward elimination.

Results
The hazardous effects were: male, MI, CAD, CVS, CKD, diabetes, hypertension, smoking, cardiovascular drugs, and obesity up to age 70. The protective effects were: pre-hypertension, hypercholesterolemia, lipid-regulating drugs, overweight, obesity at age 75, alcohol, and later year of birth. The hazard ratio (HR) of timing of MI (≤5, 5-10, >10 years ago) ranged from 1.33 to 2.15, with an average of 1.55. There was no overall trend in timing of MI on survival across the cohorts.

Discussion
The models provide accurate estimates of HR for 60+ year olds. There is 68-72% concordance and a shrinkage slope of 2-4%. The results generally agree with previous research that was not age-specific.

Random effect meta-analysis of individual patient time-to-event outcomes

Wirda Nisar, Catrin Tudur-Smith, Ruwanthi Kolamunnage-Dona
University of Liverpool, Liverpool, UK

Objectives: Random-effect meta-analysis is considered as a powerful tool for investigating the possible sources of heterogeneity caused by unmeasured covariates. The meta-analysis of IPD can be performed by employing either a one-stage or a two-stage approach. We first investigate the current methods. We explore the performance of the one-stage and two-stage methods under different conditions and compare methods for estimating the parameters in a random effect model for time-to-event data.

Methods: Articles are searched and retrieved from the database of MEDLINE (Ovid) using keywords random effect, survival data, survival outcome time-to-event, cluster, frailty and multi centre effect.

We apply the common one-stage (Vaida, F. and R. Xu 2000, Ripatti, S. and J. Palmgren 2000, Abrahantes, J.C. and T. Burzykowski 2005, Tudur-smith et al 2005, Bowden et al 2011, Ha et al 2012, Simmonds et al 2013) and two-stage (log-rank test,Cox model) of IPD meta-analysis using 5 randomised control trials investigating the use of two anti-epileptic drugs: Carbamazepine (CBZ) and Sodium Valproate (SV).

Results: After analysing full text we have got 40 eligible studies,where studies 8 were based on parametric, 26 semi-parametric and 5 on non-parametric methods. We applied 7 one-stage and 2 two-stage methods for the estimation of the random effects models. Similar estimates were obtained for logHR and its standard error but some of the methods underestimate between trial variance parameter.

Conclusion: There are many alternative approaches for random effect meta-analysis of IPD time-to-event outcomes, and results can vary.Inferences based on the performance of the methods are proposed.

Reporting the use of Statistical Regression Models in Economic Evaluations - a Review and Good Practice Guidelines.

Benjamin Kearns[1], Roberta Ara[1], Allan Wailoo[1], Andrea Manca[2], Monica Hernández Alava[1], Keith Abrams[3], Mike Campbell[1]
[1]The University of Sheffield, Sheffield, UK, [2]The University of York, York, UK, [3]The University of Leicester, Leicester, UK

Objectives
Statistical regression analyses can play a pivotal role within economic evaluations of medicines and treatments. However, there is little guidance on how such analyses should be reported. This work had two objectives. The first was to perform a review to identify the frequency of regression models in economic evaluations, their uses, and the level of reporting detail. The second objective was to form good practice guidance, with the intention of improving practice in this area.

Methods.
The review concentrated on a random sample of economic evaluations submitted to the UK National Institute for Health and Care Excellence as part of its technology appraisal process. The results of the review were discussed by an expert working group, which led to recommendations for good practice along with a checklist for critiquing reporting standards.

Results.
The review showed that statistical regression models were used in the majority of submissions (59%, n = 79). However, there was limited reporting of fundamental information such as the sample size used and measures of uncertainty. A total of 27 recommendations covering pre-modelling considerations, model building, reporting and validation, and acknowledging uncertainty were formed and summarised in a checklist.

Conclusions.
Statistical regression models are in widespread use in economic evaluations yet reporting standards relating to basic information are poor. Increasing levels of reporting transparency is important as it will lead to increased confidence in the results of the regression analyses. The recommendations and checklist may be used by both those conducting regression analyses and those critiquing them.

Higher education: who goes, who doesn't

Leyla Bagherli
The Higher Education Funding Council for England, Bristol, UK

Participation in English higher education has risen over the last decade, and this has generally been the case for areas across the country. However, the growth has not been uniform and there is significant geographical variation.

This work explores how much of this geographical variation can and cannot be explained quantitatively, through multilevel logistic regression modelling. This modelling is based on detailed linked data that tracks pupils through from school level to whether they progress into higher education. School attainment is shown to be a key predictor of higher education participation in the modelling.

The results of this work highlight those areas with rates of young participation that are higher or lower than can be explained by the data and modelling. These results have been used to produce a set of colour coded maps for a visually clear representation of the variation across England.

Prediction of the People lost follow up on Antiretroviral Therapy (ART) Services in Nepal: A Statistical Modelling

Brijesh Sathian
Manipal College of Medical Sciences, Pokhara, Kaski, Nepal

Background

The real state about the spread of the HIV epidemic in Nepal is not clear since the details available are on the basis of repeated integrated biological and behavioural surveillance.

Objective

The objective of the study is to extract as much as information possible from available data and find out the trends of People lost follow up on ART in future.

Material and methods

A retrospective study was carried out on the data collected from the Health ministry records of Nepal, between 2006 and 2012. Descriptive statistics and statistical modelling were used for the analysis and forecasting of data.

Results

Including the constant term from the equation, the quadratic model was the best fit, for the forecasting of People lost follow up on ART. Using quadratic equation, it is estimated that 4331 reported number of People lost follow up on ART will be there in Nepal by the year 2020.

Conclusion

The People lost follow up on ART in Nepal are having an increasing trend. Estimates of the total number of People lost follow up on ART attributable to the major routes of infection make an important contribution to public health policy. They can be used for the planning of healthcare services and for contributing to estimates of the future numbers with People lost follow up on ART used for planning health promotion programmes.

Developing Robust Scoring Methods for use in Child Assessment Tools.

Phillip Gichuru[1], Gillian Lancaster[1], Andrew Titman[1], Melissa Gladstone[2]
[1]Mathematics and Statistics Department, Lancaster University, Lancaster, UK, [2]Institute of Child Health, Royal Liverpool Childrens' NHS Trust, Liverpool, UK

Background: Earlier and more sensitive diagnosis of disability reduces its detrimental effect on children and their family.

Objective: To develop more robust and highly sensitive scoring methods for Child Assessment tools which will ensure a more timely intervention of detected delayed development.

Methods: Using data from 1,446 normal children from the recent Malawi Development Assessment Tool (MDAT) study, we review and extend classical total scoring methods including simple scoring, Log Age Ratio methods and Item Response Models under different assumptions to derive normative scores in this child development context using binary responses in the Gross motor (GM) domain only.

Results: The weighting of simple scores is important as a lack of a response to all items does not necessarily imply a lack of ability. Further, smoothing of score values is beneficial when variability in certain age groups is high due to recruitment problems. The more complex methods accounting for most study design issues produce more reliable and more generalizable normative scores while correcting for the age variable. The sensitivity analysis showed that simple methods perform well in ideal situations.

Key words: binary data; disability; development assessment; scoring; item response theory.

Estimating the burden of childhood tuberculosis in the twenty-two high burden countries: a mathematical modelling study

Peter Dodd[1], Elizabeth Gardiner[2], Renia Coghlan[3], James Seddon[4]
[1]University of Sheffield, Sheffield, UK, [2]Global Alliance for TB Drug Development, New York, USA, [3]TESS Development Advisors, Geneva, Switzerland, [4]Imperial College London, London, UK

Diagnosis of tuberculosis (TB) in children (<15 years) is challenging and the extent of under-reporting is unknown. Direct incidence estimates are lacking and current burden estimates start from paediatric notifications.

Within a mechanistic mathematical model, we combined estimates of adult TB prevalence with evidence from the natural history of paediatric TB to estimate the incidence of TB infection and disease among children TB in the 22 high TB burden countries (HBCs). The effects of age, BCG vaccination and HIV infection were included. Sensitivity to assuming variation in BCG efficacy by latitude was explored.

We estimated that in the 22 HBCs 7,591,759 [IQR: 5,800,053 - 9,969,780] children became infected with Mycobacterium tuberculosis (Mtb) and 650,977 [IQR: 424,871 - 983,118] developed disease in 2010. Cumulative exposure meant 53,234,854 [IQR: 41,111,669 - 68,959,804] children harboured latent Mtb infection. The proportion of TB burden in children for each country correlated with incidence, varying between 4% and 21%. The overall paediatric case detection rate was 35% [IQR: 23% - 54%] in the 15 HBCs reporting paediatric notifications. 27% of HBC paediatric cases were predicted to occur in India. Disease incidence estimates were 27% lower if BCG efficacy was constant by latitude.

Paediatric TB notifications are lower than incidences estimated by our model, particularly in younger children. However the extent is highly variable between countries. Estimates of current household exposure and cumulative infection suggest an enormous opportunity for preventive therapy.

Mexican adult cancer patients´supportive care needs: validation of the Mexican version of the Short-Form Supportive Care Needs questionnaire (SCNS-SFM).

Svetlana Doubova[1], Rebeca Aguirre-Hernandez[1], Marcos Guitérrez De la Barrera[1], Claudia Infante Castañeda[1], Ricardo Perez Cuevas[1]
[1]Instituto Mexicano del Seguro Social, Mexico, D.F., Mexico, [2]Universidad Nacional Autonoma de Mexico, Mexico, D.F., Mexico, [3]Instituto Mexicano del Seguro Social, Mexico, D.F., Mexico, [4]Universidad Nacional Autonoma de Mexico, Mexico, D.F., Mexico, [5]Banco Interamericano de Desarrollo, Mexico, D.F., Mexico

Objective. To validate Mexican version of the Short-Form Supportive Care Needs survey (SCNS-SF-MX).**Methods.** A cross-sectional survey was conducted from June to December 2013. The study included 825 subsequent cancer patients older than 20 years with all forms of solid cancer who had prior surgical removal of cancer with histological confirmation, and attended to the outpatient consultations at a tertiary care Oncology Hospital in the Mexican Institute of Social Security. The validation of SCNS-SF-MX included: 1) content validity thorough a group of experts; 2) construct validity through an exploratory factor analysis based on the polychoric correlation matrix; 3) internal consistency by using Cronbach´s alpha; 4) Convergent validity between SCNS-SF-MX and quality of life, anxiety and depression scales by calculating Pearson´s correlation coefficient; 5) discriminant validity through analysis of MANOVAs and 6) test-retest reliability through intraclass correlation coefficient calculating. **Results.** SCNS-SF-MX has 33 items with 5 factors that account for 59% of score variance. The Cronbach's alpha values ranged from 0.78 to 0.90 among factors. SCNS-SF-MX poses good convergent validity as compared with quality of life and depression and anxiety scales; and good discriminant validity, revealing great information, psychological, support and physical-daily living needs for women, patients younger than 60 years; high physical-daily living needs for those with less than one year since cancer diagnosis, with advanced disease stages and current chemotherapy or radiotherapy. The intraclass correlation coefficient between two SCNS-SF-MX measurements was 0.9. **Conclusion.** SCNS-SF-MX has acceptable psychometric properties and is suitable to evaluate cancer patients´ supportive care needs.

Reporting of harms data in orlistat trials: A comparison between clinical study reports and journal publications.

Alex Hodkinson, Carrol Gamble, Catrin Tudur-Smith
Liverpool University, Department of Biostatistics, Liverpool, Merseyside, UK

Objective To determine whether harms data published in journal articles is consistent with the data presented in corresponding Clinical Study Reports (CSRs) using a case study of orlistat trials in obesity research.

Methods Publications related to clinical trials of orlistat were identified through comprehensive literature searches. A request was made to Roche for CSRs related to the identified orlistat trials. We compared the reported adverse events (AEs) and serious adverse events (SAEs) between both report types. Event specific numerical data were pooled across CSRs using meta-analysis and compared to corresponding analysis based on data from published journal articles. The structured reporting of harms were assessed against (CONSORT-harms) criteria and compared between two document types.

Results Journal publications with matching CSR were available for five trials. Reporting of AEs in journal publications was poor, with three reporting only a total 5%, 4% and 0% events. The corresponding CSRs report with high consistency 95%, 100% and 100%. Reporting of SAEs was similar with no events reported in the same three trial journals that were in the CSR. The structured reporting of harms criteria proved more reliable in the CSRs, with higher numbers of items satisfied. Where criteria were satisfied in both report types the CSR was more detailed. Meta-analysis of harms data from CSRs compared to journal publications will be presented.

Conclusions In this case study, journal publications provided insufficient information on patient-relevant harm outcomes of clinical trials. CSRs present considerably more adverse event data than publications including reports of SAEs.

Measuring wellbeing: great idea, but what's the use?

Paul Allin
Imperial College, London, UK

In 2010, David Cameron talked of measuring the progress of the country "not just by how our economy is growing, but by how our lives are improving", by quality of life, not just standard of living. This was at the launch of ONS's measuring national programme, one of a number of initiatives that aim to measure wellbeing and progress by more than just economic statistics. It was a great idea and much welcomed. Recently this idea has re-surfaced in calls for 'responsible' or 'inclusive' capitalism. But it can also be traced back, not only to the 2010 Stiglitz Report ('Mismeasuring our lives') but also, for example, to Robert Kennedy's 1968 speech reminding us that 'there is more to life' than GDP. The similar idea of measuring 'sustainable development' continues to find support, including in discussions to replace the Millennium Development Goals. We note that work on new measures is not completed. In particular, the question of whether there should be a single, overall measure of national wellbeing is not resolved. However, our main concern is that while good measurement is necessary it is not sufficient. The wider measures need to be used - in public debate about the kind of society we want, but also to question if we need to change our behaviour. Drawing on the ONS framework for measuring national wellbeing we identify and explore the roles of various players, including national and local government ministers and officials, businesses, NGOs, the media and, ultimately, individuals and households.

Modelling the abundance of *Culex pipiens* in Portugal

Marilia Antunes[1,2], Patricia de Zea Bermudez[1,2], Maria da Conceição Proença[1,3], Maria Teresa Rebelo[1,4], Maria João Alves[5,6], Hugo Osório[5,6]
[1]Universidade de Lisboa, Lisboa, Portugal, [2]CEAUL, Lisboa, Portugal, [3]LOLS, Lisboa, Portugal, [4]CESAM, Lisboa, Portugal, [5]INSA, Lisboa, Portugal, [6]ARS, Lisboa, Portugal

Arthropod are the transmission vectors for the arbovirus, of which the *Culex pipiens* is the most common species. In order to assess which species are most commonly observed in Portugal and their distribution throughout the country, the National Health Institute, along with the regional and national health authorities, developed a vectors surveillance program. Mosquitoes are captured using traps, and both ecological and meteorological variables, as well as the geographical coordinates are recorded at the location of the capture. The geographical coordinates allow to obtain georeferenced information on variables that may affect the presence and abundance of the mosquitoes, such as the proximity of water masses. The data were recorded between May and October, 2006-2012. The observations are neither spatially nor temporally regularly distributed and, consequently, the data does not cover the whole country. A first approach consists on fitting a model to the expected number of specimens captured using a bayesian framework. The spatial dependence is introduced in the model by means of a structured random component. The data exhibits many zeros and over-dispersion. The traditional zero inflated models are considered, although some other count distributions, such as the generalized Poisson, are also used. Abundance maps are presented for several meteorological and temporal scenarios.

This work was financially supported by national funds from FCT: project PTDC/SAU-SAP/119199/2010; projects PEst-OE/MAT/UI 0006/2014 and PTDC/MAT/118335/2010

Chain Event Graphs for assessing information in missing data

Jane L Hutton[1], Lorna Barclay[2], Jim Q Smith[1]
[1]University of Warwick, Coventry, UK, [2]Dunhamby, London, UK

Chain event graphs (CEGs) extend graphical models to address situations in which, after one variable takes a particular value, possible values of future variables differ from those following alternative values (Smith and Anderson, 2008, Thwaites et al 2010). These graphs are a useful framework for modelling discrete processes which exhibit strong asymmetric dependence structures, and are derived from probability trees by merging the vertices in the trees together whose associated conditional probabilities are the same.

We exploit this framework to develop new classes of models where missingness is influential and data are unlikely to be missing at random. Context-specific symmetries are captured by the CEG. As models can be scored efficiently and in closed form, standard Bayesian selection methods can be used to search over a range of models. The selected maximum a posteriori model can be easily read back to the client in a graphically transparent way.

The efficacy of our methods are illustrated using a longitudinal study from birth to age 25 of children in New Zealand, and a geographical cohort of people with cerebral palsy.

Smith, J.Q., Anderson, P.E, and Liverani, S. (2008) "Separation Measures and the Geometry of Bayes factor selection for Classification" J Roy. Statist. Soc. B, Vol. 70, Part 5, 957 - 980

Thwaites, P. Smith, J.Q. and Riccomagno, E. (2010) "Causal Analysis with Chain Event Graphs" Artificial Intelligence, 174, 889–909

Barclay, L.M., Hutton, J.L., and Smith J.Q. (2014) "Chain Event Graphs for Informed Missingness." Bayesian Analysis. 9:53-76.

African Institute for Mathematical Sciences: opportunities for statistical capacity buildling.

Jane Hutton
The University of Warwick, Coventry, UK

The African Institutes for Mathematical Sciences (AIMS) are centres for tertiary education and research, which promote mathematics and science in Africa. AIMS trains talented students and teachers in order to build capacity for African education, research, and technology. The first centre, by the sea and mountains at Muizenberg, Cape Town opened in 2003. AIMS Senegal opened September 2011 in MBour, within a seaside nature reserve. Courses are given either in French or in English. AIMS Ghana launched in 2012 and AIMS Cameroon in 2014. AIMS has already trained about 500 people, of whom a third are women, from more than 35 African countries.

The programme has introductory skills courses, and then a series of six three-week blocks in which students choose two out of three review courses. At present, the majority of courses offered are in theoretical physics, traditional applied mathematics and pure mathematics, which reflects the impressive work of the founders Neil Turok and Fritz Hahne. My vision is that a statistics course is always one of the skills courses, and that in each review block there is a statistics course. The main barrier to the provision of substantial statistics capacity building through AIMS is that there have not been enough volunteers offering courses and supervision of essays.

One advantage of teaching in Africa is reduced travel and living costs. Another is that more than half of students trained at AIMS remain in Africa to support Africa's developmental growth. I hope to inspire you to contribute.

High Density Linkage Mapping using Multidimensional Scaling

Katharine Preedy, Christine Hackett, Thanasis Vogogias
Biomathematics and Statistics Scotland, Scotland, UK

Genetic linkage maps are crucial for locating genes responsible for observable traits in plants and animals. A linkage map is a set of genetic markers in order along each chromosome, with relative locations, and due to the development of high throughput DNA technologies, the sizes of marker data sets are increasing. Therefore, rapid methods are needed for ordering these markers and estimating the map. Markers are ordered using pairwise estimates of the frequency of recombination between them, but it is necessary to allow for the varying precision of the estimates, which depend on the parental genotypes. This is a particular challenge in genetically complex plant species such as blackcurrant and autotetraploid potato. The strength of each pairwise linkage is measured by the log of the odds ratio, or LOD score. Here we propose a method of constructing linkage maps using metric multidimensional scaling (MDS), weighted by the LOD score. We propose two procedures; the first procedure involves constrained MDS followed by projection of markers onto an arc, and the second involves unconstrained MDS followed projection of markers onto a principal curve. The first procedure is extremely robust, but does not allow the user to bring prior knowledge to the projection. The second procedure is sensitive to the penalty imposed for deviations from linearity in the principal curve, but allows the user to incorporate prior knowledge about the maps. Both procedures are considerably faster than current methods and give good estimates of marker order and inter-marker distances.

Modelling the foraging behaviour of terns

Jacqueline Potts, Mark Brewer
Biomathematics and Statistics Scotland, Aberdeen, UK

The UK holds internationally important populations of breeding seabirds, and the Joint Nature Conservation Committee (JNCC) is responsible for coordinating government policy to protect these populations in accordance with international agreements. Our work aimed to develop models of habitat usage based on environmental covariates by four species of terns around their breeding colonies to inform the designation of marine Special Protection Areas. This is the first time that UK-wide models have been developed for terns.

JNCC collected the data on foraging behaviour by following individual birds in a boat and recording locations along a track using a GPS, together with the behaviour of the bird at the time. These data represent presence-only data and we therefore simulated pseudo-absences within the foraging range for use as controls. Some previous similar work has used random effects modelling, but we chose instead to use a weighted generalised linear model with the weighting accounting for the tracks. We also investigated the use of INLA (Integrated Nested Laplace Approximation) software to account for spatial autocorrelation.

The most important covariate was distance from the breeding colony, with distance from the shore also being important for Sandwich terns in particular. Associations were also found with other environmental variables such as depth, chlorophyll concentration and salinity. Many of the environmental covariates were quite highly correlated, many models had similar predictive power, and so the final choice of models was influenced by an assessment of biological plausibility as well as by the statistical properties of competing models.

An approach for summarising the association of multiple correlated features

Marina Evangelou[1], John Todd[1], Chris Wallace[1,2]
[1]University of Cambridge, Cambridge Institute for Medical Research, Cambridge, UK, [2]MRC-Biostatistics Unit, Cambridge, UK

Summarizing the association of multiple correlated features with a single response variable is a commonly faced challenge in the area of statistical genomics. Several methods have been proposed for combining the association of multiple independent features into a single statistic. For example the Fisher's product method is a powerful method, but the null distribution of its statistic does not hold for correlated features. Permutation procedures are usually employed for finding the null distribution of the chosen statistic which can be computationally intensive and require access to the raw data, which are not always available.

We have adapted an alternative method for finding the null distribution of the chosen statistic. The correlation structure of the tested features is considered as the covariance matrix of a multivariate Normal distribution. Z-scores for these features are then drawn from this distribution, P-values are subsequently calculated and the chosen statistic is re-computed using these simulated P-values.

We have explored this approach in the setting of a genomewide association study, where the association of a gene is found by combining the association of correlated SNPs located near the gene with the phenotype of interest. We demonstrate a very high correspondence between the results found through permutation and our proposed approach (Spearman correlation rho>=0.90). Alternative areas of application include gene expression experiments where the interest is the association of modules of correlated genes with the phenotype. We compare the simulation approach to established alternatives based on generating a univariate summary for each module, usually through principal components.

A Proposed Modification to the Smith-Satterthwaite Test for Autocorrelated Data

Nigel James
Sigmametrics Consulting, Sheffield, UK

Almost by definition, data derived from pathology specimens in medicine are both heteroscedastic and autocorrelated. Achieving a diagnosis is often dependent upon laboratory results being clearly outside the normal range of variation. Where there have been investigations in only a small number of clinical cases, there are often unequal heteroscedastic small sizes which cannot usually be increased for ethical reasons, so the use of a Smith-Satterthwaite test (Miller and Freund, 1977) is most appropriate.

For the testing of differences between means of heteroscedastic samples of differing sizes, the Smith-Satterthwaite tests is commanly used in the physical sciences but little used, if ever, in medicine. The Smith-Satterthwaite test conveniently allows for the testing of real life diagnostic samples in comparison with normal values or post-treatment samples.

However, although both the Student t-test (Alber's test, 1975) and z-tests have long been modified for use in environmental studies using autocorrelated data, it is thought that the Smith-Satterthwaite test has not previously been modified for autocorrelated data. Using autocorrelated data without specific modification eventually lead to false significance levels.

The proposal for modification of the Smith-Satterthwaite test uses the method to allow for autocorrelation for spatial data (Cliff and Ord, 1975) in which the sample numbers (N) in variance error terms in denominators and *df* calculations are replaced by a simple term N(1 - r) where r is the autocorrelation.

Albers, W. (1978) *Annals of Statistics*, **6**: 1337.

Cliff A. and Ord J. *J Roy Stat Soc,* **37B**: 297

Miller, I and Freund J. (1977) *Statistics for Engineers*, Prentice-Hall, USA.

Emerging Topic: Energy Efficiency

Graham Johnson
EEVS Insight Ltd, London, UK

More statisticians are needed to help unlock the massive potential the energy efficiency industry has to bridge the gap between supply and demand for energy.

This poster is designed to raise awareness of this emerging area of statistics to the statistical community and the career opportunities for statisticians in this area.

In particular, the field of Measurement and Verification (‘M&V') focusses on the certainty around quantifying the savings achieved by energy efficiency initiatives.

Statistical Approach to MESS Epilepsy Data

Boryana Lopez[1], Jennifer Rogers[2,1], Jane Hutton[1]
[1]University of Warwick, Coventry, Warwickshire, UK, [2]London School of Hygiene and Tropical Medicine, London, Keppel Street, UK

The Multicentre study of early Epilepsy and Single Seizures (MESS) consists of a study and its resulting data set. Patients in the study had a history of epileptic seizures and their clinicians were unsure of the need for an anti-epileptic drug (AED). Patients were recruited for over five years, and randomized to one of two policies: deferred or immediate treatment. This study was developed as an attempt to assess which policies are optimal for the diverse groups of epilepsy patients. It is of particular interest to understand what the risk of recurrence is, once a first seizure has happened, and how the treatment alters that risk.

The model presented here was developed by J. Rogers and J.L.Hutton(2012), and proposes a Poisson-Gamma mixture model for the times to seizure recurrence, considering the existence of an underlying process for each patient and a multiplicative change of seizure rate after the randomization. During the study it is observed that a proportion of the population does not experience another seizure post-randomization, indicating a remission or "cure rate", which is considered and estimated from the Poisson process model and is made to depend on covariates such as type of epilepsy and Electro Encephalogram (EEG) outcomes. The corresponding log-likelihood is then estimated by means of using the EM algorithm.

Stepwise backwards elimination concluded that the optimal joint model included treatment policy, seizure type, EEG outcome and their interactions for the change in seizure rate at randomization and following a first seizure post-randomization.

Variation in cancer incidence (1996-2010) and mortality (1997-2011) by deprivation quintile, in England

Jennifer Yiallouros[1], Nick Ormiston-Smith[1], Claudia Oehler[2], Sean McPhail[2], Lucy Elliss-Brookes[2], Monika Ciurej[1]
[1]Cancer Research UK, London, UK, [2]National Cancer Intelligence Network, London, UK

Objectives

Reducing inequalities in cancer incidence and mortality is a key goal of the Improving Outcomes Strategy for Cancer. Risk factors, including smoking, diet, drinking and exercise, affect the rate of cancer between socio-economic groups. This study builds on previous reports of cancer incidence by deprivation.

Methods

For 37 individual cancer sites and all cancers combined, incidence (1996-2000, 2001-2005, 2006-2010) and mortality (2002-2006, 2007-2011; all cancers additionally including 1997-2001)in England were analysed by deprivation quintile. Statistical significance tests were performed on deprivation trends across quintiles and changes in trend over time. For statistically significant trends, excess cases and deaths were calculated.

Results

If the more deprived had the same rates as the least deprived, there would have been around 15,300 fewer cases and 19,200 fewer deaths per year, for persons, across all cancers combined in the latest 5-year periods. Lung cancer dominates with around 11,700 excess cases and 9,900 excess deaths per year.

Deprivation trends of cancer incidence and mortality have not improved over time, with the gap reducing in 2 sites and increasing in 5 others, for incidence. Mortality saw no change.

In the latest periods, for persons, 24 sites had statistically significant deprivation trends; some cancers (including breast, prostate, melanoma) showed inverse deprivation trends, with highest rates in the least deprived quintile.

Conclusion

Inequalities have not reduced. The results could be used to identify areas which may benefit from further targeted interventions to improve outcomes for more deprived populations.

Acknowledgement: CRUK/NCIN partnership

Modelling risk of an adverse outcome in anticoagulated patients with a head injury

Joanne Rothwell, Maxine Kuczawski, Suzanne Mason, Matthew Stevenson, Michael Holmes, Shammi Ramlakhan, Steven Goodacre, Rosemary Harper, Francis Morris, Dawn Teare
University of Sheffield, Sheffield, South Yorks, UK

Objectives and background:

Existing practice in emergency departments (EDs) in the UK for managing anticoagulated patients after blunt head trauma is variable and based on anecdotal evidence. Before January 2014 NICE guidelines recommended a CT scan for patients based on clinical factors. The prospective observational multi-centre AHEAD study enrolled 3534 anticoagulated patients who attended 33 EDs in England and Scotland after blunt head trauma. Based on this data, we have identified risk factors associated with the occurrence of a serious head injury.

Methods:

Head injury complication was defined as head injury-related death, neurosurgery resulting from injury, clinically-significant CT head scan or hospital re-attendance with significant head injury complications. Factors including Glasgow Coma Score (GCS), level of anticoagulation (INR) and neurological symptoms were considered in the risk modelling.

Results and Conclusion:

An adverse outcome was found in 211(6%) of patients. Preliminary analysis suggests that the strongest predictors of risk of an adverse outcome are GCS scores below 13, vomiting and loss of consciousness. A GCS <13 and > 1 episode of vomiting form part of the NICE criteria for performing a CT scan so we would expect these to have some predictive capability. However, 92 (44%) of the adverse outcomes occurred in the 2755 patients who did not meet any of the NICE criteria. Further work will explore models to predict adverse outcome within clinically relevant substrata in the cohort.

Inferences from confidence intervals based on small numbers of events: Lessons from a Cochrane Review Group

Matthew Grainge, Jo Leonardi-Bee, Hywel Williams
University of Nottingham, Nottingham, UK

Objectives: Reviews published in the Cochrane library frequently have common outcomes (>10% of randomised participants), hence statisticians would recommend using the risk ratio as a measure of effect. However, if outcomes are rare (<5%) or the number of participants is small (<30) confidence intervals based on large sample approximations can provide misleading results. This can be problematic if firm inferences about the benefit or harm of a treatment are made on the basis of these. We explored the scale of this problem within the Cochrane Skin Group (CSG).

Methods: We took a random sample of 20 published CSG reviews. For each review we identified the number of analyses (defined as a forest plot for any one treatment comparison/outcome) which contained a single study (with a binary outcome) with either <30 randomised participants or <10 outcome events.

Results: The 20 reviews contained 742 separate analyses. Of these, 56 (7.5%) contained a single study where less than 30 participants were randomised in total and 88 (11.9%) contained a single study with <10 outcome events. In 4 instances data with potentially misleading confidence intervals were presented in the review abstract and in a separate 2 instances the implications for practice section was influenced by such analyses.

Conclusions: The CSG now request that where results are based on individual studies with small numbers, the authors should report the proportion of outcomes in each treatment group together with a p value from a Fisher's Exact test instead of confidence intervals which use large sample approximations.

Automatic generation of scientific theories to fit experimental data

Peter Sozou[1,2], Peter Lane[3], Mark Addis[4,2], Fernand Gobet[1,2]
[1]University of Liverpool, Liverpool, UK, [2]London School of Economics, London, UK, [3]University of Hertfordshire, Hatfield, UK, [4]Birmingham City University, Birmingham, UK

We describe a method for the automatic generation of scientific theories which are consistent with experimental data, in cognitive science. A mechanistic theory of a cognitive process involves a series of operations, such as comparing variables or adding to or retrieving from short-term memory. It can be expressed in precise form as an algorithm, which can be simulated as a computer program. A "good" theory can be considered to be one which makes predictions consistent with the data, whilst ideally also being parsimonious. A systematic, exhaustive search of all possible theories is generally not feasible due to the size of the search space. We use an evolutionary computational method known as genetic programming to progressively evolve theories, by minimising a cost function. The optimisation problem is analogous to regularisation in inverse problems, with a trade-off between fitting closely to data and parsimony. We will present results showing this trade-off when the method is applied to experimental results for the delayed match-to-sample problem, which is concerned with subjects' speed and accuracy in recognizing an image shown earlier, building on earlier work on applying genetic programming [1, 2]. Results for attention experiments (responses to cues) may also be presented.

1. Frias-Martinez, E., & Gobet, F. (2007). Automatic generation of cognitive theories using genetic programming. *Minds and Machines*, *17*(3), 287-309.

2. Lane, P. C., Sozou, P. D., Addis, M., & Gobet, F. Evolving Process-Based Models from Psychological Data using Genetic Programming. Proceedings of the AISB-2014 meeting, Goldsmiths, University of London.

MIM_sim: A mixed-inheritance simulation package for testing hypothesised disease models within pedigrees

Alexandra Gillett, Ammar Al-Chalabi, Cathryn Lewis
King's College London, London, UK

Pedigree simulation packages enable users to generate disease and linked markers under varying evolutionary pressures, allowing comparisons between competing linkage and association analysis methods. When exploring the genetic architecture of disease by comparing observed and simulated phenotypic summary measures, e.g. recurrence risk ratios (RRRs), such packages are computationally inefficient due to the unnecessary generation of non-causal markers. Here we present a software package for simulating three generation pedigrees ascertained on an affected proband in the last generation. Disease is generated via a user-specified mixed-inheritance model; a liability threshold approach where disease can be associated with multiple common genetic loci of small effect (a polygenic component), a large effect (major) genetic loci and/ or an environmental component. Users specify:

- Major genetic loci; risk allele frequency, mode of inheritance via a penetrance function
- Polygenic component; proportion of disease liability variance attributable to common genetic variation
- Environmental component: distribution type (normal, poisson, binomial), distribution inputs (e.g. λ if distribution type is poisson).

Summary functions are available to compute RRRs for various degrees of relatives under the pre-specified disease model.

The utility of this simulation tool is demonstrated through an application to a hexanucleotide repeat expansion within *C9orf72*, which is associated with Amyotrophic Lateral Sclerosis (ALS) but demonstrates incomplete penetrance. Using mixed-inheritance models we investigate whether this incomplete penetrance is due to an unobserved polygenic component.

This package, and its future extensions, will be used to explore the types of genetic architecture that are consistent with observed familial patterns of ALS.

ON DETERMINATION OF ECONOMIC THRESHHOLD LEVEL - A STATISTICAL APPROACH

Satyabrata Pal[1], Arunava Ghosh[2]
[1]Indian Statistical Institute, Kolkata, West Bengal, India, [2]Uttarbanga Krishi Viswavidyalaya, Coochbehar, West Bengal, India

Economic threshold level (ETL) is regarded as an important and indispensable concept (component) in pest management and control. The value of this index is determined by available formula based on economic parameters. The knowledge of ETL helps reduction of crop loss (and ensures less pesticide application), and as a sequel, profit is increased. Economic injury level (EIL) is a concomitant concept and substantial knowledge is required on the dynamics of pest population in order to determine the density at which the economic injury level (EIL) may be prevented (Weersink et al*. 1991). This paper is devoted to the development of a method (based on statistical consideration) of determination of ETL, which is defined as the density at which control measures should be determined to prevent an increasing pest population from reaching the economic injury level. The gamut in regard to the methods proposed includes obtaining the distribution of pests and modelling the dynamics of the pest population growth in order to determine ETL on real life data sets on incidences of pests (Whitefly, Blackfly and other pests) on different crops, namely, betelvine, rice, chilly, brinjal, and others, obtained from distinctive experiments designed for the purpose. The benchmark probabilities (probability levels in the range, 0.25 to 0.5) obtained from appropriate distributions determine the corresponding ETL values against the respective crops.

*Weersink A., Deen W., Weaver S. (1991): Defining and Measuring Economic Threshold Levels. Canadian Journal of Agricultural Economics 39(4): 619-625.

Meta-analysis of logistic regression coefficients

Daisuke Yoneoka[1], Masayuki Henmi[2]
[1]The Graduate University for Advanced Studies, Tokyo, Japan, [2]The Institute of Statistical Mathematics, Tokyo, Japan

In this presentation, we propose a method for meta-analysis of logistic regression coefficients.

As Becker and Wu (2007) pointed out, there are some problems in meta-analysis of regression coefficients. One of them is that a set of covariates is not necessarily common in each study included in meta-analysis. If each study uses a common set of covariates to perform a regression analysis, then combining estimates of regression coefficients can be done by the technique of multivariate meta-analysis. If not, however, synthesis of regression coefficients is not so straightforward since each estimate of the coefficient for the same covariate can be biased. First of all, we derive a formula for adjustment of this bias in logistic regression models, and then propose a nonlinear weighted least square methods for meta-analysis of regression coefficients by incorporating the bias formula. The performance of this method is evaluated by both theory and simulation studies, and is shown to be better compared to meta-analysis using studies only with a common set of covariates and meta-analysis using imputation for missing coefficients. This research is motivated by meta-analysis of clinical prediction models, which have been recently accumulated for various diseases such as cancer and diabetes. Therefore, we focus on logistic regression models, but the idea of our method can be applied to other types of regression models.

Some Problematic Matters in Meta-Analysis

Paul Marchant[1,2]
[1]Leeds Metropolitan University, Leeds, West Yorkshire, UK, [2]University of Leeds, Leeds, West Yorkshire, UK

The logical perfection of combining RCTs can be violated when dealing with data from actual trials. This poster examines a meta-analysis which claims large benefits of installing new street-lights. The studies are synthesised as Controlled Before-After trials. There are several problems to be resolved. The events (crimes) are not statistically independent thereby causing the problem of overdispersion and potentially masking heterogeneity between studies. Additionally there are the effects of regression towards the mean and publication bias. The aim of the work is to estimate the potential impact of such matters.

The effect of non-independence can be partly addressed by including overdispersion in a GLZM. However the fact that some studies have more than just two (Before and After) time points, potentially allows for separate overdispersion measures and hence adjusted weights in the meta-analysis. As regards the effect of regression towards the mean, data from another crime study done on a small geographical scale suggests the before and after distribution from which the bias can be estimated. Publication bias is as ever a tricky issue but it can cause a considerable overestimate of an effect, as is known from health-care.

This is work in progress, however the magnitude of the biasing effects estimated here may well account for the discrepancy between the promise of beneficial effect from the original meta-analysis and its observed absence in a large scale roll out of the intervention.

Temporal and Spatial Analysis of TB in Singapore

Sourav Das
National University of Singapore, Singapore, Singapore

Tuberculosis(TB) has long been a major cause of death across the globe. In 1993 the World Health Organization (WHO) declared TB a global health emergency. TB was also highlighted in the United Nation's millennium development goals. Much has been achieved since then; especially the reduction of 45 % in mortality rates. However challenges persist in terms of prevalence of the disease. With 29% of the reported global TB cases South East Asia remains a significant contributor to global TB incidence and Singapore, a major economic hub in the region, has witnessed a sudden increase
in TB incidence contrasting the trends among advanced economies. In this two part analysis we investigate the temporal and spatial patterns of TB risk in Singapore. We model the monthly temporal risk of TB data using Seasonal Autoregressive Moving Average (SARIMA) and Generalized Linear Autoregressive Moving Average (GLARMA) methodology. However TB is a communicable disease driven by several socio-economic and environmental factors leading to a random spread (as also the count) of the disease in a geographical area. Within a fixed time period, assuming TB incidence to be a spatial point process we construct risk maps of TB in Singapore over a period of seventeen years (1995-2011), using standard point process methodology. Based on temporal dynamics of these maps we propose hypotheses that need further study. We classify Singapore's administrative districts based on the interpolated spatial risk. We construct a Monte-carlo test to assess the statistical significance of change in overall risk of the disease.

Keywords: intensity function, Ripley's K-function, aggregated spatial pattern.

Modelling Stroke Outcome Data over time

Jessica Potts
Keele University, Keele, UK

A stroke occurs when blood to the brain is obstructed, leading to brain damage or even death. For those who have had a stroke whether they will recover and how long that process will take is important. I am interested in the longitudinal analysis of the modified Rankin Scale (mRS) (Rankin, Scott Med J, 1957).

The mRS is a hierarchical ordinal scale with 6 scores (0-5) that ranges from a patient having no disability (0) to severe disability (5). The mRS is most the most common outcome used in clinical trials and recently a score of 6 has been included to represent death. By including a score of 6 for death in the mRS, it can be argued that the ordinality of the scale is removed.

The challenge is to find a justification for including death as a point on the scale and apply appropriate longitudinal methods. A proportional odds model could be used. We could ignore the ordinal structure and apply a multinomial model, comparing the scores to a reference category. Alternatively it may be more appropriate to treat death as a separate state and consider Latent Transition Models or Multi State Models, or even adopt a time to event analysis using Cox regression.

The mRS is not the only scale that includes death in the scale. The Neurological Disability Scale (Swank, Lancet, 1990) starts with no impairment (0) and goes up to death (6). How we treat these extremes will determine the most appropriate modelling strategy.

Variable Selection for Latent Class Analysis

Michael Fop, Thomas Brendan Murphy
University College Dublin - Insight Centre for Data Analytics, Dublin, Ireland

A variable selection method for latent class cluster analysis is proposed. At each step of the selection procedure the usefulness of a variable is assessed comparing two models. In one model the variable adds further information about the clusters, and in the other model it does not but it can be related to the already selected clustering variables. The method is capable of discarding not only those variables that do not contain any information about the clusters, but also those that are superfluous, leading to a parsimonious selection. The models are compared via an approximation to the Bayes factor and the search over the model space is performed using a backward-stepwise greedy algorithm. An application to medical data related to musculoskeletal pain is presented, and the proposed method recovers the true group structure with a small number of variables.

Index

N

O

P

R

S

T

U

W

Y

Z