Introduction

Welcome to the RSS 2008 International Conference in Nottingham.

The abstracts of the nearly 200 presentations being given during this conference have been compiled in this booklet in the following order:

- Plenary speakers in order of presentation
- Invited, Contributed and Special sessions in the order that they appear in the programme
- The Poster presentations in the numbered order of display.

The posters will be on display in the refreshments and exhibition area of the Conference Centre throughout the week with a dedicated wine reception on Wednesday evening at which the presenters will be available to discuss their work.

Finally please note that the allocation of contributed papers to topics may not be in accord with the speaker's request but has been made by the Conference Programme Committee to try and ensure as coherent a set of sessions as possible.

I hope you enjoy the conference.

Paul Gentry
RSS Theme Manager for Meetings & Conferences

P.S. Please don't forget to visit the stands in the Exhibition Hall and chat to the various organisations exhibiting during the conference

Plenary Speakers

Plenary 1

Professor Jerome H Friedman

Stanford University

JHF@stanford.edu

Sequoia Hall, 390 Serra mall, Stanford, CA 94305 USA

Fast Sparse Regression and Classification

Regularized regression and classification methods fit a linear model to data, based on some loss criterion, subject to a constraint on the coefficient values. As special cases, ridge regression, the lasso, and subset selection all use squared-error loss with different particular constraint choices. For large problems the general choice of loss/constraint combinations is usually limited by the computation required to obtain the corresponding solution estimates, especially when non convex constraints are used to induce very sparse solutions. A fast algorithm is presented that produces solutions that closely approximate those for any convex loss and a wide variety of convex and non convex constraints, permitting application to very large problems. The benefits of this generality are illustrated by examples.

Jerome H. Friedman is Professor of Statistics, Stanford University. He received both AB and Ph. D degrees from the University of California, Berkeley. He has authored or co-authored over 50 papers in major statistical and data mining journals as well as two books on Data Mining. He has invented or co-invented several widely used data mining procedures.

Plenary 2

Professor Martin Hasler

Ecole Polytechnique Fédérale de Lausanne (EPFL)

martin.hasler@epfl.ch

School of Computer and Communication Sciences
Station 14, CH-1015 Lausanne, Switzerland

Asymptotic behaviour of blinking (stochastically switched) dynamical systems

We discuss the behavior of continuous-time dynamical systems that have external input signals that are constant in small time intervals and that can only take the values 0 and 1. They can be interpreted as switching between 2m different dynamical systems if there are m such input signals. Switching is supposed to be fast with respect to the time constants of the (non-switched) different systems. Therefore, one expects that the switched (blinking) system behaves like the time-averaged system. More precisely, we suppose the switching to be stochastic such that the value of external signal at a certain time interval is a random variable and that all these random variables are independent.

In general, the solutions of the blinking and the averaged system starting from the same initial state, stay close together if the switching is fast, but this property holds only for finite time. However, if the solution of the averaged system converges to an attractor, this is also true for the corresponding rapidly switched blinking system under some weak hypotheses. In general, the solutions do not stay close together forever, but they converge to the same attractor.

Strictly speaking, this is only possible if the attractor of the averaged system is an invariant set to the blinking system. However, if this is not the case, the solution of the blinking system will still come close to attractor of the averaged system and stay close in a probabilistic sense. Furthermore, if the averaged system has more than one attractor (multi-stability), there is a small, but positive probability that the solution of the blinking system converges to another attractor than the solution of the averaged system.

Hence, there are 4 cases to distinguish, according to whether or not the averaged system has more than one attractor and whether or not the attractor(s) of the averaged system is (are) invariant under the blinking system. We give for each case an example and we prove a theorem that characterizes the relation between the asymptotic behavior of the averaged and the blinking system. Depending on the case, convergence of the blinking system to the attractor of the averaged system is either in the strong sense (for almost all switching sequences) or in the weak sense (with probability converging to one) if switching is fast.

Martin Hasler received the Diploma in 1969 and the PhD degree in 1973 from the Swiss Federal Institute of Technology, Zurich, both in physics. He continued research in mathematical physics at Bedford College, University of London, from 1973 to 1974. At

Plenary 3

Alan Agresti

University of Florida

aa@stat.ufl.edu

1632 NW 24 Street
Gainesville, Florida 32605, USA

Good Confidence Intervals for Discrete Statistical Models

We survey good methods for constructing confidence intervals for parameters in discrete statistical models, with emphasis on categorical data. The method of inverting score tests for parameter values performs well, usually much better than inverting Wald tests and often better than inverting likelihood-ratio tests. Exact small-sample methods are conservative inferentially, but inverting a test using the mid-P value provides a sensible compromise. For some models ordinary score inferences are impractical, such as when the likelihood function is not an explicit function of the model parameters. For such cases, we propose pseudo-score inference based on a Pearson-type chi-squared statistic that compares fitted values for a working model with fitted values of the model when the parameter of interest takes a fixed value. Finally, we briefly summarize a different pseudo-score approach that approximates score intervals for proportions and their differences by adding artificial observations before forming simple Wald confidence intervals.

Alan Agresti is Distinguished Professor Emeritus of Statistics at the University of Florida. He earned a doctorate in Statistics from the University of Wisconsin in 1972. He is author or co-author of five textbooks, including "Categorical Data Analysis"

Invited and Contributed Sessions

1A Statistical Shape Analysis (Invited)

Professor John Kent

University of Leeds

J.T.Kent@leeds.ac.uk

Department of Statistics, University of Leeds, Leeds LS2 9JT, UK

New developments in projective shape analysis

Projective shape refers to information recorded on a camera image that is invariant under changes of the camera view. It is an important tool in machine vision for identifying common features in images of the same scene taken from different camera angles. The simplest example is the cross ratio for 4 points on a line. In this paper we describe the beginnings of a metric theory for projective shape which provides the tools needed to estimate shape averages and shape variability. The methodology is analogous to the more familiar Procrustes methodology for similarity shape analysis.

1A Statistical Shape Analysis (Invited)

Adrian Bowman

Department of Statistics, The University of Glasgow

adrian@stats.gla.ac.uk

Department of Statistics, The University of Glasgow
Glasgow G12 8QQ

Statistics with a human face

Some forms of medical imaging allow the extraction of three-dimensional surface data, which poses interesting problems of statistical modelling. One example is in facial surgery, where there is interest in describing the facial shape and growth of healthy children and contrasting this with the shape and growth of children who have been born with a cleft lip and/or palate and who have subsequently undergone surgical repair. An immediate issue is how the information in these images should best be extracted. Anatomical landmarks provide a useful starting point and there are well developed methods of analysis for this kind of data. Facial curves, with clear anatomical meaning, can also be extracted, to exploit the much richer information present in each digitised face. Standardised meshes, whose nodes correspond across individuals, can also be fitted to represent the surface as a whole. Some of the issues involved in analysing data of these types will be discussed and illustrated on the facial growth study. These include the extraction and representation of information, its graphical exploration, and issues of statistical modelling, including longitudinal structures.

1A Statistical Shape Analysis (Invited)

Dr Chris Brignell

University of Nottingham, UK

chris.brignell@nottingham.ac.uk

School of Mathematical Sciences, University of Nottingham, Nottingham, NG7 2RD, UK

Joint authors: Ian L. Dryden

Surface shape analysis, with an application to brain cortical surface analysis in schizophrenia

We introduce some methods for the statistical analysis of surface shapes and asymmetry, with particular focus on the investigation of large-scale cortical surface shape differences between control and schizophrenia patients. Key aspects of shape analysis are to remove nuisance transformations by registration, and to identify which parts of one object correspond with the parts of another object. We introduce Bayesian and maximum likelihood methods which are particularly appropriate for registering brain images and providing large-scale correspondences of the cortical surfaces. Size and shape analysis of the registered surfaces is then carried out, focusing on overall size and shape and certain differences in asymmetry (called torque). Initially whole cortical surface analysis is considered, with dimension reduction carried out using principal and independent components analysis. Some small but significant findings are observed. We then investigate asymmetry, and significant differences in asymmetry are observed between the control and patient groups, which concur with findings from the literature. Further investigation of the midline plane location in the two groups, and the fitting of non-planar curved midlines are also considered.

1B Statistics in Finance (Contributed)

Jonathan. D Rees MEng AMIMechE

Queens University Belfast (Systems Engineering Doctorate Centre) & Rolls-Royce Defence Aerospace

Jon.Rees@Rolls-Royce.com

PO Box 3 (WH-51), Rolls-Royce Defence Aerospace, Filton, Bristol, BS34 7QE

Joint authors: Dr Richard Curran - Senior Lecturer, Queens University Belfast

"Pricing for Risk: An Industry Perspective"

What's the difference between uncertainty and risk?… It's probably the most sadistic question one could ask in any risk management meeting, but it's an important one. This understanding was especially significant when one company tried to implement a unified corporate strategy, for the risk management of complex power system services. In practice, controversy and a lack of synergy had previously led to a wasting of both human and financial resources, ultimately leading to reduction in quality.

Although this strategy is still in its infancy, it is set to have a huge influence on business, both internal and external, none more so than that illustrated through life cycle costing. This paper therefore, is going to explore two perspectives on risk modelling, firstly looking inward to see how the company is developing proactive risk managed services, and secondly, reflecting on the customers perception of 'value added'.

Whilst much of the analytical process and detail is only outlined, this paper aims to illustrate how those involved in all aspects of risk management, from data capture to business management, can contribute to high quality, 'win-win' type contracting, i.e. low price / low risk.

1B Statistics in Finance (Contributed)

Alan Forrest

HBoS plc

alanforrest@hbosplc.com

Group Credit Risk Analytics, HBoS plc., Level 2, Teviot House,
41 South Gyle Crescent, Edinburgh EH9 3ET

Statistics and Financial Regulation – Basel II and Low Default Portfolios

On 1st Jan 2007, the EU Capital Requirements Directive (CRD) came into effect. This regulation, which enforces the 2005 Basel II Accord, requires all EU Lending Institutions to set aside money (Regulatory Capital) to offset unexpected losses at the 99.9% quantile.

Meeting, or even understanding these regulations requires an appreciation of statistical concepts such as correlated uncertainty and extreme events, and, to take greatest advantage of the CRD's benefits, many UK Financial Institutions have found they must bring new statistical understanding to their Regulatory Capital calculations.

This talk aims to show how statistical thinking has influenced the content and implementation of the CRD, and will illustrate three general points with concrete examples:

1. Statistical concepts are essential to modern international financial regulation.
2. Financial Regulators need statisticians to help define their rules and to monitor compliance.
3. Financial Institutions need statisticians to interpret the regulations and to put them into practice.

Examples will be drawn from the author's direct experience of CRD implementation in UK Financial Institutions over the last 5 years, and from his active involvement in the UK Financial Services Authority's Low Default Portfolios Expert Group.

1B Statistics in Finance (Contributed)

Adam. R Brentnall

Imperial College London

a.brentnall@imperial.ac.uk

Institute for Mathematical Sciences, 53 Prince's Gate, South Kensington, London SW7 2PG

Joint authors: Martin J. Crowder, David J. Hand

Model-based monitoring and prediction of individuals' cash machine usage

In this talk, statistical models for the time, amount, type of an individual account's automated teller machine transactions are described. Such models might be used within financial institutions' customer management strategies through predictions about likely behaviour, and by flagging occasions when usage was different from that expected. To predict, an empirical distribution of individual maximum-likelihood estimates is used to approximate the random-effects distribution of model parameters. To monitor for shifts in behaviour, accounts are ranked using likelihood-ratio change-point statistics. The results of applying the models and methods to a real data set show that the random-effects predictions can be better than projecting past behaviour forwards, and that monitoring behaviour may provide useful management information

1B Statistics in Finance (Contributed)

Jonathan Hill

University of North Carolina- Chapel Hill

jbhill@email.unc

Dept of Economics CB 3305, University of North Carolina, Chapel Hill, NC 27599-3305

Robust Non- Parametric Tests of Extremal Dependence

Dependence between extremes is predominantlymeasured by assuming a parametric joint distribution functional form, and almost always otherwise marginally iid processes. We develop non-parmetric tests of bivariate tall dependence for possibly heavy - tailed time series based on tall exceedancesand events. The tests capture extremal dependance decay over time, obtain asympototic power of one against infinitessimal deviations from tall independence, and apply to to dependent, heterogeneous processes with or without extremal dependence, irrespective of the degree of tall thickness and non-extremal properties. We do not require a joint distribution tall specification, and weak limits on a cadlag metric space are established allowing for sub-sample analysis. The tests are applied to analyze inter-market extremal associations within and between equity and foreign exchange rate markets.

Dr Gopalakrishnan Netuveli

Research Fellow, ReseDepartment of Primary Care and Social Medicine, Imperial College London

g.netuveli@imperial.ac.uk

3rd Floor, Reynolds Building, St Dunstan's Road, London W6 8RP

Using sequence matching to draw inferences about trajectories: the fascis analysis

Recently, optimal matching is increasingly being used to explore and describe trajectories. While the derived classes from optimal matching can be used as variables in various analyses, they cannot be used to make inferences because there are many ways of validly clustering the sequences which can potentially lead to conflicting results. In this paper we describe an approach, we call fascis analysis which uses sequence matching to draw inferences about trajectories. Fascis (Latin for bundle) analysis draws its name from the fact that trajectories are treated as 'bundles' with in groups. In establishing an association between grouping variables, we measure how 'tightly' the trajectories are bundled with in each group. The difference from cluster analysis is that our method defines bundles using an external variable. We define the tightness bundle in terms of equality of distribution of sequence distances with in groups compared to between groups. In this respect the logic is similar to the analysis of variance. We use Gini coefficient as a measure of inequality. The Gini coefficient can be decomposed as between groups, within groups and overlap. It has no distributional assumptions about the variable being used. Similarity of this procedure with ANOVA has lead to it being called ANoGi (Frick et al. 2004). In drawing inferences about trajectories, the usual practice of alignment in optimal matching can be problematic because each point in our trajectories represents an age, a period and a cohort influence. In searching for the longest common sequence, these influences are forgotten. In addition, there is the question of the distance to be used. In optimal matching the distance matrix is computed for all possible pairs. In our analysis we have the problem of determining whether the all the pairs or pairs within each group are to be used. We avoid it by calculating the distance from a reference sequence. This paper presents a proof of concept analysis using data on health and employment trajectories in BHPS.

Karen. L Smith

Centre for Statistics in Medicine

karen.smith@csm.ox.ac.uk

Centre for Statistics in Medicine, Wolfson College, University of Oxford, Linton Road, Oxford OX2 6UD

Joint authors: Dr Richard J. Gadsden

SASSY: One-to-One Tuition in Statistics

Sigma, the Centre for Excellence in Teaching and Learning in Mathematics and Statistics Support at Loughborough and Coventry Universities, has run a statistical advisory service for students since 2005. Over 500 students, undertaking final year undergraduate projects, master's dissertations or research degrees, have received advice through the service to date.

The service aims to provide support to students during both study design and data analysis phases. Unlike professional consultancy, in which the focus is on the best design to answer the question of interest and the most appropriate analysis, we frequently have to advise students to take a relatively simple approach to design and perhaps carry out a somewhat naïve analysis. This is to take account of the background and ability of the students in question. In addition we have to consider this to be a teaching opportunity and guide the students in developing the necessary understanding to explain and interpret their analysis in a written report and, in the case of research students, to defend the approach in a viva. This presents many challenges.

In this paper we will give practical examples of some of the issues that have arisen, and the advice we have given to the students. In particular, we will focus on those study designs that we deal with on a regular basis, such as crossover studies, pre-post designs, reliability, issues with multiple regression and correct handling of categorical data from surveys.

We will also discuss some of the difficulties that have arisen with host departments, who in the main will be marking the students' work, with some of the activities that we have undertaken and others that are planned to help address these. Curriculum content of the initial statistics teaching is key to resolving some of the issues, as is the pedagogical approach, and we will discuss some of the pertinent issues that are highlighted by our activities.

Dr Neil .H Spencer

University of Hertfordshire

N.H.Spencer@herts.ac.uk

Business School, University of Hertfordshire, de Havilland campus, Hatfield, Herts., AL10 9AB

Coping with Missing Levels in Multilevel Modelling

An assumption of standard multiple regression analysis is that the cases are independent. If not, standard errors of the coefficients will be underestimated, inappropriate models may be fitted and inappropriate conclusions drawn. To overcome this, multilevel modelling can be used, incorporating the hierarchical structure of the data into the modelling process.

However, multilevel modelling can only take into account those levels of a hierarchy which are known to the researcher. If a dataset lacks information on a level of the hierarchy then the multilevel modelling process which is undertaken can still lead to inappropriate models being fitted and inappropriate conclusions being drawn. Research on this "missing level" issue has demonstrated these effects. However, the only suggested remedy is that efforts are made to ensure that information on important levels of the data hierarchy are recorded at the time of data collection.

However, it may be the case that information on all levels of a data hierarchy is not available. This may be because the analysis uses data that has already been collected and it is not possible to add the relevant information on the missing levels. It may also be because when the data was collected, it was not thought important to collect information about a certain level of the hierarchy, and only with hindsight is this considered to be a potentially important aspect of the modelling. For researchers finding themselves in this position, the advice to collect information about the missing levels is not useful. What they need is a way of coping with missing levels. This is what is presented in this paper.

Another scenario where the work presented in this paper will be of use is when there might be unknown levels in the hierarchy of which the researcher is not aware (e.g. friendship groups in schools). The methods presented in this paper allow the researcher to investigate whether or not there may be missing levels, and what impact they might have on conclusions drawn.
This paper tackles the missing levels problem by attempting to construct potentially missing levels between known levels of the hierarchy. It uses model-based clustering of the residuals from multilevel models which have been constructed using the known levels. Constraints are placed on the model-based clustering and a best-fit model is obtained. The outcomes of this model-based clustering are then used to (re)construct the potentially missing levels. The effects of including these new levels of the hierarchy on the results of the multilevel modelling are examined. Based on these results from the revised multilevel model and the model-based clustering, a decision can be made whether to include the new (no longer missing) levels or exclude them from the analysis

1C General (Contributed) 1

Carlos Cuevas-Covarrubias

Anahuac University, Mexico.

ccuevas@anahuac.mx

Universidad Anahuac, Escuela de Actuaria,
Av. Lomas Anáhuac s/n, Lomas Anáhuac,
Huixquilucan Edo. Mex., C.P. 52786. Mexico

Joint authors: Arturo Cervantes-Trejo, M.D. and Ph.D.,
Anáhuac University, Mexico

Principal Components and ROC Curves: an alternative approach to reduction of dimensionality.

Principal Components and Discriminant Analysis are usually discussed separately without exploring its simultaneous application; this work, however, presents an original combination of both. Its main contribution is the proposal of a new criterion to control reduction of dimensionality. Based on the multivariate bi-normal model for ROC curves, our method simultaneously obtains principal components for two heteroskedastic independent samples; the reduction of dimensionality is assessed in terms of the area under the ROC curve of an optimal discriminant function. In order to illustrate its application, we present an example where four psychiatric variables are analyzed in two groups of teenage girls: sexually active and non-sexually active. The results of the example suggest that this methodology has a worth considering practical potential

1D Six Sigma - What are the current issues? (Invited)

Professor John Oakland Error! Bookmark not defined.
Oakland Consulting plc

johnoakland@oaklandconsulting.com
33 Park Square West, Leeds LS1 2PF

Strategies and capability development for Lean – Six Sigma; real improvements in on-quality, on-time, on-cost delivery

In this paper Professor Oakland will describe the design, development and implementation of successful performance improvement programmes in some of Europe's largest companies. He will describe how a framework has been developed and implemented to ensure customers, processes, people and supply chains can be brought together to deliver improvements in on-time, on-quality, on-cost performance.

The paper will show how to successfully:
• construct a strategic approach to quality and operational excellence
• tailor approaches such as six sigma and lean for the organisation
• obtain senior management buy-in to the approach, including business units in different sectors and geographic areas
• review customer perceptions in a positively active way
• design an integrated skills development programme for the people
• assess the supply chains to determine improvement opportunities
• roll out the programme to deliver real measurable business benefits.

1D Six Sigma - What are the current issues? (Invited)

Professor Tony Bendell

Services Ltd. and Abu Dhabi International Centre for Organisational Excellence

tony@servicesltd.co.uk/t.bendell@ioe.ae

Services Ltd., Quality & Reliability House
82 Trent Boulevard, West Bridgford, Nottingham, NG2 5BL

The Correct and Erroneous Statistical and Non-Statistical Basis for Six Sigma – Common Methodological Errors, Omissions and Overclaiming

Six Sigma is good. It is providing the basis for substantial improvement to both manufacturing and service processes in areas as diverse as healthcare, financial services and electronics.

Whilst statistically-based, the Six Sigma approach is a lot more than just about applying statistical methods to process problems, its success is also due in part to its emphasis on top management support, programme deployment and project management.

But there is bad Six Sigma out there too; incorrectly applied statistical methods, dated statistical tools, implicit assumptions and training-by-rote. Compounded with this is a cost-down focus masquerading as quality improvement. Nor is the Six Sigma deployment model appropriate in all application areas.

This talk reviews the correct and incorrect basis and claims for Six Sigma, and it identifies common errors, omissions and exaggerated claims. However, it concludes by focussing on what Six Sigma can and is contributing to business and other processes, and how statisticians can help, not hinder.

1E Meta-analysis (Contributed) 1

Stephan Morgenthaler

Imperial College London and EPFL, Lausanne

e.kulinskaya@imperial.ic.uk

Statistical Advisory Service, Imperial College,
8 Princes Gardens, South Kensington Campus, London SW7 1NA

Joint authors: Elena Kulinskaya, Robert G. Staudte

Variance stabilisation: overview

Statistical inference is heavily based on the large sample normality of the various statistics. Let $S_n \sim N(\mu, \sigma^2/n)$ with unknown μ and known σ . We want to test H_0: $\mu = \mu_0$ vs H_1: $\mu > \mu_0$. An effect is $\theta = \mu - \mu_0$ and the standardised effect is $\delta = \theta/\sigma$. Unfortunately, the variance σ^2 of a statistic S often depends on its mean μ.
This makes the inference more complicated and the convergence to normality slower.

Variance stabilisation helps to resolve both these problems. A variance stabilizing transformation (vst) $h(x)$ is needed so that for $T_n = h(S_n)$, the variance $Var[T_n] = 1 + o(1/n)$. Then $T = h(S)$ is approximately $N(E(T),1)$. The transformed statistic T is ideally on a unit normal scale for all values of parameters. The ideal is approached rapidly for a wide range of parameters occurring in practice.

Variance stabilization is a well established and very useful statistical technique with numerous publications on particular transformations in 1940s and 1950s, such as Anscombe (1948) for binomial and Azorin (1953) for t distribution, followed by more theoretical papers by Holland(1974), Efron(1982) and DiCiccio, Monti and Young (2006). Obtained approximations are of an amazing quality and they can easily be used for sample size calculation in lieu of the use of professional statistical packages.

This talk gives an overview of the methodology of variance stabilisation, including several examples, existence of the vst and large sample properties of the transformed statistics, as well as the rationale for using this technique in meta analysis.

References

Anscombe, F.J. The transformation of Poisson, binomial and negative binomial data, *Biometrika*,1948, **35**, 266-254

Azorin, P.F. Sobre la distribucion t no central I, II. *Trabajos de Estadistica*, 1953, 4, 173-198 and 307-337

Efron, B., Transformation Theory: How Normal is a Family of Distributions? *The Annals of Statistics*, 1982, **10**, 2, 323-339

Holland, P.W. Covariance stabilizing transformations, *Annals of Mathematical Statistics*, 1974, **1**,74-92.

DiCiccio, T. J., Monti, A. C. and Young, G. A. Variance stabilization for a scalar parameter, *Journal of the Royal Statistical Society, B*, 2006, **68**, 5, 873–875

1E Meta-analysis (Contributed) 1

Elena Kulinskaya

Imperial College London

e.kulinskaya@imperial.ac.uk

Statistical Advisory Service, Imperial College,
8 Princes Gardens, South Kensington Campus, London SW7 1NA

Joint authors: Stephan Morgenthaler, Robert G. Staudte

Variance stabilisation: combining the evidence

In the traditional fixed effects model (FEM) of meta analysis, given the estimated effects from K studies $\theta_1,\ldots, \theta_K$, with $\theta_i \sim N(\theta, \sigma_i^2)$, the combined effect θ is estimated as the weighted mean $\theta_{est} =(w_i \theta_1+\ldots+ w_K \theta_K)/W \sim N(\theta ,1/ W)$, where $w_i=\sigma_i^{-2}$ and $W= (w_1+\ldots+ w_K)$. If the homogeneity of the effects is rejected, the random effects model can be used: $\theta_i \sim N(\theta, \sigma_i^{-2} +\tau^2)$. (Sutton et al, 2000).

When the variance stabilizing transformation (vst) is applied to the estimated effects, we deal instead with the transformed standardised effects $K(\delta_i)$. They are estimated by $\kappa_i =n_i^{-1/2}h(S_i) \sim N(K(\delta),1/n_i)$ and can be added with known weights n_i in meta-analysis. (Kulinskaya, Morgenthaler and Staudte, 2008)

Given variance stabilized statistics from K studies $T_1,\ldots,T_K$, with $T_1 \sim N(n_i^{1/2} \kappa ,1)$, the combined effect $\kappa_{est}=(n_1 \kappa_1+\ldots+n_K \kappa_K))/ N \sim N(K(\delta),1/ N)$ where $N= n_1 +\ldots+n_K$. The back-transformation is used to obtain the inference on the standardised effects δ. If the homogeneity of the transformed effects is rejected, the random transformed effects model can be used:
$\kappa_i \sim N(\kappa , n_i^{-1} +\tau^2)$. Inference for both models is further discussed in the talk.

When there are no nuisance parameters (as in the 1-sample Binomial or Poisson case) these two approaches to meta analysis are equivalent. In the general case, the variance stabilization approach can be used even when the inference on the original, non-standardised effects is of primary interest. In this case the optimal weights depend on the nuisance parameters. An example is the variance stabilizing arcsine transformation for the difference in absolute risks, with the average risk as the nuisance parameter.

References

E. Kulinskaya, S. Morgenthaler and R.G. Staudte. *Meta-analysis: A Guide to Calibrating and Combining Statistical Evidence*. Wiley Series in Probability & Statistics. Wiley, Chichester, 2008.

Sutton, A.J. , Abrams, K.R.,d Jones, D.R.,Sheldon, T.A. and Song, Fujian. Methods for Meta-Analysis in Medical Research, John Wiley & Sons, 2000

1E Meta-analysis (Contributed) 1

James Carpenter

London School of Hygiene and Tropical Medicine

james.carpenter@lshtm.ac.uk

Medical Statistics Unit, London School of Hygience and Tropical Medicine, London WC1E 7HT

Joint authors: Gerta Rucker, Guido Schwarzer

Arcsine test for publication bias in meta-analyses with binary outcomes

In meta-analyses, it sometimes happens that smaller trials show different, often larger, treatment effects. One possible reason for such small study effects is publication bias. This is said to occur when the chance of a smaller study being published is increased if it shows a stronger effect. Assuming no other small study effects, under the null hypothesis of no publication bias, there should be no association between effect size and effect precision (e.g. inverse standard error) among the trials in a meta-analysis.

A number of tests for small study effects/publication bias have been developed. These use either a non-parametric test or a regression test for association between effect size and precision. However, when the outcome is binary, the effect is summarized by the log-risk ratio or log-odds ratio (log OR). Unfortunately, these measures are not independent of their estimated standard error. Consequently, established tests reject the null hypothesis too frequently.

We propose new tests based on the arcsine transformation, which stabilizes the variance of binomial random variables. We report results of a simulation study under the Copas model (on the log OR scale) for publication bias, which evaluates tests so far proposed in the literature. This shows that: (i) the size of one of the new tests is comparable to those of the best existing tests, including those recently published; and (ii) among such tests it has slightly greater power, especially when the effect size is small and heterogeneity is present. Arcsine tests have additional advantages that they can include trials with zero events in both arms and that they can be very easily performed using the existing software for regression tests.

2A Shape Analysis (Contributed)

Hanna Heikkinen

Univ of Oulu

hanna.heikkinen_at_oulu.fi

Dept of Math Sciences, PO Box 3000, FI-90014 Univ of Oulu, Finland

Joint authors: Kenneth Nordström

3D landmark-based shape distributions: is progress possible?

Suppose that an object can be summarized by a finite number of distinguished points known as landmarks, and suppose further that the shape of an object corresponds to invariance under the group of similarity transformations. For such shape data, an elegant and useful statistical methodology has evolved since the mid-1980s, building on the seminal work of the late David Kendall and Fred Bookstein. Indeed, this methodology, described at length in the well known book by Dryden and Mardia, has had a profound effect on the way shape is studied in various fields. Traditional multivariate morphometrics, for example, has been largely superseded by this methodology of 'geometric morphometrics', which has been said to have revolutionized quantitative morphometrics.

A striking feature of this body of methods is that it is largely tailored for planar objects. Indeed, essentially all the known probability models are for shapes of two-dimensional objects, the recommendation for three and higher dimensions being Procrustes analysis and tangent space inference. However, for three-dimensional objects (and this includes most objects!) exhibiting considerable variation in shape, no landmark-based methodology appears to be available. With the increase in number and decline in price of 3D imaging devices, this can be seen as a drawback of current landmark-based shape methodology.

In this talk we'll outline some of the issues that need to be addressed when trying to extend current shape methodology to 3D objects. In particular, we'll indicate possible approaches to deriving shape distributions for 3D data, and present some tentative results on marginal shape distributions from landmarks in three dimensions.

2A Shape Analysis (Contributed)

Arne Kovac

University of Bristol

a.kovac@bristol.ac.uk

University of Bristol, Department of Mathematics, University Walk, Bristol, BS8 1TW

Shape constraints and multiresolution

We consider various settings of the nonparametric regression and image analysis problem under shape constraints. Our overall aim is to find smooth functions with smallest number of local extreme values that satisfy a multiresolution criterion that ensures good approximation of the fitted function. The criterion requires an approximation f to be sufficiently close to the data in the sense that sums of residuals on various scales and locations are not larger than what would be expected from noise. This strategy easily leads to minimisation problems that can be very difficult to solve, hence the design of efficient algorithms is crucial.

We explore this concept in the usual regression context, but also expand it to inverse problems, estimation of parameters in differential equations, bivariate curves and online data.

2A Shape Analysis (Contributed)

Dr Marta Garcia-Finana

University of Liverpool

martaf@liv.ac.uk

Centre for Medical Statistics and Health Evaluation, Shelley's Cottage,
Brownlow Street, Liverpool L69 3GS

Unbiased estimation of geometrical parameters in Brain Research

The identification and quantification of morphological changes that occur in the brain due to neurological disease, development and ageing, is of special interest in medical research. Design-based stereological methods have been widely applied in combination with magnetic resonance imaging to estimate the volume and surface area of brain structures. In this talk, we will explore several sampling methods that are currently applied in Neurosciences (e.g. to estimate the surface area of human cerebral cortex). The question of how to predict the precision of these methods will be here described. Finally, we will present some of the latest results to predict the precision of the volume estimator and how this is applied to quantify changes of the hippocampal volume in a patient with epilepsy.

2B Statistics in Finance (Invited)

Paul Embrechts

ETH Zurich

embrecht@math.ethz.ch

Statistical extremes and Quantitative Risk Management: a difficult marriage

An overview will be given of the qualitative regulatory guidelines in use within Quantitative Risk Management (QRM) for the financial industry. Special attention will be given to the estimation of extremal events within QRM: high quantile estimation (Value-at-Risk), the modelling of extreme tail dependence,.... In particular, the current subprime crisis will be discussed from a statistical modelling point of view and some relevant current statistical research topics within QRM will be highlighted.

2B Statistics in Finance (Invited)

Anthony Ledford

AHL, Man Investments

ALedford@maninvestments.com,

Man Research Laboratory, Blue Boar Court,
9 Alfred Street, Oxford, OX1 4EH, UK

An overview of AHL: model based automated trading

AHL is part of Man Investments and is one of the world's largest managed futures style hedge-funds, managing over USD 20 billion of client capital. AHL has a 20-year history of profitably trading global financial and commodity markets using model-based, purely systematic algorithmic approaches. The trading models we research and deploy monitor live market data from around the world and make real-time trading and investment decisions based on distributional forecasts of how prices will evolve over a range of timescales. Rigorous multidisciplinary research underpins the development, implementation and monitoring of these trading systems, with statistical modelling and detailed data analysis at the heart of our activities.

I will illustrate this talk with some examples of the data we model and the evolving issues we face when doing so. I will explain some of the specific types of predictability our models aim to capture and discuss how we deploy them over a wide spectrum of traded instruments in order to construct a robust diversified portfolio. Our historical track-record will be used to provide an objective assessment of the success of our approach and will be used to illustrate how statistical reasoning informs the day-to-day management of our investment process as well us underpins our research activities.

2C Statistics and the Law (Invited)

Roberto Puch-Solis

Forensic Science Service

Trident Court, Solihull Parkway, Birmingham Business Park, Birmingham, B37 7YN

Scientific aspects of statistical evaluation of evidence

In the aftermath of a crime, material linking the crime scene with the person that committed the crime can be recovered. Usually, at this point, no one has been arrested. One of the main tasks of forensic science is that of investigation, where a forensic scientist support the police to detect a suspect. Once a person has been arrested, the role of the forensic scientist is that of evaluation: material recovered from the crime scene is compared to material recovered from the suspect. If a DNA profile has been recovered in the crime scene and another DNA profile has been obtained from a suspect, the forensic scientist compares a pair of hypotheses of the kind:

Hp: The suspect is the donor of the DNA profile
Hd: Someone else, unrelated to the suspect, is the donor of the DNA profile

DNA is extracted from organic material, which can be degraded, generally rendering partial profiles. In this case, the statistical evaluation and interaction with a criminal court can become more complex. Similar situations can arise with other evidence types such as fingerprints and glass. In this talk, aspects related to statistical variation and noisy observations will be discussed.

2C Statistics & the Law (Invited)

James Curran

Department of Statistics, University of Auckland

j.curran@auckland.ac.nz

Statistics in Forensic Science

This talk will provide an introduction to the type of problems that get statisticians involved in forensic science. It will cover the various approaches that have been made over time to answer the questions that the court is interested in with regard to forensic evidence, and will cover more recent problems that need statistical methodology.

2C Statistics & the Law (Invited)

Professor Paul Roberts

University of Nottingham

Paul.Roberts@nottingham.ac.uk

University of Nottingham, School of Law, NOTTINGHAM NG7 2RD.

Statistics, Expert Witnesses and Criminal Justice

This is going to be a general review of key issues, for a non-legal audience.

Roland Caulcutt

Caulcutt Associates

rcaulcutt@btinternet.com

59 Treffry Road, Truro, Cornwall, UK TR1 1WL

Six Sigma could be even more successful

Many organisations claim to have achieved enormous benefits from the adoption of a six sigma approach to management. However, in practice the achieved benefits vary considerably from company to company and, indeed, from project to project. Why are some organisations less successful than others?

Almost all six sigma organisations have employees with "blackbelt" as their job title. It is widely accepted that these people play a vital role as they lead the process improvement projects that are expected to yield a net gain for the organisation by producing benefits for stakeholders.

Causes of underperformance can be identified by considering the following elements of the blackbelt role within the six sigma approach:

Selection of blackbelts
Selection of projects for blackbelts
Expected benefits from blackbelt projects
Statistics in the training of blackbelts
Human dynamics in the training of blackbelts
The support of blackbelts
The monitoring of blackbelt projects.

2D Six Sigma (Contributed)

Shirley Coleman

ISRU, Newcastle University, UK

Shirley.coleman@ncl.ac.uk

Joint authors: Ron S. Kenett

Statistical Aspects of Quality by Design (QbD) Applications to Analytical Methods in the Process Industries

Process Analytical Technologies (PAT) guidelines are proving to be an opportunity for process industries to achieve substantial benefits from increased process understanding. Quality by Design (QbD) offers a methodological approach to tackle process improvement at a fundamental level. QbD is a systematic approach to development that begins with predefined objectives and emphasizes product and process understanding and process control based on sound science and quality risk management. The FDA and ICH have recently started promoting QbD in an attempt to limit rising development costs. It is also expected that QbD will help PAT guidelines reduce regulatory barriers to innovation and creativity [1, 5, 7, 8]. Process industries are characterized by having extensive multivariate data. QbD is partially based on the application of multivariate statistical methods [2, 3, 4, 7] and a statistical Design of Experiments strategy [6, 7] to aid the development of both analytical methods and new formulations. A process is well understood when all critical sources of variability are identified and explained, variability is managed by the process, and product quality attributes can be accurately and reliably predicted over the design space. In this talk we will review the basics of QbD and emphasize their impact on the development of analytical measurement methods in the process industries. Various aspects of measurement uncertainty relevant to QbD will be addressed, with case studies.

2D Six Sigma (Contributed)

Dr Carolyn Craggs

CCSL – consultancy & training

carolyn@carolyncraggs.co.uk

Wood Hill, Eggleston, Barnard Castle, DL12 0DH

Six Sigma and Pharmaceutical Manufacture

To statisticians, Six Sigma is often seen of as a 'tool kit' which comprises some straightforward statistical tools with perhaps a few other techniques such as FMEA. This presentation will give an insight into the benefits Six Sigma can achieve – and some of the challenges that can be faced in pharmaceutical manufacture.

2E Meta-analysis (Contributed) 2

Cathal Walsh

Trinity College Dublin

Cathal.walsh@tcd.ie

Dept Statistics, Trinity College Dublin, Dublin 2, IE

Joint authors: Dr Killian O'Rourke, Mater Misericordiae Hospital, University College Dublin.

Prior Elicitation and Evidence Combination

The Bayesian literature has evolved to the stage where we can answer real, ordinary, everyday questions that are raised by researchers in the health sciences, without having to devote weeks or months to the fitting of complex models; knowing we can only publish them in a dedicated statistics journal. Nonetheless, persuading clinical referees and editors of medical journals can still be a (minor) challenge.

In this presentation, we describe a situation where routine studies of modest size have been discussed with the research team and formulated as Bayesian analyses. This has been in the area of effectiveness of treatment for Multiple Sclerosis.

Prior elicitation has been done by careful identification of the model parameters, and using the knowledge of experts in the field to identify appropriate references from which to extract a family of priors.

The data available from the clinical setting has been naturally incorporated in the likelihood, examining the impact of study design by adjustment where necessary.

The experience of working with clinicians who had previously been unexposed to Bayesian thinking will be discussed.

The comments obtained from the clinical and statistical referees, before we persuaded the editor to publish the work also merits comment.

2E Meta-analysis (Contributed) 2

Claudia Lozada-Can

EPFL

claudia.lozadacan@epfl.ch

École Polytechnique Fédérale de Lausanne, Institute of Mathematics,
STAT-IMA-FSB-EPFL, Station 8, CH 1015 Lausanne, Switzerland

Joint authors: Professor Anthony Davison

Usual inferences based on maximum likelihood estimates are accurate to first order. Recent developments on likelihood inferences result in third order accuracy for continuous data, and second order accuracy for discrete outcomes. In this paper, we show an application of higher order methods to meta-analysis based on the modified likelihood root. We briefly introduce the theory of accurate likelihood inference, and we show an example of how it could be used in the context of meta-analysis.

2E Meta-analysis (Contributed) 2

Professor J.L Hutton

University of Warwick

j.l.hutton@warwick.ac.uk

Statistics Department, University of Warwick, Coventry CV4 7AL

Joint authors: Dr K Hemming

Common, severe adverse events: meta-analysis and missing data.

Vigabatrin, a licensed anti-epileptic drug, has been found to be associated with severe and asymptomatic, visual field constrictions in about 50% of those prescribed the drug. A systematic review of observational studies has identified univariate links with age, duration of use and cumulative dose using the inefficient, but common meta-regression approach to missing data: any studies with incomplete information on a particular covariate were excluded in an available case analysis. A complete case approach to joint influences of covariates would be biased and limited by a very small number of eligible studies.

To maximise the available data, we propose a joint likelihood approach, factorising the density for the likelihood of the adverse event, conditional on the various covariates, with the joint distribution for the covariates, specified as a series of conditional distributions. This joint likelihood approach assumes the covariate data are missing at random. Clinical expertise informs the prior densities. The implications for design of studies to facilitate early confirmation of suspected severe, asymptomic adverse events are considered

3A The Best Of Significance (Invited)

David Spiegelhalter

University of Cambridge and MRC Biostatistics unit.

david_spiegelhalter@yahoo.co.uk

Telling Statistical Stories

Hans Rosling has led the way in showing how a combination of a strong narrative and atttactive animated graphics can bring statistical insights to a wide audience. The Youtube era brings with it a possible audience for short statistical 'stories' that nevertheless can convey important information. We shall demonstrate some prototype examples covering games, sport, health risks and coincidences.

3A The Best Of Significance (Invited)

Michael Blastland

Beauty and the Beast: Numbers and the Media

“Good with words, c**p with numbers,” said the famous broadsheet editor, summarising the capabilities of journalists. And many of us can cite unflattering examples of their handiwork.

This session will enjoy some of the best (or worst, depending how you see it). It will also suggest ways for us to understand a little better what goes wrong, why and what to do about it. In particular, it will look at the idea of the story, which is the template for most news reporting, and compare it with the tools of understanding more likely to be used by statisticians.

Michael Blastland is a former senior BBC journalist and a graduate in English Literature. Too late in life, he says, he learnt how to count, then made a radio series about it: “More or Less”, followed by a book: “The Tiger that Isn’t – Seeing Through a World of Numbers”. However, he will argue that the clash of cultures is not defined quite as neatly as it seems.

3B Emerging Geometries for Statistical Science (Contributed)

Paul K Marriott and Karim Anaya-Izquierdo

University of Waterloo, Canada and The Open University

pmarriot@math.uwaterloo.ca K.Anaya@open.ac.uk

Waterloo, Ontario, N2L 3G1, Canada
and:
Walton Hall, Milton Keynes, MK7 6AA

Joint authors: Frank Critchley, The Open University
Paul W Vos, East Carolina University, USA

Sensitivity analysis for statistical science: some new developments exploiting computational geometry

Sensitivity analysis in statistical science studies how scientifically relevant changes in the way we formulate problems affect answers to our questions of interest. There are two aspects to such changes, associated with models and data respectively:

1. In statistical science, the complex mechanism generating the data is described by some simpler, but still realistic, model highlighting specific aspects of interest. Such models are not 'true' in any absolute sense. Rather, since 'all models are wrong, but some are useful', it makes sense to explore a range of scientifically reasonable models around the currently accepted one.

And:

2. Exploration of the data can reveal unanticipated features. These may reflect previously unknown structure, requiring us to elaborate our models. Or, again, subsets of observations having disproportionately large influence on our results, requiring reference back to the scientist to determine if they are highly informative cases meriting further study or, at the other extreme, the result of measurement or recording errors requiring down-weighting or deletion.

Since such changes in problem formulation are pertinent, sensitivity analyses are sensible.

Overall, scientific motivation is, thus, threefold:

1. Stability: If small changes have small effects, you gain the reassurance of stability.
2. Robustness: If large changes have small effects, you know your analysis is robust.
3. Warning: If small changes have large effects, you want to know about it!

New advances on the geometry of the space of models allow us to build a rigourous framework in which to investigate these problems and develop insightful computational tools, including new diagnostic measures and plots. Examples are given where the resulting sensitivity analyses indicate the need for specific model elaboration or data re-examination

3B Emerging Geometries for Statistical Science (Contributed)

Henry P Wynn

London School of Economics

h.wynn@lse.ac.uk

Houghton Street, London, WC1X 9BA

Statistical applications of computational algebraic geometry: an overview

Algebraic Statistics is the use of computational algebraic geometry in statistics. The reason for its recent success is two-fold. First, many statistical models can be described in algebraic terms, from the fitting of both univariate and multivariate polynomial models in a regression context to product-type models for categorical data. The second reason is that computational algebra has itself developed very fast so that many of these models can be better understood as geometric objects.

The talk will concentrate on 'what the technology can do for you'. A brief catalogue is as follows:
• identifiablility and aliasing problems
• tracking the equivalence between explicit and implicit models
• the geometry and enumeration of maximum likelihood solutions
• statistical simulation for exact tests and multiple tests
• limit or boundary models
• graphical and causal models.

There are also large areas which have only just begun to be studied, such as applications to asymptotics and information geometry. The talk will demonstrate the ideas via a collection of readily understood examples, together with prompts to the use of computational algebra packages.

3B Emerging Geometries for Statistical Science (Contributed)

John B Copas

University of Warwick

jbc@stats.warwick.ac.uk

Coventry CV4 7AL

Joint authors: Shinto Eguchi, ISM, Tokyo, Japan

A geometrical look at likelihood and robustness

We can think of a statistical model as a line in the space of all possible distributions – different parameter values correspond to different points on this line. The likelihood function for a given model and for a given set of data is therefore a function defined on a line.

But in practice a model is only a working model, which may well be wrong. There may well be other models which give an equally good fit to the data – geometrically we can think of these as distributions belonging to a "tubular neighbourhood" surrounding the line of our working model. How can we redefine likelihood to take account of this model uncertainty? A crucial question is the "object of inference" – a parameter only has meaning in the context of a model, and so once we depart from this model we have to be clear exactly what it is that we are trying to estimate. Practical examples are used for illustration.

3B Emerging Geometries for Statistical Science (Contributed)

Frank Critchley

The Open University

F.Critchley@open.ac.uk

Walton Hall, Milton Keynes, MK7 6AA

Emerging geometries for statistical science

Euclidean geometry is deeply embedded in statistical thinking, as is the maxim that a picture is worth a thousand words. We all know the power of good graphics, while fundamental statistical ideas – from 'between plus within' decompositions of the total sum of squares to Karl Pearson's 'On lines and planes of closest fit' – express and exploit Pythagoras' theorem, orthogonal projection providing the necessary right-angle.

At the same time, many important practical problems – such as assessing the adequacy of a generalised linear model, coping with model uncertainty, and addressing identifiability and aliasing issues – can require geometries other than Euclid's. New geometries for statistical science have been emerging in recent years to meet this challenge, enabled by a growing realisation of the potential of computational geometry, and its associated graphics, to deliver operational tools. The talks in this session review on-going progress in this area, under the sub-title:

Look what geometry could do for you!

Professor Julius Sim

Keele University

j.sim@keele.ac.uk

Room Mac2.04, Keele University, Staffordshire ST5 5BG

Joint authors: Dr Angus Dawson

Ethical implications of cluster randomized trials

In the cluster randomized trial (CRT), the unit of randomization is a group or cluster, rather than an individual as in a clinical trial using individual randomization. Within a CRT, the intervention under test may be delivered either at the level of the cluster ('cluster-cluster' CRT) or at the level of the individual ('individual-cluster' CRT).

Methodological and statistical aspects of the CRT have recently attracted considerable attention in the literature, but associated ethical implications have received less scrutiny. This paper will examine the ethical issues that arise in CRTs in relation to equipoise and consent.

Equipoise is the ethical principle that requires genuine uncertainty as to which is the superior intervention at the outset of a trial, and is normally interpreted in terms of the (non)optimality of an intervention for the individual patient. This interpretation raises difficulties in both types of CRT. In a cluster-cluster CRT, equipoise has to be interpreted in terms of the 'average' patient, and for some patients a given intervention will thus almost inevitably be a priori suboptimal. In an individual-cluster trial, there may be no available alternative intervention for members of a cluster, with a similar implication. The demands of equipoise must therefore be reconciled with the nature of the CRT, if this design is to be ethical justifiable.

Informed consent is generally seen as an ethical prerequisite for a clinical trial; such consent can be i) to the study occurring, and ii) to the receipt of a particular treatment within the trial. In a cluster-cluster CRT, however, there is little scope for patients to opt out of either the trial or the treatment to be received. It could be argued, however, that one would not normally expect to give consent for treatment policies implemented at a practice or hospital level. The difficulty of obtaining individual consent may therefore be counterbalanced by a reduced need for such consent in a cluster-cluster CRT. In an individual-cluster CRT, patients cannot meaningfully consent to the trial occurring, but can give or withhold consent to a treatment within it delivered at an individual level. However, alternative interventions may be unavailable or logistically problematic, consent from controls may induce contamination, and special problems arise with non-competent members of a cluster. Thus, individual consent to treatment is possible, but may be inert or give rise to methodological difficulties. Hence, alternatives to the traditional model of informed consent may be required in the CRT.

In conclusion, the need to reconcile the methodological and ethical demands of the CRT poses important challenges for statisticians involved in designing such studies.

3C Statistics and the Law (Contributed)

Joseph L Gastwirth

Department of Statistics, George Washington University

jlgast@gwu.edu

Department of Statistics, George Washington University, Washington DC 20052-0001 USA

Do you agree with the interpretation of a measure of relative disparity in educational funding made by the U.S. Supreme Court?

In areas where the U.S. government has substantial facilities, e.g. military base, which cannot be taxed by local authorities, the government provides the local school system with funds to support the education of children connected with the government presence. This talk describes a measure of relative disparity based on two percentiles that is used to decide whether educational funding in a state system is sufficiently uniform or equal that the state rather than the affected school districts receive federal funding. The talk will differ from the standard one as the audience will be asked to read the law and choose between the various interpretations. Then the speaker will demonstrate that a statement concerning the validity of an approximation inherent in the government's interpretation that was made by the government's attorney at the hearing is statistically inaccurate.

3C Statistics and the Law (Contributed)

Dr Daniel Ramos

Universidad Autonoma de Madrid

daniel.ramos@uam.es

Escuela Politecnica Superior, Ciudad Universitaria de Cantoblanco
Calle Francisco Tomás y Valiente, 11, 28049 - Madrid (SPAIN)

Joint authors: Dr. Joaquin Gonzalez-Rodriguez, Dr. Grzegorz Zadora, Prof. Colin G. G. Aitken.

Assessing and comparing evidence evaluation methods using information theory.

The debate about scientific procedures among forensic disciplines has highlighted the necessity for several important developments in forensic science. In particular, transparent assessment and reporting of the accuracy of a given forensic discipline is seen as critical to ensure the scientific nature of the discipline. This contribution addresses the topic of the assessment and comparison of likelihood-ratio-based evidence evaluation methods. Statistics have been successfully used in the last few years in order to help with this, and different assessment frameworks such as Tippett plots have been considered. The paper focuses on an information-theoretical method recently proposed by the authors and based on cross-entropy.

This procedure is closely related to the use of strictly proper scoring rules for the evaluation of subjective opinions from forecasters, widely used in the statistics literature. The link between the proposed information-theoretical procedure and the use of strictly proper scoring rules is shown in this contribution, and the concepts of calibration and refinement are explored from an information-theoretical point of view. The adequacy of the proposed assessment method is illustrated by experiments in likelihood ratio computation for evidence evaluation, both in forensic automatic speaker recognition and in forensic glass analysis.

3C Statistics and the Law (Contributed)

Dr Dimitrios Mavridis

The University of Edinburgh

dmavridi@staffmail.ed.ac.uk

School of Mathematics, King's Buildings, Mayfield Road, Edinburgh EH9 3JZ

Joint authors: Professor C.G.G. Aitken

Consider a population of discrete items with a categorical response. An example in forensic science is a consignment of pills, the categorization of which may be the content of pill which could be legal, ecstacy or LSD. The accurate estimation of the proportions of each category of pill seized is important in any criminal investigation but resource constraints make it impractical to examine the whole consignment.

A sequential Bayesian sampling procedure based on a Dirichlet prior and a multinomial likelihood is described. The procedure recommends stopping when the joint probability interval or ellipsoid for the parameter estimates is less than a given threshold in size, e.g., the volume of the 99% ellipsoid is less than 5% of the area of the parameter space. For trinomial data the procedure is illustrated with ternary diagrams with ellipses formed around the cumulative sample point. A comparison of results with traditional procedures shows that a considerable reduction in sample size may be achieved through use of the Bayesian procedure, especially when prior information is included.

3D Six Sigma – the Great Debate

Principle speakers will be **John Oakland** and **Tony Bendell**.

Seconders will be **Roland Caulcutt** and **Ron Kenett**.

The Chair for the debate will be John Shrouder (Chair, RSS Six Sigma Study Group)

Professional statisticians today do not effectively support Lean – Six Sigma approaches to performance improvement

Six Sigma and Lean Systems are now key components of performance improvement in many businesses and public sector organisations. Executives and senior managers in these organisations often seek the help of consultants and trainers who can guide them to successful implementation of "lean-six-sigma." Although many of these managers are aware that the use of statistics is essential in six-sigma projects, they are reluctant to turn to professional statisticians for support.

Unfortunately, statistics and statisticians have a poor reputation in the implementation of Six Sigma and Lean. For example, many practicing managers have serious concerns about the statistical content and use of statistical tools within Six Sigma approaches. Many believe that the training is too heavily biased towards statistics and statistical tools. It is necessary, therefore, to 'play down' the (essential) use of statistical tools for data analysis in many companies, and certainly the public sector, to get engagement at all – even in engineering companies.

The question is not one of how to persuade executives and managers that statisticians and the statistical tools they use are vital for effective improvement, but one of how to persuade staticians and the profession to engage with the real world of business and management – the economic pressures of today – and make their approaches understandable and digestible.

The evidence is that, at the moment, the statistics profession is just not doing this.

Objectives of the debate

To hear the arguments from highly respected Lean Six Sigma experts for and against an important issue in the application of Lean Six Sigma. The debate will consider a subject of interest to both Six Sigma practitioners and to statisticians that has elements of constructive tension and should provoke lively discussion through a substantive motion with opposing points of view from the debaters.

3E Meta-analysis (Contributed) 3

Tim Friede

Warwick Medical School, The University of Warwick

t.friede@warwick.ac.uk

Warwick Medical School, Gibbet Hill Road,
Coventry CV4 7AL

Joint authors: Beat Neuenschwander, Novartis Pharma AG, Basel
Alan Moore, Novartis Pharma AG, Basel

Evidence synthesis of adverse event data

In drug development safety analyses combining the evidence across clinical trials are increasingly important in order to detect even small risks of adverse reactions to the experimental drug under investigation. However, regulatory guidelines are vague regarding the methods to be used to summarize data across trials and simple pooling of the data is not uncommon (McEntegart 2000). Furthermore, unlike in standard meta-analysis the contrasts of interest usually cannot be estimated in every single study. Therefore for evidence synthesis in the context of indirect comparisons random effects models allowing for between-trial variation are used (see for example Spiegelhalter et al 2004, Whitehead 2002, Lumley 2002, Lu and Ades 2004). Safety events are often defined as pathological levels of blood or urine parameters with the samples routinely taken at discrete time points, say every three months, during the course of the study. Based on a discrete time-to-event model as described for example by Clayton and Hills (1993) we develop a hierarchical model synthesizing the evidence across trials. Model fitting is considered from a maximum likelihood perspective as well as from a Bayesian perspective. In a simulation study motivated by real-life data we investigate the properties of the proposed procedures.

References

Clayton D, Hills M (1993) Statistical models in epidemiology. Oxford University Press, Oxford. Chapter 4.

Lu G, Ades AE (2004) Combination of direct and indirect evidence in mixed treatment comparisons. Statistics in Medicine 23: 3105–3124.

Lumley T (2002) Network meta-analysis for indirect treatment comparisons. Statistics in Medicine 21: 2313–2324.

McEntegart DJ (2000) Pooling in integrated safety databases. Drug Information Journal 34: 495–499.

Spiegelhalter D, Abrams KR, Myles JP (2004) Bayesian approaches to clinical trials and health-care evaluation. Wiley, Chichester. Chapter 8.

Whitehead A (2002) Meta-analysis of controlled clinical trials. Wiley, Chichester. Chapter 10.

3E Meta-analysis (Contributed) 3

Dr Evangelos Kontopantelis

National Primary Care Research and Development Centre, University of Manchester

e.kontopantelis@man.ac.uk

NPCRDC, University of Manchester, Williamson building 5th floor, M13 9PL

Joint authors: Dr David Reeves

A comparison of Random Effects meta-analysis methods when study effects are non-normally distributed

Background
A primary concern for meta-analysts is heterogeneity in effect sizes across the studies included in an analysis, which can be attributed to clinical and/or methodological diversity. When heterogeneity is present, meta-analysis can be performed using a random-effects (RE) model, which assumes that the true effect sizes vary across studies. A number of different RE methods have been proposed, but almost all are based on the assumption that the true study effects are drawn from a normally distributed population. Very little is known about how the various RE methods perform when this assumption does not hold, even though this may more-often-than-not be the case.

Methods
Using simulations we investigated the performance of six RE methods using a variety of different distributions for the study effects, while also varying the number of studies and the ratio of between- to within-study variance. The impact of normal, skewed-normal, uniform, and "extreme" bimodal, beta and U-shaped distributions on DerSimonian & Laird, Maximum Likelihood, Profile Likelihood, Permutation, "Q-based" and a simple T-Test model was examined. Methods were evaluated for coverage, power, overall error (Type I plus Type II error) and confidence-interval (CI) estimation (accuracy of the CI around the mean effect).

Results
Within any given method, results were highly consistent across all types of distribution shape, including the more extreme bimodal and U-shaped distributions. If the analyst's priority is to maintain an accurate Type I error rate, then the simple t-test method outperforms all other methods. If control of the Type II error rate is equally important the Maximum Likelihood method has a slight edge. All methods except for Profile Likelihood considerably underestimated the width of the CI around the mean effect.

Conclusions
The findings importantly demonstrate that RE methods are highly robust against violations of the assumption of normally distributed effect sizes, even when study numbers are small. The DerSimonian & Laird method is the approach most commonly applied in practice, yet this method was outperformed by others on all criteria – coverage, power, overall error, and confidence interval. The optimum choice of method in any application depends upon the relative weight given to each of these criteria.

3E Meta-analysis (Contributed) 3

Dr Alex J Sutton

University of Leicester

ajs22@le.ac.uk

Department of Health Sciences University of Leicester 2nd Floor (Room 214e) Adrian Building University Road Leicester LE1 7RH

Joint authors: Dr Denise Kendrick
Dr Richard Riley

Use of Individual Participant Data (IPD) in meta-analysis of effectiveness studies

Background: While meta-analysis of individual participant data (IPD) is acknowledged as the gold-standard approach to synthesis, meta-analysis is still conducted using summary data in the vast majority of cases. The statistical benefits of IPD over summary approaches are under-researched and under-appreciated. Further, little work has been done on statistical approaches for the combined synthesis of IPD and summary data; despite IPD only being available for a proportion of the relevant studies in many instances.

Objectives: This talk will briefly review the literature on meta-analysis involving IPD. Then focus will turn to a case-study concerning the effectiveness of home safety education and the provision of safety equipment for the prevention of childhood poisoning. This review had an emphasis on examining potential interactions with socio-demographic characteristics previously shown to be associated with injury risk. IPD was only available from a proportion of the relevant studies. A number of different approaches to synthesis for this dataset are described, compared and contrasted. The relative contribution of the summary and IPD data to the estimation of model parameters is then considered.

To conclude, consideration is given to the implications the results of this extended case-study have on how meta-analyses of intervention studies should be conducted in the future. Specific consideration is given to decision making contexts including when such syntheses are used to inform economic (cost-effectiveness) decision models.

3E Meta-analysis (Contributed) 3

Dr Richard Riley

Centre for Medical Statistics & Health Evaluation, University of Liverpool

richard.riley@liv.ac.uk

Centre for Medical Statistics & Health Evaluation
Faculty of Medicine, University of Liverpool
Shelley's Cottage, Brownlow Street, Liverpool. L69 3GS

Joint authors: Mrs Susanna Dodd, Dr Jean Craig, Prof Paula Williamson

Meta-analysis of diagnostic test studies using individual patient data and aggregate data

Background
A meta-analysis of diagnostic test studies provides evidence-based results regarding the accuracy of a particular test, and usually involves synthesising aggregate data (AD) from each study, such as the two by two tables of diagnostic accuracy. A bivariate random-effects meta-analysis (BRMA) can appropriately synthesise these tables [1], and leads to clinical results such as the mean sensitivity and mean specificity across studies. However, translating such results into practice may be limited by between-study heterogeneity and that they relate to some 'average' patient across studies.

Objectives
This talk will describe how the meta-analysis of individual patient data (IPD) from diagnostic studies can lead to more clinically meaningful results tailored to the individual patient. IPD models will be introduced that extend the BRMA framework to include study-level covariates, which help explain the between-study heterogeneity, and also patient-level covariates, which allow the interaction between test accuracy and patient characteristics to be assessed. It will be shown that the inclusion of patient-level covariates requires careful separation of within-study and across-study accuracy-covariate interactions, as the latter are particularly prone to confounding. The models will be assessed through simulation, and are extended to allow IPD studies to be combined with AD studies, as IPD are not always available for all studies [2]. Application is shown to 23 studies assessing the accuracy of ear temperature for diagnosing fever in children, with 16 IPD studies and 7 AD studies. The models reveal that between-study heterogeneity is partly explained by the use of different measurement devices, and importantly there is no evidence that individual age modifies diagnostic accuracy.

[1] Chu H, Cole SR: Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. J Clin Epi 2006, 59:1331-1332.
[2] Riley RD, Lambert PC, et al. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. Stat Med 2008; 27: 1870-1893.

4A Papers from the RSS Journals (Invited)

Simon Wood

University of Bath

s.wood@bath.ac.uk; www.maths.bath.ac.uk/~sw283

Department of Mathematical Sciences, University of Bath, Claverton Down, Bath, BA2 7AY, UK

Joint authors: Mark V. Brabington,CSIRO, Hobart, Australia
Sharon L. Hedley, University of St Andrews

Soap film smoothing

Conventional smoothing methods sometimes perform badly when used to smooth data over complex domains, by smoothing inappropriately across boundary features, such as peninsulas. Solutions to this smoothing problem tend to be computationally complex, and not to provide model smooth functions which are appropriate for incorporating as components of other models, such as generalized additive models or mixed additive models. We propose a class of smoothers that are appropriate for smoothing over difficult regions of R2 which can be represented in terms of a low rank basis and one or two quadratic penalties. The key features of these smoothers are that they do not 'smooth across' boundary features, that their representation in terms of a basis and penalties allows straightforward incorporation as components of generalized additive models, mixed models and other non-standard models, that smoothness selection for these model components is straightforward to accomplish in a computationally efficient manner via generalized cross-validation, Akaike's information criterion or restricted maximum likelihood, for example, and that their low rank means that their use is computationally efficient.

4A Papers from the RSS Journals (Invited)

Byron Morgan

University of Kent, UK

B.J.T.Morgan@kent.ac.uk

Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent, CT2 7NF, UK.

Joint authors: M.S. Ridout, University of Kent

A new mixture model for capture heterogeneity

We propose a mixture of binomial and beta–binomial distributions for estimating the size of closed populations. The new mixture model is applied to several real capture–recapture data sets and is shown to provide a convenient, objective framework for model selection. The new model is compared with three alternative models in a simulation study, and the results shed light on the general performance of models in this area. The new model provides a robust flexible analysis, which automatically deals with small capture probabilities.

Kaatje Bollaerts

Hasselt University Belgium

kaatje.bollaerts@uhasselt.be

Agoralaan, Gebouw D 3590 Diepenbeek

Joint authors: Prof Dr. Marc Aerts, Dr. Stefaan Ribbens, Dr. Yves van der Stede, Ides Boone, Dr. Koen Mintiens

Identification of Salmonella high risk pig-herds in Belgium by using semiparametric quantile regression

Consumption of pork contaminated with Salmonella is an important source of human Salmonellosis worldwide. To control and prevent Salmonellosis, Belgian pig herds with high Salmonella infection burden are encouraged to take part in a control programme supporting the implementation of control measures. The Belgian government decided that only the 10% pig herds with the highest Salmonella infection burden (shortly high risk herds) can participate. To identify these herds, serological data reported as SP-ratios are collected. However, SP-ratios have an extremely skewed distribution and are heavily subject to confounding seasonal and animal age effects. Therefore, we propose to identify the 10% high risk herds by using semi-parametric quantile regression with P-splines. In particular, quantile curves of animal SP-ratios are estimated as a function of sampling time and animal age. Then, pigs are classified into low and high risk animals with high risk animals having an SP-ratio larger than the corresponding estimated upper quantile. Finally, for each herd, the number of high risk animals is calculated as well as the beta-binomial p-value reflecting the hypothesis that the Salmonella infection burden is higher in that herd compared to the other herds. The 10% pig herds with the lowest p-values are then identified as high risk herds. In addition, since high risk herds are supported to implement control measures, a risk factor analysis is conducted using binomial Generalized Linear Mixed Models to investigate factors associated with decreased or increased Salmonella infection burden. Finally, since the choice of a specific upper quantile is to a certain extent arbitrary, a sensitivity analysis is conducted comparing different choices of upper quantiles.

Robert Kunst

Institute for Advanced Studies and University of Vienna

kunst@ihs.ac.at

Institute for Advanced Studies, Stumpergasse 56, 1060 Wien (Vienna), Austria

A nonparametric test for seasonal unit roots

We consider a nonparametric test for the null of seasonal unit roots in quarterly or monthly time series under the name of RURS (for range unit root seasonal) that generalizes the idea of the RUR test by Aparicio, Escribano, and Sipols to seasonal analysis. We find that the test concept is more promising than a formalization of visual aids such as plots by period. In order to cope with the sensitivity of the original RUR test to autocorrelation under its null of a unit root, we suggest an augmentation step by autoregression. We present some evidence on the size and power of our procedure and we provide empirical illustrations by applications to a commodity price and to an unemployment rate.

AMS Subject code: 62M10, 62G10, 91B84

Idris. A Eckley

Lancaster University

I.Eckley@lancaster.ac.uk

Department of Mathematics and Statistics,
Fylde College, Lancaster University, Lancaster, UK, LA1 4YF.

Joint authors: Guy P. Nason, University of Bristol

A wavelet-based adaptive sampling regime

In this talk we introduce a wavelet-based approach which could be used to develop an efficient on-line adaptive sampling regime. The approach is based on the locally stationary wavelet (LSW) process model. The LSW modelling approach permits a location-scale decomposition of power in the form of the evolutionary wavelet spectrum. We use this spectrum to identify changes within the structure of the collected signal which can in turn trigger an increase or decrease in the sampling rate, thus leading to the adaptive sampling regime.

4B Time Series (Contributed)

Fabio Rigat

CRiSM, Department of Statistics, University of Warwick

f.rigat@warwick.ac.uk

1 Cryfield Cottage, Gibbet Hill road, CV4 7AL, Coventry

Joint authors: Jim Q. Smith

Non-parametric change-point estimation for time series models

This paper illustrates a non-parametric method to detect significant changes of the parameters of time series models. Rather than relying on a parametric specification of the coefficients' evolution through time, their dynamics are assessed within a hypothesis testing framework as a sequential non-parametric change-point problem. The Kullback-Leibler (KL) divergence between the posterior distributions of successive sets of data under the same model is proposed and justified as a test statistic. Markov chain Monte Carlo posterior simulation is used to approximate the value of the KL divergence and its critical region under the null hypothesis of no change. For exponential family models we show that the KL statistic has a closed form.

The method is illustrated using examples in finance and in the analysis of neural data using various statistical models. In finance, parameter changes signal abrupt market shifts or sudden variations among the correlations among the prices of a given set of stocks. In the analysis of neural data parameter changes are shown to mark modifications of the plastic functional connectivity of neuronal networks or pathological change in the brain activity preceding the onset of focal epileptic seizures.

Ta-Hsin Li

IBM T. J. Watson Research Center

thl@us.ibm.com

Dept. of Mathematical Sciences, IBM T. J. Watson Research Center
Yorktown Heights, NY 10589-0218, USA

A Nonlinear Method for Robust Spectral Analysis

By replacing the least-squares criterion with an Lp-norm criterion in the regression formulation of the ordinary periodogram, a new type of periodogram, called the Lp-norm periodogram, is obtained. It is shown that the Lp-norm periodogram has a relationship with the so-called fractional autocorrelation spectrum of a random process in the same way as the ordinary periodogram has with the ordinary autocorrelation spectrum (or normalized power spectrum). It is also shown that the Lp-norm periodogram can strike a balance between the robustness inherited from Lp-norm regression against heavy-tailed noise and the effectiveness like the ordinary periodogram under normal conditions. Simulation results and real-data examples are provided to demonstrate the performance of the Lp-norm periodogram for spectral analysis and signal detection.

Dr Toby Prevost

University of Cambridge

toby.prevost@phpc.cam.ac.uk

Institute of Public Health, Forvie Site, Robinson Way, Cambridge CB2 0SR

Joint authors: Professor David Spiegelhalter

Multivariate meta-analysis for modelling studies of behaviour change

This work was motivated by a randomised trial of a complex intervention aiming to change the physical activity in a sample of participants identified through primary care (reference). The intervention was based on the psychological Theory of Planned Behaviour which specifies inter-relationships between attitudes and intentions towards the behaviour, and the behaviour itself. The six variables are targeted by the professional delivering the intervention.

We identified and programmed eight meta-analysis methods for pooling the 15 pairwise correlations between the variables to assess the strength of the assumed inter-relationships in longitudinal physical activity studies. We found some support for the theory in this context.

We also found that inferences on individual pooled correlations were broadly similar whichever method was used within random- and fixed-effects classes; whether Bayesian or classical; maximum likelihood or least squares. However, methods which ignored the implicit mathematical association between multiple correlations gave misleading conclusions for contrasts and global tests that require the involvement of multiple correlations.

We extended the methods to assess compatibility of a new study with the existing meta-analysis; wider potential applications include the planning of trials of interventions that act on multiple related variables.

Reference:

Kinmonth AL, Wareham NJ, Hardeman W, Sutton S, Prevost AT, Fanshawe T, Williams K, Ekelund U, Spiegelhalter D, Griffin SJ. Efficacy of a theory-based behavioural intervention to increase physical activity in an at-risk group in primary care (ProActive UK): a randomised trial. Lancet. 2008; 371:41-48.

Stephen Walters

University of Sheffield

s.j.walters@sheffield.ac.uk

ScHARR, University of Sheffield, Regent Court, 30 Regent St, Sheffield, S1 4DA

Joint authors: Lucy Radford

Models for analysing data from individually randomised trials in primary care with clustering effects due to health professional variation

The majority of statistical analyses, in individually randomised controlled trials (iRCTs), assume that the outcomes on different patients are independent. In reality there are situations, particularly in primary care, where there is some doubt about the validity of this assumption. One example is when the intervention is delivered by a health professional (such as a General Practitioner) and a number of patients receive the intervention from each professional. The success of the intervention can depend on the professional delivering it, so that outcomes of patients treated by the same professional may be correlated or "clustered". The aim of this presentation is to investigate statistical methods of adjusting for clustering, in the context of iRCTs, where the intervention is delivered by a health professional.

There are essentially three strategies to allow for clustering:
1. Cluster level analysis – analysis is carried out at the cluster level, using aggregate summary data (such as the mean outcome per cluster).
2. Marginal or population-averaged approach, with model coefficients estimated using generalised estimating equations (GEEs).
3. Random-effects (R-E) or cluster specific approach.

In practice both Marginal and R-E models provide valid methods for the analysis of clustered data, although the two approaches lead to different interpretations of the treatment effect.

This presentation will compare and contrast the three approaches, using a variety of example data, with binary and continuous outcomes, from four iRCTs in primary care with a therapist led intervention including: specialist clinics for the treatment of venous leg ulcers; acupuncture for low back pain; homeopathy treatment for chronic fatigue syndrome; pre and post operative physiotherapy for total knee replacement

Elinor Curnow

UK Transplant

Elinor.Curnow@uktransplant.nhs.uk

UK Transplant, Fox Den Road, Stoke Gifford, Bristol, BS34 8RR

Joint authors: Professor Dave Collett

Monitoring outcomes following organ transplantation

Background

Organ transplantation is a life-enhancing and life-saving procedure with highly successful results in the UK; it is estimated that nationally 96% of adult patients are alive one year after their first deceased heartbeating donor kidney transplant, based on 2002-2005 data from the National Transplant Database. Results are similar for other organs. One-year adult patient survival following a first deceased donor heart transplant is estimated at 80% and one-year adult patient survival following a first elective deceased heartbeating donor liver transplant is estimated at 89%. Although early post-transplant patient mortality rates and graft failure rates are generally low, it is important that robust monitoring systems are in place to compare outcomes within and between transplant centres. Prospective monitoring also provides an early-warning system to enable avoidable causes of patient mortality or graft failure to be quickly identified and remedied.

Monitoring Process

Two types of cumulative sum (CUSUM) chart are used to monitor performance within individual centres. Observed – Expected (O – E) charts provide a visual comparison of observed and expected outcomes and tabular CUSUM charts are used to identify a change in the underlying mortality rate or graft failure rate. The implementation of these monitoring procedures for kidney transplantation will be described, including a description of measures to be taken when a change in the underlying failure rate has been detected.

Funnel plots are used to identify differences in performance between kidney transplant centres. In addition, cross-validated hierarchical logistic regression models with random centre effects can be used to quantify these differences. Such models also provide probabilistic statements about the extent to which any one centre's results differ from the rest. In this way, centres performing better or worse than expected can be identified.

Results and Conclusions

CUSUM charts, funnel plots and hierarchical logistic regression models provide a suitable toolkit for identifying differences in performance following organ transplantation within and between centres, enabling action to be taken if there appears to be an increase in patient mortality or graft failure rates. Moreover, the continuous monitoring of centre outcomes provides centres with information for self-assessment, and enables any change in performance to be investigated in a timely manner.

Giancarlo Manzi

MRC Biostatistics Unit

giancarlo.manzi@mrc-bsu.cam.ac.uk

MRC Biostatistics Unit, Institute of Public Health, Robinson Way, CB2 0SR, Cambridge, U.K.

Joint authors: Professor David J. Spiegelhalter, Dr Julian Flowers, Dr Rebecca M. Turner and Professor Simon G. Thompson

Combining small-area smoking prevalence estimates from multiple surveys

Combining information from multiple surveys can improve the quality of small-area estimates. Customary approaches include the multiple-frame method and the statistical matching method. However, these techniques require individual data, whereas in practice often only aggregate estimates are available. Commercial surveys usually produce aggregate estimates without clear description of the methodology used. In this context bias modelling is crucial, for which we propose a series of Bayesian hierarchical models. These allow for additive biases, which are exchangeable between small-areas within surveys, and include the possibility of estimating correlations between data sources and trends over time. Our objective is to obtain combined estimates of smoking prevalence in each of the 48 local authorities across the East of England from seven data sources, which provide smoking prevalence estimates at the local authority level, but vary by time, sample size and methodology. The estimates adjust for the biases in commercial surveys but incorporate useful information from all the sources to provide more accurate and precise estimates. Our approach is more general than other methods and uses prevalence rates rather than individual data. It provides estimates of smoking prevalence in each area, based essentially on meta-analysis of synthetic estimates, and tools to evaluate the amount of bias in each data source.

4D Gaussian Processes (Invited)

Dr Dan Cornford

NCRG, Aston University

d.cornford@aston.ac.uk

Computer Science, Aston University, Birmingham, B4 7ET

Joint authors: Dr Yuan Shen, Michael Vrettas, Dr Cedric Archambeau, Prof Manfred Opper

Gaussian process based approximate inference for diffusion processes

In this talk I will describe the variationally formulated approximate inference methods we have recently developed for diffusions. The work is motivated by the data assimilation problem in atmospheric science, which will be briefly described. In this work we approximate the posterior over the state of a dynamical system using a Gaussian process, defined by a time varying linear dynamical system. This approximating Gaussian process distribution, formulated in continuous time, is then adapted by minimising the relative entropy between the approximation and the true posterior. The algorithm will be explained and results shown. The algorithm has some limitations: in particular the marginal variance tends to be underestimated, and the approximation is poor where the Gaussian assumption is not appropriate. One solution to this is to use the approximation as a proposal distribution in a Markov Chain Monte Carlo setting. The results show faster mixing compared to a state of the art hybrid Monte Carlo method. Additionally, in the variational formulation it can be shown that the “free energy” that is minimised in the optimisation method bounds the marginal likelihood. To demonstrate the effectiveness of this bound, and illustrate some generic problems with parameter inference in diffusion processes, several examples of maximum likelihood type II estimation of the parameters in both the drift and the diffusion will be shown. The presentation will conclude with some suggestions of further directions that are currently being explored

4D Gaussian Processes (Invited)

Michael Osborne

University of Oxford

mosb@robots.ox.ac.uk

Information Engineering Building, Department of Engineering Science,
University of Oxford, Parks Road, Oxford OX1 3PJ UK

Joint authors: Stephen J. Roberts

Gaussian processes for Bayesian multi-sensor time-series prediction

We propose a powerful prediction algorithm built upon Gaussian processes (GPs). They are particularly useful for their flexibility, facilitating accurate prediction even in the absence of strong physical models.

GPs allow us to work within a completely Bayesian framework. In particular, we show how the hyperparameters of our system can be marginalised by use of Bayesian Monte Carlo, a principled method of approximate integration. We employ the error bars of the GP's prediction as a means to select only the most informative observations to store. This allows us to introduce an iterative formulation of the GP to give a dynamic, on-line algorithm. We also show how our error bars can be used to perform active data selection, allowing the GP to select where and when it should next take a measurement.

We demonstrate how our methods can be applied to multi-sensor prediction problems where data may be missing, delayed and/or correlated. In particular, we present a real network of weather sensors, Bramblemet, as a testbed for our algorithm.

4D Gaussian Processes (Invited)

Chris Williams

University of Edinburgh

ckiw@inf.ed.ac.uk

Institute for Adaptive and Neural Computation School of Informatics, University of Edinburgh 10 Crichton Street, Edinburgh EH8 9AB, UK

Joint authors: Kian Ming Chai and Edwin Bonilla

Multi-task Learning with Gaussian Processes

We consider the problem of multi-task learning, i.e. the setup where there are multiple related prediction problems (tasks), and we seek to improve predictive performance by sharing information across the different tasks. We address this problem using Gaussian process (GP) predictors, using a model that learns a shared covariance function on input-dependent features and a ``free-form'' covariance matrix that specifies inter-task similarity. We discuss the application of the method to a number of real-world problems such as compiler performance prediction and learning robot inverse dynamics.

4E Modern Approaches to Causaility (Invited)

Michael Eichler

University of Maastricht

m.eichler@ke.unimaasl.nl,

P.O.Box 616, 6200 MD Maastricht, The Netherlands

Causal learning in multivariate time series

In time series analysis, inference about cause-effect relationships among multiple time series is commonly based on the concept of Granger causality, which exploits temporal structure to achieve causal ordering of dependent variables. One major and well known problem in the application of Granger causality for the identification of causal relationships is the possible presence of latent variables that affect the measured components and thus lead to so-called spurious causalities. In this paper, we present a new graphical approach for describing and analysing Granger-causal relationships in multivariate time series that are possibly affected by latent variables. It is based on mixed graphs in which directed edges represent direct influences among the variables while dashed edges---directed or undirected---indicate associations that are induced by latent variables. We show how such representations can be used for inductive causal learning from time series and discuss the underlying assumptions and their implications for causal learning.

4E Modern Approaches to Causaility (Invited)

Vanessa Didelez

University of Bristol

vanessa.didelez@bristol.ac.uk

A decision theoretic approach to causality

This talk will discuss how causal inference can be formalised as a decision theoretic problem, and contrast it to counterfactual and structural frameworks. It will be shown how graphs, or more precisely decision diagrams, can help to formulate, understand, and check assumptions, such as “no unmeasured confounders”. The special (and closely related) cases of inference about direct/indirect effects as well as about sequential decisions will serve as illustration. In both these cases, standard or naïve adjustment for covariates will lead to biased results, while g-computation and inverse probability of treatment weighting methods are consistent.

5A Best of RSC2008
(competition winners at the 2008 Research Students Conference)

Gemma Stephenson

National Oceanography Centre, Southampton (NOCS)

gs7@soton.ac.uk

National Oceanography Centre, Southampton, University of Southampton, Waterfront Campus, European Way, Southampton S014 3ZH

Joint authors: Mr Peter Challenor

Using Derivative Information in the Statistical Analysis of Computer Models

Complex models are used in many areas, such as engineering and environmental science, to simulate the behaviour of real-world systems. One application of the models is to predict how the real-world system may behave in the future. These models are written as computer codes and referred to as simulators. The simulators are deterministic, for each time they are run with the same inputs they will produce the same output.

Complex models may take an appreciable amount of computing time to run and in this sense they are expensive to execute. Performing analyses such as sensitivity and uncertainty analysis can require many runs of the simulator and this quickly becomes impractical with a computationally expensive model. Hence an emulator, which is a statistical approximator to the simulator, can be built to provide greater efficiency. A common approach is to model the simulator by a Gaussian process model and the emulator is built based on data collected from running the simulator at a specified, small number of input points.

It is possible to obtain derivatives of model outputs, for example through the adjoint of the model. The value of learning derivatives when building emulators is being investigated to determine whether additional efficiency can be achieved.

Results from this investigation applied to the climate model, C-GOLDSTEIN, will be presented.

5A Best of RSC2008
(competition winners at the 2008 Research Students Conference)

James Miller

Department of Statistics, The University of Glasgow

jamesm@stats.gla.ac.uk

Department of Statistics, The University of Glasgow,
Glasgow. G12 8QQ

In recent times there have been many advances in the field of digital imaging. These advances allow the production of images which give much greater precision and finer detail than was previously available. One example where such advances in digital imaging technology have been used to great effect is in facial surgery. Children born with a cleft lip and/or palate generally undergo an operation to correct the cleft at 3 months and there is an interest in investigating the contrast in facial shape between these children and healthy children. Digital images of the children are taken and the challenge for statisticians is to produce more refined techniques with which to analyse the data provided by these digital images.

Much of the previous analysis on this type of data has been done using Procrustes methods to investigate the position of anatomically important landmarks. An alternative to this is to extract facial curves with clear anatomical meaning and describe the amount of bending the curves experience. The amount of bending can be represented by a single function for a two-dimensional curve and by two functions for a three-dimensional curve. Functional data analysis techniques can then be used to compare these functions of bending across and within groups of children. An interesting aside is that it is often logical to align the functions according to important anatomical landmarks before looking for differences between groups. This enables a comparison between the relative merits of performing analysis on the aligned curves as opposed to the raw curves.

5A Best of RSC2008
(competition winners at the 2008 Research Students Conference)

Ben Parker

Queen Mary, University of London

b.parker@qmul.ac.uk

School of Mathematical Sciences, Queen Mary University of London,
Mile End Road, London E1 4NS

Joint authors: S. G. Gilmour; J.Schormans

Design of Experiments for Markov Chains, or how often should we open the box?

Suppose we have a system that we can measure a fixed number of times, but at any chosen interval. Motivated by an example of probing data networks, we model this as a black box system: we can either chose to open the box or not at any time period, our aim to find out how the system evolves over time.

We use the statistical principles of design of experiments to model numerical experiments that can be designed optimally. We demonstrate how to analyse the evolution of a system as a Markov Chain, and deduce its likelihood function, and hence the Fisher information matrix. From this, numerical results provide a guide to the best design for the experiment for different values of input parameters, and we show that we can find estimators whose variance is close to the minimum variance possible. We further develop our ideas to show what happens when we take into account the effect of the observations interfering with the experiment, as would always be the case with packet probing. We present examples, and demonstrate how this could be useful to many fields, with particular reference to experiments on data networks.

Guy Abel

University of Southampton

g.j.abel@soton.ac.uk

Division of Social Statistics, School of Social Sciences, University of Southampton, Southampton, SO17 1BJ

Harmonization and Estimation of Missing Cells in International Migration Flow Tables

International migration flows may be missing, reported by the sending country, reported by the receiving country or reported by both the sending and receiving country. For the last situation in which two sources of information are possible for one particular flow, the data often do not resemble each other because of differences in definitions and data collection systems. In this presentation, a model-based approach is developed to (i) harmonize migration data and (ii) to estimate missing patterns. The first is done by adjusting data based on comparisons of data provided by both the receiving and sending countries against data that are known to be accurate and correct in terms of definition. The Expectation-Maximization (EM) algorithm is then employed to estimate cells for which no reliable reported flows exist by using covariate information drawn from migration theory. Finally, measures of variability of all estimated cells are then derived using the Supplemented EM algorithm. Recent data on international migration between countries in Northern Europe are used to illustrate the methodology. The results represent a complete table of harmonized flows that can be used by regional policy makers and social scientist alike.

Peter Congdon

QMUL

p.congdon@qmul.ac.uk,

Dept of Geography, Queen Mary, Mile End Rd, London E1 4NS

Random Effects Models for Migration Attractivities: a Bayesian Methodology

Analysis of interregional migration flows is important for economic planning and allocation of resources for housing, education and health. Existing models for interregional flows are typically based on fixed effects regression via frequentist estimation. For example, Poisson or lognormal regression models may be applied to estimate the impacts of economic variables on migration flows (Devillanova & García-Fontes, 2004) or the migrant attractivities of different areas (Fotheringham et al, 2000), while fixed effects log-linear models may be applied to estimate migrant origin, destination and interaction effects (Raymer, 2007).

By contrast, this talk considers the benefits in terms of parameterisation level and model fit of applying random effect hierarchical models to inter-area migration. In particular, a fully Bayesian approach via MCMC estimation is used to model migration push and pull factors. The approach used controls for the migration context of particular areas (e.g. the population sizes of neighbouring areas), and allows both for spatial correlation in migration push/pull patterns over areas, and correlation of push & pull factors within areas. Effective parameterisation is assessed via the method of Spiegelhalter et al (2002). An application considers migration between 354 areas of England using data from the 2001 Census.

References

Devillanova C, García-Fontes W (2004) Migration Across Spanish Provinces, Investigaciones Económicas, 28: 461-487

Fotheringham A, Champion T, Wymer C, Coombes M (2000) Measuring destination attractivity: a migration example. International Journal of Population Geography, 6:391-422

Raymer J (2007) The estimation of international migration flows: a general technique focused on the origin-destination association structure. Environment and Planning A, 39:985-995.

Spiegelhalter D, Best N, Carlin B, van der Linde, A (2002) Bayesian measures of model complexity and fit. J Roy Stat Soc B, 64: 583–639.

Beata Nowok

Population Research Centre at the University of Groningen; Netherlands Interdisciplinary Demographic Institute

nowok@nidi.nl

Netherlands Interdisciplinary Demographic Institute
P.O. Box 11650, 2502 AR The Hague, The Netherlands

Joint authors: Professor Frans Willekens

A probabilistic approach towards harmonisation of migration statistics

Inadequate and inconsistent data is a common problem in the field of migration. Its persistence led to policymakers' recognition of statistical methods as a potential tool for producing comparable data. The new Regulation on Community statistics on international migration (EC 862/2007), which obliges countries to supply harmonised statistics, provides for possibility of using estimation methods to adapt statistics based on national definitions to comply with the required one-year duration of stay definition. Harmonisation of the available flow statistics requires identifying different types of migration measures and finding relations between them. We tackle these issues using a probabilistic perspective on migratory movements. Since migration is a random event, a probabilistic approach is natural and has been widely used in modelling migration. The novelty consists in applying probability theory to harmonisation problems. In general, different migration statistics, which are always put in a discrete time framework, represent the same continuous process generating the data. They differ depending on how the data happened to be collected and how the statistics happened to be produced. We assume that migration events are generated by a homogenous Poisson process, which may be generalized by allowing migration rate to be variable. The application of a simply model amenable to probabilistic calculations, allows expressing different migration measures in terms of the underlying process and to find relations between them. We consider event data and status data. The main focus is put, however, on the time criterion used in migration definition. It refers to duration of stay following relocation, which is specified very differently among countries (e.g. three months, one year, or "permanent") and constitutes the main source of discrepancies in operationalization of migration concept in the European Union states.

Professor Peter. W. F Smith

University of Southampton

pws@soton.ac.uk

Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ

Joint authors: Dr James Raymer, Professor Jonathan J. Forster

An Overview of Statistical Issues in the Estimation of Migration Flows

Migration flows are essential for understanding population change and population planning. The obtainment of accurate, timely and comparable information on migration flows involves (i) overcoming limitations in available data sources and (ii) consolidating between various data sources. Internationally, migration data are not comparable because countries collect migration data with different definitions and with different data-collection techniques. Within a particular country, migration data may be available from surveys, censuses and population registers. Each of these sources provides partial information about migration flows over time and at various levels of disaggregation. After reviewing some of the statistical problems when estimating of migration flows cross-classified by origin and destination, we provide some possible solutions and identify areas where further work is required. Our examples are based on our recent work on estimating migration data between countries in northern Europe (Brierley et al. 2008) and between areas in England and Wales (Raymer et al. 2007). The first uses a flexible Bayesian modelling framework, which is capable of dealing with flows of varying quality and missingness to provide reliable estimates with meaningful measures of precision. The second uses a log-linear model to combine detailed data from the 2001 Census with aggregate data from the health service population register to provide estimates of more recent and detailed migration flows.

References

Raymer, J., Abel, G. and Smith, P.W.F. (2007) Combining census and registration data to estimate detailed elderly migration flows in England and Wales. Journal of the Royal Statistical Society, A, 170, 891-908.
Brierley, M.J., Forster, J.J., McDonald, J.W. and Smith, P.W.F. (2008) Bayesian estimation of migration flows. In International Migration in Europe: Data, Models and Estimates, Raymer, J. and Willekens, F., eds., pp. 149-174. Wiley: Chichester.

5C Issues in the detection, reporting & analysis of adverse drug reactions (Invited)

Munir Pirmohamed

University of Liverpool

munirp@liv.ac.uk

Department of Pharmacology, The University of Liverpool, Ashton Street, Liverpool, L69 3GE

The current burden of adverse drug reactions on the NHS

Adverse drug reactions are a common clinical problem. Until recently, however, we have had to rely on either data generated more than 20 years ago and/or US data. Over the last few years, we have conducted two major studies to determine the incidence of ADRs in NHS hospitals. These have utilised a prospective design to ensure capture of the relevant information about the adverse reactions, and elucidate factors contributing to the reactions, and their possible avoidability. The studies can be summarised as follows:

1. ADRs causing hospital admission – this showed that 6.5% of all admissions were related to ADRs, at an annual cost to the NHS of £0.5 billion per year (Pirmohamed et al, 2004; *BMJ*. 329: 15-19).
2. ADRs occurring after hospital admission – this study has just been completed, and assessed 3695 patient episodes, showing that 14.7% of patients experience one or more ADRs. Fortunately, the majority of these ADRs were mild, but nevertheless caused discomfort and affected patient quality of life.

Taken together, it can be estimated that approximately 10 800-bed NHS hospitals are currently occupied by patients with ADRs at annual (conservative) cost of about £637 million. These studies have focused on adult patients, but similar studies are currently on-going in children (as part of a programme funded by the NIHR at Alder Hey Hospital, Liverpool) where the off-label and unlicensed use of medicines is much higher than in adults.

A consistent finding from many of the studies performed to date is that approximately 70% of adverse drug reaction are potentially avoidable (shown in a systematic review by Howard et al, *BJCP*, 2007; 63: 136-147). Many ADRs are due to poor prescribing including the use of inappropriate doses or inappropriate drug combinations. Interestingly, the commonest drugs causing ADRs are amongst the oldest available to prescribers, which may reflect their high usage, but also highlights that despite the experience that we have had with these drugs, we still have not developed practices to improve their benefit-harm ratios.

The ultimate aim of our studies was to provide a baseline of the current burden of ADRs in NHS hospitals. There is now a need to develop novel intervention strategies which can be used to reduce this burden of iatrogenic disease.

5C Issues in the detection, reporting & analysis of adverse drug reactions (Invited)

Dr Lesley Wise

MHRA

Lesley.Wise@mhra.gsi.gov.uk

From signals to action: Pharmacovigilance at the MHRA

This presentation will discuss how pharmacovigilance is carried out at the MHRA, including the assessment of spontaneous reports and monitoring of literature. Issues that arise in signal detection from spontaneous reports using data mining algorithms, and the evaluation and actions following signal detection will be discussed. There will also be a short presentation on the important developments in risk management of medicines aimed at proactive pharmacovigilance

Dr Owen.D Lyne

IMSAS, University of Kent

o.d.lyne@kent.ac.uk

Institute of Mathematics, Statistics & Actuarial Science,
University of Kent, Canterbury, Kent CT2 7NF. UK

Inference from noisy data for stochastic SIR epidemics in a community of households

This paper is concerned with a stochastic model for the spread of an SIR (susceptible to infected to removed) epidemic among a closed, finite population partitioned into households.

This model permits heterogeneity of infection rates for between-household and within-household infections.

The threshold behaviour of the model is briefly outlined and methods for making statistical inferences about the parameters governing such epidemics from final outcome data are described.

In reality, such data will be subject to measurement errors – there may be both false positives and false negatives.

Inference in the presence of such noise will be rather more challenging.

A simple model for this noise is developed and illustrated on both real and simulated data.

Jarno Vanhatalo

Helsinki University of Technology

jarno.vanhatalo@tkk.fi

Helsinki University of Technology
Department of Biomedical Engineering and Computational Science (BECS)
P.O. Box 9203
FI-02015 TKK, FINLAND

Joint authors: Dr Aki Vehtari

Approximate Inference in Disease mapping with sparse log Gaussian process priors

Gaussian processes (GP) are an attractive manner to construct intensity surfaces for the purposes of spatial epidemiology. The intensity surfaces are naturally smoothed by GP, and the spatial correlations between areas can be included in an explicit and natural way into the model via a correlation function. The drawback with using GP is the computational burden of the covariance matrix calculations and memory requirements. These limit a direct implementation of GP to problems with at most a few thousand of data points. Fortunately, there are a number of sparse GP models that reduce the computational burden.

In this work we consider fully (FIC) and partially independent conditional (PIC) sparse approximations. We analyse properties and performance of FIC and PIC approximations in disease mapping problem. We also propose an additive sparse GP model that can model both long and short length-scale phenomena. This is done by adding the covariance function induced by FIC with a compact support (CS) covariance function. The FIC part is used for long and CS for short length-scales. In CS covariance function the covariance becomes exactly zero when the distance between points exceeds a certain threshold. This gives rise to naturally sparse covariance matrices that lead to savings in computational time and memory requirements. We show that the proposed model outperforms additive models with FIC and PIC approximations and is computationally feasible for data sets with over 10 000 data points.

The disease mapping model consists of Poisson likelihood and log GP prior, for which the inference is not analytically tractable. A common choice for approximate inference have been Markov chain Monte Carlo (MCMC) methods. The problem with MCMC is, however, that sampling from full posterior is exceedingly slow even for moderate size data sets and becomes unpractical for large data sets. In this work we use Laplace approximation and expectation propagation (EP) to conduct the posterior inference. We compare them to full MCMC scheme and show that they give as good results as MCMC in a lot shorter time. The comparisons of Laplace and EP approximations to MCMC are done with data sets with over 3000 data points.

M.Sc. Jaakkoo Riihimaki

Helsinki University of Technology

Jaakko.Riihimaki@tkk.fi

P.O.Box 9203, FI-02015 TKK, FINLAND

Joint authors: Dr.Tech. Aki Vehtari

Variable selection for Gaussian processes using Kullback-Leibler projections

Variable selection can be used to reduce the number of explanatory variables in a model while trying to preserve the performance of the reduced model. Selecting a subset of variables makes the model more explainable and measurement costs may be lower in the future, since only the most relevant variables need to be measured. In this work, a variable selection strategy based on the Kullback-Leibler (KL) projections is applied for Gaussian process (GP) models within a Bayesian framework. The nonparametric GP models are attractive since they allow possible nonlinear effects and implicit interactions.

To avoid a selection bias, the choice of variables is done by finding the submodel closest to the full model in the sense of the KL divergence. The hyperparameters of the full model are sampled using Markov chain Monte Carlo simulation techniques, and for each sample the projection of the full model to the lower dimensional parameter space of the submodel is computed. The variable selection method is tested for the GP in regression and binary classification cases alike, and predictive performance is compared to a model where the same relevant variables are included but the selection bias is ignored. In the classification task, the distribution of latent variables is approximated with an iterative Expectation Propagation algorithm, and the projection is done by conditioning to the latent variables. To manage the computational burden of the GP for large data sets, a sparse approximation based on a small set of pseudo-input points is used.

5E Analysis of Streaming Data (Invited)

Theodoros Tsagaris

Imperial College & Bluecrest Capital

theodoros.tsagaris05@imperial.ac.uk

Department of Mathematics, Statistics Section, Imperial College London, London SW7 2AZ

Joint authors: Niall Adams, Giovanni Montana, Ajay Jasra

Online Mean-Variance Asset Allocation

We present an online robust approach to regression for data streams. The motivation for our approach is within the context of algorithmic trading, where financial data comprising asset prices are accumulated from different sources into massive data streams. Such trading demands fast algorithms to avoid latencies occurring from computing infrastructure.

In particular, we look at a recursive version of the mean-variance portfolio optimization problem. Due to the link of this approach with least squares minimization, we propose a recursive least squares algorithm, complemented by a low rank approximation of the data, for asset allocation. The approach is evaluated against known asset allocation techniques, using spot foreign exchange data. The allocation system performs favourably in the out-of-sample period when evaluated using the Sharpe performance ratio.

5E Analysis of Streaming Data (Invited)

Dimitris K Tasoulis

Institute for Mathematical Sciences, Imperial College London

d.tasoulis@imperial.ac.uk,

53 Prince's Gate, South Kensington London SW7 2PG United Kingdom

Analysing Streaming Data

Analysing roaming data streams attracts continue interest from the data mining research community. This new data format introduced new challenges since we cannot rely any more on assumptions like the off-line availability of the data, and non-variable statistical properties. In this talk we are going to show how classical statistical analysis like regression and density estimation can be formulated to address these emerging challenges. Using appropriate methods of ``forgetting'' we can develop data driven models that adapt to the changing dynamics of the stream. Real life applications will also be demonstrated to expose the potential of the proposed methods.

6A Epidemics: Statistics and Modelling (Invited)

Dr Theodore Kypraios

University of Nottingham

theodore.kypraios@nottingham.ac.uk

School of Mathematical Sciences, University Park,
University of Nottingham, Nottingham NG7 2RD

Joint authors: Prof Philip D. O'Neill (University of Nottingham)
Dr Ben Cooper (Health Protection Agency, London)

Modelling Healthcare Associated Infections: A Bayesian Approach.

There are large knowledge gaps in both the epidemiology and population biology of major nosocomial pathogens such as methicillin-resistant Staphylococcus aureus (MRSA) and glycopeptide-resistant enterococci (GRE). We are interested in answering questions such as: what value do specific control measures have? how is transmission within a ward related with ``colonisation pressure''? what effects do different antibiotics play? Is it of material benefit to increase or decrease the frequency of the swab tests? What enables some strain to spread more rapidly than others?

Most approaches in the literature to answering questions such those listed above are based on coarse aggregations of the data. Although using aggregated data is not necessarily inappropriate, the limitations of such an approach have been well-documented in the literature. In addition, when individual-level data are available, at present most authors simply assume outcomes to be independent. First, such independence assumptions can rarely be justified; moreover, it has been shown that failing to account for such dependencies in the data will result in incorrect inferences and lead to major errors in interpretation.

Our approach is to construct biologically meaningful stochastic epidemic models to overcome unrealistic assumptions of methods which have been previously used in the literature, include real-life features and provide a better understanding of the dynamics of the spread of such major nosocomial pathogens within hospital wards. We implement Markov Chain Monte Carlo (MCMC) methods to efficiently draw inference for the model parameters which govern transmission. Moreover, the extent to which the data support specific scientific hypotheses is investigated by considering different models. Trans-dimensional MCMC algorithms are employed for Bayesian model choice. The developed methodology is illustrated by analysing highly detailed individual-level data from a hospital in Boston.

6A Epidemics: Statistics and Modelling (Invited)

Peter Neal

University of Manchester

p.neal-2@manchester.ac.uk

School of Mathematics, University of Manchester,
Alan Turing Building, Oxford Rd, Manchester M13 9PL

Joint authors: Ms Yinghui Wei

Cute furry animals: Statistical analysis of endemic diseases in rodents

The understanding of infectious disease spread is important from a public health perspective. Most analysis of infectious diseases has focussed upon epidemic outbreaks and their control. However there are many endemic diseases both within humans and animals which we would like to get a better understanding of in order to develop efficient control strategies and hopefully eventually eradicate. This talk will consider the spread of cowpox amongst rodents with the data coming from a capture-recapture experiment

Andrew Garratt

Royal Statistical Society

a.garratt@rss.org.uk

12 Errol Street, London EC1Y 8LX

Raising awareness, changing perceptions – assessing UK Statistical Authority activity as a public relations campaign

The first few months of UK Statistics Authority activity are examined from the perspective of a public relations campaign.

Measures of confidence in official statistics have come from surveys (ONS) and in-depth interviews (MORI for the Statistics Commission). While it is not unusual for such techniques to be used by the PR industry to evaluate campaign success, their cost often rules them out for frequent use to provide ongoing assessments.

Media monitoring and evaluation are well-established practices in the PR industry, given the media's key role in informing the public and its effect in influencing opinion. Increasingly, the monitoring and evaluation of the 'blogosphere' is also undertaken – accessing the views of 'stakeholders' which have been traditionally difficult or impossible to measure. Some bloggers have become significantly influential – their views being followed by opinion formers, policy makers and the media.

Internet search engines, RSS (Really Simple Syndication) technology and email alerts minimise the effort involved in monitoring web based material, especially where it changes frequently.

Results are to be presented of evaluating coverage in online media and the blogosphere in the first four months from the formal establishment of the Authority on 1 April 2008, with particular emphasis on assessing:

• the extent to which the Authority has established awareness of its existence and of its objectives
• the extent to which articles reflect key messages about the quality of official statistics and of the independence of their production from government

Dr Tiziana Leone

London School of Economics

T.Leone@lse.ac.uk

Dept Social Policy LSE Houghton St, London WC2A 2AE

Measuring maternal mortality using census data in developing countries

Despite methodological advancements in formal demography, maternal mortality in developing countries remains hard to calculate and often relies on 'guestimates'. The last few years have seen an increasing demand for more accurate estimates and in particular the emphasis on Millennium Development Goal (MDG) number five which urges reduction in maternal mortality has highlighted the importance of providing good and accurate estimates of mortality that would enable developing countries to keep track of their achievements in the area of maternal mortality. Given the lack of coverage in vital statistics most countries have relied on the sisterhood method from DHS data. However, this method is only useful to assess the extent of the problem ad not to get an accurate estimate of the maternal mortality ratio. In addition given the rarity of the event, sample surveys are not appropriate for estimating differential mortality as the estimates have high confidence intervals.

Census data has been recommended by international organisations as the best way forward to estimate maternal mortality in absence of a complete vital registration system. The aim of this paper is to investigate demographic indirect techniques and statistical smoothing modelling, for the estimation of differential maternal mortality by age and household wealth using census data from South Africa and Nicaragua. These two countries give two different settings for patterns of mortality (the former heavily affected by HIV/AIDS deaths) and household dynamics that could affect the levels of under or over-reporting of deaths. Their data will be analysed according to these patters and the relative adjustments will be undertaken. This paper is set within a wider need for methodological advancement in light of the upcoming 2010 census round.

Timothy Duke

Office for National Statistics

tim.duke@ons.gsi.gov.uk

Room: 1.156, ONS, Government Buildings, Cardiff Road, Newport Wales NP10 8XG

The Inter-Departmental Business Register (IDBR) holds comprehensive information on businesses in the UK and is used by government for sampling purposes. It is also a key data source for analysis of business activity. After a major revision of the UK Standard Industrial Classification of economic activities (SIC), the IDBR updated its business classification in January 2008 moving from SIC(2003) to SIC(2007). This was motivated by the need to adapt classifications of businesses to changes in the world economy. The revision was carried out in parallel with other European Union member states, and is consistent with the European industrial classification system, NACE Rev. 2.

The change in SIC is also relevant to government organisations that make use of the IDBR and this paper will highlight ONS's approach to the revision and comment on the latest developments and updates. In practice, the revision to the SIC means historic time series must be converted from the old industrial classification to the new one and this paper looks at the methods for construction of these series. With the dual-coding of the IDBR, conversion matrices can be created between the classifications and it is these matrices that are used to combine series from the old classification to the new one. The paper will describe the conversion matrices themselves, how they are created and applied, and the quality implications in their use for historical back series.

David Boniface

University College London

d.boniface@ucl.ac.uk

HBU, Department of Epidemiology, UCL, Brook House, 2-16 Torrington Place
London WC1E 6BT

Percentile-based modelling of trends in the obesity epidemic

Introduction
Whereas only 8.8% of men and women in England aged 18-64y were obese in 1980, 23.9% were obese in 2006. This paper considers the history of the epidemic and by constructing a model, predicts its development to 2020.

Method
Body mass index (BMI) measurements of men and women aged 18-64 were obtained from the National Survey of Heights and Weights (1980, n=7690), the Health and Lifestyle Survey (1984-85, n=5135), the Diet and Nutritional Survey of British Adults (1987, n =1917), and the Health Survey for England (1991 to 2006, average n=8400). Tukey mean-difference plots with bootstrap confidence intervals were used to examine changes in the shape of the distribution of BMI. The epidemic was modelled with standard logistic growth curves. The cumulative of the asymmetric BMI distribution in each survey was modelled in terms of parameters representing the mode and spread. In turn, these 19 sets of parameters were modelled with respect to the year of the survey. This model was used to predict the parameter values and hence future shape of the BMI distribution. Obesity rates were estimated for 2010 and 2020 with corresponding 95% confidence intervals.

Results
Near linear increases in mean BMI began around 1984. At the same time changes took place in the shape of the distribution with upper percentiles increasing more than lower percentiles. The model for the BMI distribution estimated future obesity rates of 25.0% for 2010 and 26.1% for 2020. These predicted rates for obesity prevalence were lower than other published forecasts.

Conclusions
The modelling approach developed for the BMI distribution conveniently dealt with the particular asymmetric shape and its evolution over the years 1980 – 2006 and provided an upper asymptote for the mode and spread. The technique may be generalised to other situations. The resultant forecast future values of mean BMI and obesity prevalence were lower than those of other estimates. Adjustment should be applied to allow for future increases in population age.

Wicher Bergsma

London School of Economics and Political Science

W.P.Bergsma@lse.ac.uk

Houghton Street, London WC2A 2AE

New sign correlations related to Kendall's tau and Spearman's rho for measuring arbitrary forms of association

Kendall's tau and Spearman's rho are sign correlation coefficients which measure monotonic relationships in bivariate data sets. In this paper we introduce two new coefficients, related to these, which measure arbitrary forms of association. The first one is based on "concordance" and "discordance" of four points in the plane, the latter one of six points. We propose tests of independence based on either of the new coefficients as an alternative to the chi-square test for ordinal categorical and continuous data, and argue that this leads to a significant increase in power in most practical situations

Velo Suthar

Institute for Mathematical Research, UPM, Malaysia

vsutahar@yahoo.co.uk

Institute for Mathematical Research, UPM, 43400 Serdang, Selangor, Malaysia

Joint authors: Dr Habshah Midi, Head of the Laboratory of Computational Statistics and Applied, Institute for Mathematical Research, UPM, Malaysia

Randomised response variable in logistic regression using survey data of Sindh, Pakistan

The maximum likelihood estimation of the parameter vector β in the logistic regression model, where some of the covariates are subject to randomized response is discussed. Randomized response (RR) is an interview technique that can be used when sensitive questions have to be asked and respondents are reluctant to answer directly. RR variables are described as misclassified categorical variables where conditional misclassification probabilities are known. The likelihood of the univariate logistic regression model with RR covariates is derived to obtain maximum likelihood estimates. The univariate model is revisited and is presented as a generalized linear model. Standard software can be easily adjusted to take into account the RR design. The approach is illustrated by analyzing RR data taken from a sample study in regulatory non-compliance regarding unemployment benefit

Paolo Rocchi

IBM

paolorocchi@it.ibm.com

Via Shangai 53 00144 Roma ITALY

Joint authors: Lucia Rocchi (University of Perugia)

Eclectic Authors in the Statistical Field

Nowadays statisticians see the theoretical opposition between the frequentist school and the Bayesian school, but a significant circle of authors openly holds that both the methods should be used. In this work I call ‘dualist’ writers those who ground the dual view of probability upon philosophical considerations. Dualists are kept apart form ‘eclectic’ writers who rarely enter into the philosophical nodes. The present paper illustrates a survey upon 70 books written by dualist and eclectic authors in English, French, Spanish and Italian.

6D Statistics in Higher Education (Contributed)

John McColl, Ewan Crawford

Maths, Stats & OR Network (H.E.Academy) and University of Glasgow

john@stats.gla.ac.uk, ewan@stats.gla.ac.uk

Department of Statistics, University Gardens,
University of Glasgow, Glasgow G12 8QW

Joint authors: Adrian Bowman

Technology in Statistical Education

The Maths, Stats & OR Network (part of the Higher Education Academy) has a number of projects which aim to make profitable use of appropriate technology in educational settings. This paper will discuss three of these, with a variety of illustrations.

rpanel is a recently developed R package which allows rapid development of graphical user interfaces. These can be used to good effect in teaching, to produce dynamic graphics to illustrate statistical concepts and issues.

Model Choice is an internet based system for presenting multiple choice questions to students. The current name is due to the nature of the questions in the first application which tests knowledge of probability models. What sets it apart from other systems is the tailored feedback over which some considerable care is taken. The next phase of this project aims to use the same careful approach to feedback with other question types.

The STEPS Glossary, a very popular web based glossary of statistical terms, has been available for some time. A new expanded version with greatly increased size and scope will be available shortly.

Phil Ansell

Newcastle University

p.s.ansell@ncl.ac.uk

School of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne, NE1 7RU

Development of teaching resources in statistics using the language R

In July 2005, the School of Mathematics & Statistics at Newcastle University began a three year programme to develop new statistical teaching material.

The aim of the project was to replace the use of MINITAB in the delivery of its undergraduate teaching with a more powerful, flexible and free package, namely, the statistical computing package R.
With the project coming to an end in July 2008, this paper will discuss the outcomes and will giving specific examples of the materials that have been produced.

The benefit of experience has given some insight into the advantages and pitfalls of using R in undergraduate teaching and these will be highlighted alongside some of the ways that specific problems have been overcome.

David Lucy

Lancaster University

d.lucy@lancaster.ac.uk

Department of Mathematics & Statistics, Lancaster University,
Lancaster LA1 4YF

A possible model for the statistical education of European forensic Scientists

Two main issues arose about the statistical education of forensic scientists during the June 2007 European Network of Forensic Science Institutes (ENFSI) sponsored meeting in Crakow, Poland.

1. In most European countries forensic scientists are selected from individuals who have degrees in some specialist scientific field, and consequently statistical education, where it had been a feature of the undergraduate course, was related more to the specialist subject area, and not to the demands of forensic science.

2. That the statistical sciences had two distinct roles for forensic scientists. These are:

(a) where where the scientist needs to quantify the results from laboratory experiments,
(b) and where the scientist needs to evaluate numerical evidence for presentation in court.

These two uses of statistical science are for the forensic scientist fundamentally similar, but because of paradigmatic change can be regarded in different lights. It is the second, and argueably more important, area of statistical usage, which tends to be specific to forensic science, and with which forensic scientists are least familiar.

This talk will outline the areas of statistics which are used in forensic science, and suggest possible programmes of statistical education which may be presented to newly recruited forensic scientists in Europe.

Dr Neil. H Spencer

University of Hertfordshire

N.H.Spencer@herts.ac.uk

Business School, University of Hertfordshire, de Havilland campus, Hatfield, Herts., AL10 9AB

An Adaptive, Automated, Individualised Assessment System for Introductory Statistics

This contribution concerns the use of individualised computer-assisted assessment for a first year undergraduate module taken by students studying for a degree in Marketing. For this module, it was decided that students should be assessed by a series of regular tutorial sheets that they had to complete in their own time. This was to better mirror situations in which marketing students might undertake quantitative work in future employment. Also it was hoped that it would encourage students to fully engage with the learning required at an early stage of the learning process. To overcome issues of plagiarism, it was decided each student should receive a tutorial sheet containing different data on which calculations had to be performed. In order to create these for a significant number of students, it was recognised that a computer-based system was needed. It was also recognised that if feedback was to be given to students in a timely fashion, a computer-based system of marking was necessary.

This paper describes how this system of assessment was developed and subsequently improved, covering issues of distribution of the individualised assessments, how students submit their work, how the marking is carried out and how results are fed back to students, as well as measures taken to minimise the chances of plagiarism taking place.

An important way in which the system described in this paper differs from other methods of implementing individualised assessment is in the marking of work submitted by students. In a traditional examination, marks are awarded not just for achieving the correct final answer, but for using correct methods in obtaining this answer. For instance, a marker would be able to see that a student had made a relatively small error (e.g. using a variance instead of a standard deviation in a calculation) and give some marks for getting the rest of the method correct. An automated marking system that simply compares the answer provided by a student with the correct answer cannot replicate this. The automated marking system described in this paper overcomes this deficit by not only checking whether the student's answer is correct, but also checking a range of alternative answers that might be produced by a student applying a method that is almost, but not quite, correct.

Naturally, it is impossible to anticipate all the ways in which a student might get an answer wrong, but it is possible to guess a good number. In addition, the e-mailed feedback to students encourages students to challenge the marking system. If students can demonstrate that they got their wrong answer by doing something that was deserving of some marks then the system can be programmed to additionally look for these new "wrong methods" and award marks appropriately. The challenging of marking strategies by students is not something that is generally encouraged in education. However, the author's experience of this has been that students have appreciated the openness with which the marking has been carried out.

6E Uncertainty in Complex Models (Invited)

Neil Crout

School of Biosciences, Nottingham University

neil.crout@nottingham.ac.uk

www.nottingham.ac.uk/environmental-modelling

School of Biosciences, Biology Building, University Park, Nottingham NG7 5RD

Joint authors: Andy Wood, Jim Craigon, Davide Tarsitano
University of Nottingham

Using model reduction to assess model structural uncertainty

In many areas of the natural sciences the use of mechanistic or, process based, models is ubiquitous. Typically such models attempt to represent the details of the processes at work within a system, encoding a wide range of discipline specific understanding. A feature of this type of model is that they generally have a large number of components which are inter-dependent in a complex and problem specific fashion.

While mechanistic models tend to be detailed, they are less detailed than the real systems they seek to describe, so judgements are being made about the appropriate level of detail within the process of model development. These judgements are difficult to test. Consequently it is easy for models to become over-parameterised, potentially increasing uncertainty in predictions. The work we describe is a step towards addressing these difficulties. We propose and implement a method which explores a family of simpler models obtained by replacing model variables with constants (Model Reduction by Variable Replacement). The procedure iteratively searches the simpler model formulations and compares models in terms of their ability to predict observed data, evaluated within a Bayesian framework. The results can be summarised as posterior model probabilities and replacement probabilities for individual variables which lend themselves to mechanistic interpretation. This provides powerful diagnostic information to support model development, and can identify areas of model over-parameterisation with implications for interpretation of model results.

We present our ideas with respect to several example models. In each case reduced models are identified which outperform the original full model in terms of comparisons to observations, suggesting some over-parameterisation has occurred during model development.

We argue that the proposed approach is relevant to anyone involved in the development or use of process based mathematical models, especially those where understanding is encoded via empirically based relationships.

6E Uncertainty in Complex Models (Invited)

Hugo Maruri-Aguilar

Department of Statistics, London School of Economics

H.Maruri-Aguilar@lse.ac.uk

Room B709, Columbia House, Houghton Street, London WC2A 2AE

Joint authors: Alexis Boukouvalas (Aston University) and John Paul Gosling (Sheffield University)

Designs for screening in computer experiments

Computer experiments differ to traditional experiments in that replications of the experiment give the same value and thus they add no extra information to the analysis. The methods by Morris (1991) and more recently, Campolongo et al. (2006) provide efficient design alternatives for detecting important factors in a computer experiment. I intend to review designs for screening and compare with a sequential design strategy to detect factors with linear effects.

6E Uncertainty in Complex Models (Invited)

Ian Vernon

University of Durham

i.r.vernon@durham.ac.uk

Calibrating the Universe: an Uncertainty Analysis for a Galaxy Formation Simulation

Many scientific areas now employ large physical computer simulations in order to understand complex real world systems. Such simulations contain a large number of input and output parameters and this leads to several problems of interest. A subset of the outputs can be compared with observed data of the real system, and a natural question to ask is: which choices of input parameters to the model will give rise to `acceptable' matches between the outputs and the data. This is known as Calibration and is one of the most important problems when dealing with models of this size and complexity.

This talk will describe a version of Calibration known as History Matching in the context of a Galaxy Formation simulation that attempts to model the evolution of one million galaxies from the beginning of the Universe until the present day. The process of Emulation (the statistical modelling of a deterministic function) and the method of History Matching via an Implausibility measure will be described. Communicating the results of a successful Calibration is also a non-trivial matter as this involves describing a complex shape in a high-dimensional space, so various visualisation issues and techniques will also be discussed.

7A The Statistical Challenge of Legally Induced Abortion (Contributed)

Professor Joel Brind

Baruch College, CUNY, New York

joelbrind@yahoo.com

Dept. of Natural Sciences, Box A0506,
1 Bernard Baruch Way, New York, NY 10010 USA

Statistical Issues in the literature on Breast Cancer and Abortion.

The data selection and statistical methods used in papers reporting that abortion is not a significant risk factor in breast cancer are to be discussed. [3, 13]

Under-reporting of abortions is a major issue. The numbers of women, found to have had abortions in some of the published studies are compared to what can be expected from national statistics.

National data on breast cancer incidence and abortion incidence in several countries including England & Wales, Scotland, Denmark and Sweden is used as a backdrop against which estimation errors in published studies are highlighted.

Most studies published so far concentrate on younger women due to the comparatively recent introduction of legal abortion in most western countries. Focus on the age group below age 45 where breast cancer is relatively uncommon has led to underestimation of breast cancer risks post abortion over age 45. The modern epidemic is apparent over age 45. National trends in breast cancer incidence below age 45 in the countries named are examined in relation to published findings.

The issue of left censoring, whereby abortions at an early age, before abortions are recorded, are not counted in cohort studies is to be examined in published papers. Whereas the protective effect against breast cancer of full term pregnancies at an early age is acknowledged, the additionally carcinogenic effect of early abortions could also be admitted as a relevant factor.

References

3. Carroll P. The. Breast Cancer Epidemic: Modeling and Forecasts Based on Abortion and Other Risk Factors. JPANDS Journal of American Physicians and Surgeons Vol12 Fall 2007 pp72-78
http://www.jpands.org/vol12no3/carroll.pdf
13. Brind J. Induced abortion as an independent risk factor for breast cancer: A critical review of recent studies based on prospective data. J Am Phys Surg Vol. 10, No. 4 (Winter 2005) 105-110.
http://www.jpands.org/vol10no4/brind.pdf

7A The Statistical Challenge of Legally Induced Abortion (Contributed)

Patrick Carroll

PAPRI

papriresearch@btconnect.com

35 Canonbury Road, London N1 2DG

The Demographic Impact of Legally Induced Abortion

All abortions lower the birth rate. Abortion can largely account for the decline in British fertility to below replacement level.[1,2]

Sex selective abortions raise the replacement level by increasing the ratio of male to female live births. As a consequence some countries are more seriously below replacement level. This applies to some Asian countries where selective abortions are more numerous.

In all countries, there are prospects of population decline and resulting increases in the burden of the aged in respect of pensions[2] and healthcare provision.

Abortions here contribute to a decline in family life along side a decline in marriage, and increases in births outside wedlock and single parenting. Links between these trends can be explored with reference to the British preponderance of nulliparous abortions[1].

Abortions contribute to certain trends in mortality and morbidity that are now apparent. Abortion complications are more likely in late abortions, permitted up to 24 weeks gestational age here. Post abortion women may experience: depressive illnesses, impaired fertility, infertility, premature births, infant morbidity and increased breast cancer risks. All these adverse health sequelae tend to be aggravated in cases of nulliparous abortions.

A cohort analysis of breast cancer incidence with modelling, using cumulated cohort birth rates and abortion rates as explanatory variables, can provide useful forecasts of national incidence.[3]

Yet there is supposed to be some health gain from British abortions. They are medically approved by two doctors for health reasons and termed Therapeutic, on Scottish abortion statistics. Certain cases of maternal mortality in full term pregnancies, and some suicides related to unwanted pregnancies may be averted by abortions. Ectopic pregnancies, which are usually treated by a procedure resembling an abortion, can be viewed as part of the bigger picture of abortion today.

References

1. Abortion Statistics published annually by ONS and from 2002 by the Department of Health for England & Wales and by ISD-NHS Information and Statistics Division of the National Health in Scotland for Scotland.
2. Carroll P. Assessing the Damage. The Demographic Impact on Society and Consequences for the Health of Women of the 1967 Abortion Act over 40 Years. PAPRI & The Medical Education Trust. 2007. Publication accessible on the web sites
http://home.btconnect.com/papri and www.mededtrust.org.uk
3. Carroll P. The. Breast Cancer Epidemic: Modeling and Forecasts Based on Abortion and Other Risk Factors. JPANDS Journal of American Physicians and Surgeons Vol12 Fall 2007 pp72-78
http://www.jpands.org/vol12no3/carroll.pdf

7A The Statistical Challenge of Legally Induced Abortion (Contributed)

Mika Gissler

STAKES

Mika.Gissler@stakes.fi

Lintulahdenkuja 4, P O Box 220, FI-00531, Helsinki, Finland

Mental Health post Abortion

Pregnancy associated mortality from external causes of injury and poisoning is higher after a spontaneous or induced abortion compared to a live birth. According to Finnish data, post induced abortion, suicide and homicide rates are elevated, especially for young women, aged below 25 years.[3] The phenomenon is partly explained by socio-economic status and risk factors, in common between induced abortion and deaths from external causes. Whereas British data is lacking to indicate this connection, there is more data available in Finland.

Other evidence to connect induced abortion and mental health problems is lacking for paucity of data and skewness of the available data.

Abortion - spontaneous or induced - can aggravate certain mental conditions. Post abortion mental health problems may be more often depressive, possibly related to pre-existing condition. More women in the UK are now being treated for depression, but there are no studies, if more prescriptions of anti-depressives are written for women who have had a previous abortion.

7A The Statistical Challenge of Legally Induced Abortion (Contributed)

Margaret A Eames

The Acorns Public Health Research Unit , Hatfield

meames.ph@btinternet.com

The Acorns PHRU,6 ,Honeysuckle Gardens, Hatfield, Herts, AL10 8PJ

Joint authors: Dr G. Gardner, GP, Cape Hill Medical Centre, Smethwick, W. Midlands

Counting the Cost of Abortion in England over the last 40 years: The economic effects of terminations on the NHS –evidence of post-termination premature births and other apparent side-effects.

Evidence of increased rates of premature birth among women following termination:

Two important detailed reviews of the scientific literature have been published recently on the evidence of association between induced abortion (termination) and subsequent preterm birth. In Jan 2003 Thorp6 et al reviewed 24 peer-reviewed published studies, concluding that abortion increased the risk of subsequent pre-term birth (Risk Ratios between 1.3 and 2.0). Three large (1990's) cohort studies all showed an increased risk of preterm delivery with higher risk for women who had had more than one abortion.

Similarly Rooney and Calhoun (7) published a literature review of 57 papers published between 1970 and 2003 on this subject. Of these, 49 papers reported an increased risk of premature birth following a termination, with the same dose-response effect as Thorpe noted. The authors included a measure of the degree of prematurity in their summary of papers. This showed the greatest increased risk (RR = 5.6), of extreme premature delivery (below 28 weeks gestation) was associated with 3 or more abortions. This increased risk is substantially greater than that associated with maternal age, marital status, parity or socioeconomic status.

Two other large case-control studies, including results from 10 countries in Europe (Moreau8, Ancel9) also show strong evidence of increased risk. The statistical methods used in these studies will be critiqued and discussed.

Cost to the NHS

Extremely preterm delivery is associated with high risk of brain damage or neonatal death. Preterm births lead to increased risk (15-20%) of permanent disability including cerebral palsy, sensory impairment and severe learning difficulties. The increase in premature births associated with abortion and the survival of such children inevitably has increased costs not only to the NHS, but to social services, and the education system.10 In England and Wales currently 5% of infants are born at <37 weeks gestation, (33,500 per year); 1.4% at <33 weeks; (9400 per year) and 0.4% at <28 weeks (2,700). 11,12

Ancel (9) estimates that14% of all women in the UK who give birth have had a previous induced abortion. Using this and a conservative estimate of relative risk of 1.3, gives an attributable risk of approximately 400 additional preterm infants (< 33 weeks gestation), and 230 extremely preterm infants (<28 weeks) born in the UK per year, due to abortion.6 The acute care costs to the NHS of delivering an extremely pre-term infant have been estimated at £30,000-£100,000. 10There is also the 15%-20% chance of life-long disability with life-time estimated costs approaching £1million per child (NHS negligence claim). We discuss the total costs to the NHS that pre-term births incur per year, attributable to abortions. (10) Costs from other possible side-effects of termination eg. depression and infertility will also be mentioned in discussion.

6. Thorp JM, Jr., Hartmann KE, Shadigian E. Long-term physical and psychological health consequences of induced abortion: review of the evidence. Obstet Gynecol Surv 2003;58(1):67-79.
7. Rooney B, Calhoun BC. Induced abortion and risk of later premature birth. J Am Phys Surg 2003;8:46-49.
8. Moreau C, Kaminski M, Ancel PY, Bouyer J, Escande B, Thiriez G, et al. Previous induced abortions and the risk of very preterm delivery: results of the EPIPAGE study. Bjog 2005;112(4):430-7.
9. Ancel PY, Lelong N, Papiernik E, Saurel-Cubizolles MJ, Kaminski M. History of induced abortion as a risk factor for preterm birth in European countries: results of the EUROPOP survey. Hum Reprod 2004;19(3):734-40.
10. Calhoun BC, Shadigian E, Rooney B. Cost consequences of induced abortion as an attributable risk fir preterm birth and impact on informed consent. J Reprod Med 2007;52:929-37.
11. Moser K, Hilder L. Assessing quality of NHS numbers for babies' data and providing gestational age statistics. Health Statistics Quarterly 2008;37:15-23.
12. ONS Birth statistics 2005.

Chris Cannings

University of Sheffield

c.cannings@shef.ac.uk

School of Medicine, University of Sheffield, Beech Hill Road, Sheffield S10 2RX

Joint authors: Mr Orstis Chrysafis

Preferential Attachment Models for Network Growth.

There has been much recent interest in the growth of networks, particularly with reference to the WWW, but also in various other social and in biological contexts. The Preferential Attachment model (due to Simon(1954) but usually accredited to Barabasi and Albert(0)) envisages new nodes being added to an existing network one at a time, and each new node is then linked to some m pre-existing nodes which are chosen with probabilities proportional to their degree. Such a process leads to a power law distribution for the degree distribution (Barabasi and Riorden, Jordan).

Here we explore certain features of the above model and introduce a more general class of models in which nodes have defined strengths, the choice of links depends on those strengths and there is some percolation of strength within the network. We present some analytic results concerning the degree distribution and neighbourhood correlations, together with the output of some simulations.

Refs.

Barabasi AL & Albert R (1999) Emergence of scaling in random networks. Science, 286, 509
Bollobas B, Riordan J, Spenser J & Tusnady G (2001) The degree sequence of a scale-free random graph process. Random Struct.Alg. 16, 279.
Jordan J (2006) The degree sequence and spectra of scale-free random graphs. Random Struct.Alg. 29,226.
Simon HA (1955) On a class of skew distribution functions. Biometrika, 42, 425

Rowland R Kao

University of Glasgow

Faculty of Veterinary Medicine

r.kao@vet.gla.ac.uk

464 Bearsden Rd, Glasgow, G61 1QH

Network approaches to analyzing epidemics in well-characterised structured populations.

Social network approaches to modelling infectious disease spread has provided new insight into understanding how epidemics spread and how to better control them. The structure of the British livestock industry is uniquely well-described, with the movements of cattle, pigs and sheep recorded on a daily basis, in the case of cattle down to the individual bovine level. Combined with disease notifications, these spatio-temporal data offer the opportunity to test standard network theory and develop new perspectives on the potential for spread of diseases such as foot-and-mouth disease, bovine tuberculosis, and scrapie. Not only are these questions of practical importance, but also the questions raised by the challenge of analysing real datasets have also led us to develop new basic results on the importance of dynamics of networks to the spread of diseases on networks.

7C Surveying Children -Issues and Solutions (Invited)

Diane Beddoes

Office for Public Management

dbeddoes@opm.co.uk

OPM, 252b Gray's Inn Road,

Joint authors: Nina Mguni, Annie Hedges

Surveying young children/ collecting data from young people

Young people's views on the services and neighbourhoods in which they live are being sought on an increasingly frequent basis. Using a variety of methods, from surveys to qualitative research to fun events, and channels such as youth councils and parliaments, young people can contribute their perspective on services from housing to health, youth justice and drugs services.

In this paper, we look briefly at some of the reasons for this growing importance attributed to young people's views, including policy drivers, changes in service configuration and young people themselves. The main section of the paper will draw on some case studies, based on our own practice, that highlight some of the more – and less – effective ways of involving young people in research. Lastly, we argue that whilst effective research is essential, the capacity of organisations to respond to research findings and develop innovative ways of building on these are also crucial.

7C Surveying Children -Issues and Solutions (Invited)

John Flatley

Home Office

john.flatley@homeoffice.gsi.gov.uk

5th Floor Peel, 2 Marsham Street, London SW1P 4DF

Joint authors: Krista Jansson

Interviewing children in the British Crime Survey

The British Crime Survey (BCS) is one of the main sources of information about the extent and trends in crime in England and Wales. It also provides information about people's perceptions and attitudes to crime-related issues such as anti-social behaviour, police and the criminal justice system.

In 2007, the Home Office commissioned methodological work to examine the feasibility of extending the survey to children. The work concluded that this is feasible, and should be done by interviewing children aged 10 to 15 in households selected to take part in the main survey. The Home Office is now planning to extend the survey to under 16s from January 2009, following a consultation and further developmental and piloting work.

This paper discusses the programme of work carried out to develop the under 16s survey, which includes: exploring children's understanding of language related to crime and crime-related topics through qualitative work and question testing; examining children's experiences of crimes and the nature of crimes committed against them; assessing how the interviews would be best administered in field.

7C Surveying Children -Issues and Solutions (Invited)

Jenny Graham

National Centre for Social Research

j.graham@natcen.ac.uk

35 Northampton Square, London EC1V 0AX

Joint authors: Rachel Ormston, Research Director, Scottish Centre for Social Research

Children's views on ethical issues in survey research

Recent years have seen a big increase in research with children, accompanied by numerous research textbooks and guidelines on this topic. Many of these include discussions about the ethics of conducting research with young participants. However, somewhat ironically, relatively little of what has been written is informed by the views of children themselves about the research process. Moreover, much of the literature on the ethics of research with children focuses on qualitative research. Our presentation discusses findings from a NatCen-funded study which sought to address these gaps, by exploring the views of children on the ethics of conducting survey research with them. Drawing on focus groups with children aged seven to fifteen, it will focus particularly on children's views on two issues commonly held to differentiate research with children from research with adults:

- informed consent (and who should be involved in the consent decision), and
- confidentiality and disclosure.

We hope the presentation will contribute to an ongoing debate about ethics and best practice in conducting survey research with children.

7D Process Analytical Technologies (Invited)

Dr Theodora Kourti

GlaxoSmithKline

PAT and Multivariate Process Performance Monitoring: The Path To Real Time Release

The term Process Analytical Technology (PAT) is often taken to be synonymous with the use of real time analysers. However, the 2004 guidance from the Food and Drug Administration, gives PAT a mandate much wider than real time measurements. In this wider definition PAT is linked to process control, real time process signature monitoring, process understanding and correct technology transfer. Multivariate Analysis has played an integral part for the real time development of process analytical measurements and it is ready to face the demands of PAT in its wider definition. This session discusses the methods as well as the data quality requirements that will lead to real time release. It will be shown that from determining the acceptability of raw material entering the plant to ensuring quality of the product that leaves the plant, the multivariate analysis philosophy should govern all the operations that take the raw material and convert it to a final product in a cost efficient way, while meeting safety and environmental constraints.

7D Process Analytical Technologies (Invited)

Professor Julian Morris

Newcastle University

Julian.morris@ncl.ac.uk

Centre for process Analytics and Control Technologies (CPACT)
School of Chemical Engineering & Advanced Materials

Joint authors: Dr Zeng-ping Chen

Process Analytical Technologies and Quality by Design – a Process Systems Engineering View

On-line real-time PAT technologies such as NIR, MIR, UV-Vis, Raman, X-ray diffraction, etc., have been quite widely applied in chemical process monitoring but to a lesser extent in pharmaceuticals. However, unlike in off-line assays, in-situ on-line real-time spectroscopic measurements are almost inevitably subjected to fluctuations/variations in process and instrumental variables. The accuracy and reliability of multivariate calibration models for processes which are subject to variations in physical properties (such as particle size, samples' compactness, surface topology, etc) as well as process variables are increasingly becoming a matter of much concern. For example, the variation in the optical path-length materializing from the physical differences between samples may result in multiplicative light scattering which influences spectra in a nonlinear manner and hence lead to the poor calibration performance. This makes the task of extracting the relevant chemical information, and ultimately reliable process understanding for process modelling, control and optimisation, from spectroscopic measurements well beyond being routine.
A number of other issues also arise which can challenge the application of on-line real-time PAT and Multivariate Statistical Process Control (MSPC) based process performance monitoring. For example, most applications of Calibration models and MSPC have tended to focus upon the manufacture of a single product in a single vessel, i.e. a single formulation, grade or recipe, etc. with separate models being developed to monitor individual product types and product re-formulations. However, with process manufacturing trends being influenced by customer demands and the drive for product diversification, there has been an increase in flexible manufacturing. Thus with many companies now producing a wide variety of products as well as , there is a real need for process models which allow a range of products, grades or recipes to be monitored using a single process representation. Additionally, process dynamics can have an impact on the sensitivity of the performance monitoring charts with increased false alarm rates. These and other issues will be discussed and solutions proposed through presentations of industrial process manufacturing applications.

7D Process Analytical Technologies (Invited)

Dave Burnham

SAS Institute

SAS Institute, Wittington House, Henley Road, Medmenham, Marlow, SL7 2EB, UK

Role of Statistical Thinking in Design, Development and Manufacture of Pharmaceutical

A lean and simple approach to increasing process understanding that aligns scientific and engineering thinking with statistical principles is presented. This allows a larger community of scientific and engineering users to apply statistical methods to increase process understanding based on data and use this knowledge to increase process understanding based on data and use this knowledge to increase product quality.

7E High-Dimensional Data (Contributed)

Dr Casper Albers

The Open University

c.j.albers@open.ac.uk

Department of Mathematics and Statistics
Walton Hall
Milton Keynes MK7 6AA

Joint authors: Dr. Catriona M. Queen

A Bayesian graphical dynamic approach to forecasting and monitoring road traffic flow networks

Congestion is a major problem on many roads worldwide. Many roads now have induction loops implanted into the road surface providing real-time traffic flow data. These data can be used in a traffic management system to monitor current traffic flows in a network so that traffic can be directed and managed efficiently. Reliable short-term forecasting and monitoring models of traffic flows are crucial for the success of any traffic management system.

Traffic flow data are invariably multivariate so that the flows of traffic upstream and downstream of a particular data collection site S in the network are very informative about the flows at site S. Despite this, most of the short-term forecasting models of traffic flows are univariate and consider the flow at site S in isolation. In this paper we use a Bayesian graphical dynamic model (GDM) for forecasting traffic flow. A GDM is a multivariate model which uses a graph in which the nodes represent flows at the various data collection sites, and the links between nodes represent the conditional independence and causal structure between flows at different sites. All computation in GDMs is performed locally, so that model computation is always simple, even for arbitrarily complex road networks. This allows the model to work in real-time, as required by any traffic management system. GDMs are also non-stationary and can readily accommodate changes in traffic flows. This is an essential property for any model for use with traffic management systems where series often exhibit temporary changes due to congestion or accidents, for example. Finally, GDMs are often easily interpretable by non-statisticians, making them easy-to-use and understand.

The paper will focus on the problem of forecasting and monitoring traffic flows in two separate busy motorway networks in the UK.

7E High-Dimensional Data (Contributed)

Dr John Aston

Warwick University and Academia Sinica

j.a.d.aston@warwick.ac.uk

Dept of Statistics, University of Warwick, Coventry

Joint authors: Mr Michael Jyh-Ying Peng

Change Point Detection in fMRI through the use of algorithms to find pattern distributions.

Functional Magnetic Resonance Imaging (fMRI) is routinely used to assess brain activation in psychological and neurological experiments. Often the times of change of these brain states are of interest, but only changes that are sustained for at least a certain time should be considered. This paper will explore methods to detect unknown change-points occurring in fMRI time series when using Hidden Markov Models (HMMs) to account for the differing underlying brain states. The method will take into account some of the biological constraints on the speed and duration of changes, as well as allow for an unknown number of changes to occur. The methods will allow fast analysis of change points without the need to use sampling methods, a necessity when analyzing the large quantities of data present in routine fMRI studies.

7E High-Dimensional Data (Contributed)

Chris Harbron

AstraZeneca

Chris.Harbron@AstraZeneca.Com

AstraZeneca, Mereside, Alderley Park, Macclesfield,
Cheshire SK10 4TG

A Flexible Probe Level Approach to Improving the Quality and Relevance of Affymetrix Microarray Data

Microarrays have developed into a powerful tool in understanding biological processes at a fundamental molecular level through being able to simultaneously measure the expression levels of many thousands of genes. Unfortunately this great strength also gives rise to the greatest weakness of microarray technology, that of identifying false positives through multiple testing. This has led to many claims of discoveries which have failed to be validated in independent samples, leading critics to question the practical utility of microarray technology.

In a set of samples from a specific tissue, many genes will be unexpressed throughout and so will effectively just be contributing noise and false positives to any further analysis. An approach which would eliminate these unexpressed genes, whilst still retaining all the information containing genes would clearly be of great value. Various approaches have been applied, but they are limited by the use of arbitrary cut-offs and unrealistic simplifications about the behaviour of the technology and the underlying biology. In this presentation we will present an approach using Principal Component Analysis to identify those genes demonstrating consistent behaviour across the set of probes within an Affymetrix probeset, and demonstrate that restricting analyses to this set of consistent probes decreases the false discovery rate.

Graeme Chamberlin

Office for National Statistics

Graeme.chamberlin@ons.gov.uk

1 Myddelton Street, London, EC1R 1UW

Data uncertainty and economic policy

There is a well-known trade-off between timeliness and accuracy in economic statistics. Early vintages of data such as Gross Domestic Product (GDP) are based on limited information which are subsequently revised to reflect the availability of more complete survey data and methodological improvements in the way the economy is measured.

The Bank of England refer to these revisions as 'data uncertainty', and identify it as a pertinent risk in the operation of monetary policy. This paper sets out to give an overview of the issues and identify how large the costs of data uncertainty are to economic policy-makers in terms of policy regret, i.e. how would policy have changed if the final data were known at the time of preliminary estimates. The paper also considers how a better accuracy-timeliness trade-off may be achieved.

8A Best of YSM 2008 (competition winners at the 2008 Young Statisticians Meeting)

Mark Kelly

SEWTU, Cardiff University

kellymj1@cf.ac.uk

7th Floor, Neuadd Meirionnydd, Heath Park, Cardiff University, CF14 4YS

Joint authors: Gillespie D., Nuttall, J. and Hood, K.

A comparison of different approaches to modelling patient diary information

Introduction: Patient diaries are a useful and cost-effective way of obtaining longitudinal data on individuals, sparing respondents the inconvenience of being repeatedly followed up and allowing them to fill in the diary in their own time. There are numerous analysis methods possible with such information. Patient symptom diaries from a multi-country observational study of antibiotic prescribing will be used to illustrate some of the different methods possible.

Aims: To compare the results of different statistical approaches for analysing symptom diary data.

Method: Two main approaches will be investigated. First, simple summaries of the diary data symptom scores will be fitted using a hierarchical model (including area under the curve and symptom half-life). Secondly, the individual day scores themselves will be modelled as repeated measures data, in a hierarchical analysis. Different correlation structures will be fitted. The results from both of these methods will be compared and contrasted.

Results: Both approaches address different research questions and each has advantages and disadvantages. Broadly speaking, both approaches provide similar results. The increased simplicity of modelling symptom summaries is paid for in terms of the hypotheses that it can address.

Conclusion: The choice of statistical model to use when investigating diary data is not straightforward, and will almost certainly depend on the research question to be addressed.

8A Best of YSM 2008 (competition winners at the 2008 Young Statisticians Meeting)

Dr Daniel Farewell

Cardiff University

farewelld@cf.ac.uk

4th floor, Neuadd Meirionnydd, Heath Park, Cardiff CF14 4YS

Joint authors: Vern Farewell

In a large Welsh study of neighbourhood social cohesion, the analysis of multilevel questionnaire data focussed on variance components. In general, inference about the sampling distribution of estimates of variance components in multilevel models is difficult. These estimates tend to arise from numerical optimisation routines, rather than as closed-form functions of the data, so their properties are not immediately apparent. Use of bootstrap methodology provides a possible approach to inference, but was computationally demanding for the Welsh study. We present a novel method of inference that mimics certain features of a bootstrap, while avoiding the computational expense of refitting the multilevel model for each bootstrap realisation.

In brief, we explicitly express the updating step of an EM algorithm for the variance components as a function of the posterior second moments of the random effects. Provided the data may be partitioned into independent groups, we can imitate a non-parametric bootstrap by sampling (with replacement) from this set of posterior second moments. This replicate then updates the bootstrap maximum likelihood estimates, a procedure that is computationally trivial and so may be repeated as many times as desired for bootstrap inference. The approximation is improved by adjusting for the rate of convergence of the EM algorithm, and is over 300 times faster than full bootstrap inference.

Mr Steffen Unkel

Department of Mathematics and Statistics,

The Open University, Milton Keynes, UK.

S.Unkel@open.ac.uk

The Open University, Faculty of Mathematics, Computing & Technology, Department of Mathematics and Statistics, Walton Hall, Milton Keynes, MK7 6AA.

Joint authors: Dr. Abdel Hannachi
Dr. Nickolay T. Trendafilov

Independent Factor Analysis of Climate Data

Climate anomalies are studied by means of the Independent Factor Analysis (IFA) model. The IFA is viewed as a method of factor rotation in Exploratory Factor Analysis (EFA). First, by considering EFA as a new form of data matrix decomposition, estimates for all EFA model parameters are obtained simultaneously. Based on this initial EFA solution, an orthogonal rotation matrix is sought that maximizes the independence between the common factors. To compensate for the rotation of the scores, the initial loading matrix is rotated as well.

The proposed approach is applied to winter monthly sea-level pressure data over the Northern Hemisphere. The North Atlantic Oscillation (NAO), the North Pacific Oscillation (NPO), and the East Atlantic/Western Russia (EAWR) pattern are identified among the rotated spatial patterns with a physically interpretable structure.

Philip Jonathan

Shell Technology Centre Thornton

philip.jonathan@shell.com

PO Box 1, Chester, CH1 3SH

Joint authors: Kevin Ewans

Modelling the seasonality of extreme waves in the Gulf of Mexico

Statistics of storm peaks over threshold depend typically on a number of covariates including location, season and storm direction. Here, a non-homogeneous Poisson model is adopted to characterise storm peak events with respect to season for two Gulf of Mexico locations. The behaviour of storm peak significant wave height over threshold is characterised using a Fourier generalised Pareto model, the parameters of which vary smoothly with season using a Fourier form. The rate of occurrence of storm peaks is also modelled using a Poisson model with rate varying with season. A seasonally varying extreme value threshold is estimated independently. The degree of smoothness of extreme value shape and scale , and Poisson rate, with seaon, is regulated by roughness-penalised maximum likelihood, the optimal value of roughness selected by cross-validation. Despite the fact that only the peak significant wave height event for each storm is used for modelling , the influence of the whole period of a storm on design extremes for any seasonal interval is modelled using the concept of storm dissipation, providing a consistent means to estimate design criteria for abitary seasonal intervals. Characteristics of the 100-year storm peak significant wave height, estimated using the seasonal model , are examined and compared to those estimated using the seasonal model, are examined and compared to those estimated ignoring seasonality.

Aki Niemi

University of Jyväskylä & Finnish Forest Research Institute

aki.niemi@maths.jyu.fi

Finnish Forest Research Institute, Vantaa Research Unit,
P.O. Box 18 (Jokiniemenkuja 1), FIN-01301 Vantaa, FINLAND

Joint authors: Prof. Erkki Tomppo (Finnish Forest Research Institute)
Prof. Antti Penttinen (University of Jyväskylä)

Spatial point process modeling of unobserved trees using airborne LiDAR data

Estimation of volume, growth and quality of growing stock are classical forest inventory tasks. Modern forest monitoring systems, like the Finnish National Forest Inventory (FNFI), combine field sample plot data and information from supplementary sources such as satellite images and digital maps. Currently, a lot of research effort is put into development of inventory methods employing LiDAR (Light Detection and Ranging) measurements.

LiDAR is laser-based technology which in forest mensuration is used for airborne 3D measurement of trees. For upper canopy layer, i.e. big trees, high-pulse LiDAR measurements give tree locations and heights with high accuracy. However, most of the small trees will not be detected since the pulses return from the branches of the bigger trees before reaching the small trees. Hence, the effect of unobserved trees on the forest resource estimates needs to be corrected when LiDAR data is employed.

This talk presents an approach with spatial point process modeling for tackling the problem of missing trees when using LiDAR data.

The two key aspects in the research are: 1) modeling of the occurrence and locations of the small trees conditional on the observed bigger trees, and as a function of tree and stand variables, and 2) estimation of the observation probabilities of small trees when using LiDAR data and observed spatial patterns of big trees and modeled spatial pattern of small trees.

Based on modeling at a single-tree level, we develop methods for estimating the total number, volume and other forest parameters on aggregate level. We apply our methods to extensive LiDAR and field data sets measured in FNFI in northern Finland.

Keywords: forest inventory, inclusion probability, Light Detection and Ranging (LiDAR), missing data, spatial point process.

8C Methodological Developments for the 2011 Census (Invited)

Heather Wagstaff

Office for National Statistics

Heather.Wagstaff@ons.gov.uk

Office for National Statistics, Segensworth Road, Titchfield, Fareham, Hampshire, PO15 5RR

Joint authors: Dr. Steven Rogers and Mr. James Danielis

Census Validation: Edit and Imputation Methods

Many users of Census data consider quality to be purely a function of the volume of error in the data: if the 2011 Census data are perfectly accurate then the data will be of very high quality. A key aim of any edit and imputation method is to provide a measure of quality by validating and correcting for missing or erroneous responses in the data. In striving to achieve high quality outputs from the 2011 UK Census, the Office for National Statistics (ONS) has developed an Edit and Imputation Strategy which aims to deliver a database of complete and consistent records.

The design of the 2011 Census presents a number of technical challenges for editing and imputation which make the strategy necessarily complex. These include the use of electronic data capture techniques and internet data collection. To deal with these and other methodological challenges, we start by reviewing the lessons that were learned from the 2001 Census. We then develop and implement methodological best practice by using generalised editing and imputation tools. Early indications suggest our approach will lead to a process that is reliable, efficient and will yield corrected data of a very high quality.

8C Methodological Developments for the 2011 Census (Invited)

Ruth Wallis

ONS

Ruth.Wallis@ons.gov.uk

Office for National Statistics, Segensworth Road, Titchfield, Fareham,

Joint authors: Katie Draper, Jayne Mathias

New challenges in the Census questionnaire development

In 2011 we aim to achieve two firsts in the development of the questionnaire for the England and Wales Population Census. It will be the first where the Census questionnaire is available on the internet and the first Census questionnaire to undergo a dual language development. This paper examines some of the methodological challenges met to date in progress towards the Census dress rehearsal questionnaires.

The development and testing of an internet questionnaire has presented a range of methodological challenges. Underlying all of these is the question of improving data quality versus introducing mode effects. The internet provides a wealth of ways to improve data quality, with the use of technology to guide responses, to increase routing functionality and reduce respondent burden. However every functionality 'improvement' could result in a mode effect - a different response to that which would have been provided on the usual Census paper self-complete post-back questionnaire. This paper examines some of the guidelines that the ONS has developed to address the issue.

Dual language development methodology has been used by other countries for their Population Censuses, but previously, for the England and Wales Censuses, the English questionnaires have been only translated into Welsh. For 2011 a duel- development approach has been adopted for the Welsh language questionnaires. This paper compares the dual-development method with translation and illustrates some of the advantages and technical challenges raised.

Key words: questionnaire design / dual language development / translation / web questionnaire design / data quality / mode effects

8C Methodological Developments for the 2011 Census (Invited)

Keith Spicer

Office for National Statistics

ONS, 1 Drummond Gate, Pimlico, London, SW1V 2QQ

Joint authors: Jane Longhurst & Caroline Young, ONS
Natalie Shlomo, University of Southampton

Developing a Statistical Disclosure Control strategy for the 2011 UK Census

This paper discusses the development of a Statistical Disclosure Control (SDC) strategy for the 2011 UK Census. The 2011 Census SDC Policy was agreed in November 2006; The Registrars General of Scotland, England and Wales and Northern Ireland have agreed to aim for a common SDC methodology for 2011 Census outputs.

A high level analysis was undertaken in 2007 to address the advantages and disadvantages of a wide range of SDC methods in terms of the protection they afford together with their impact on the integrity of the data. Quantitative work is now being carried out to assess three short-listed methods: Record Swapping, Over-imputation and the (Australian Bureau of Statistics) Cell Perturbation method. The 2001 approach (Record Swapping with Small Cell Adjustment) is being used as a benchmark. We discuss the methodology behind these approaches and present some initial results on risk and utility applied to a variety of census tables.

8C Methodological Developments for the 2011 Census (Invited)

Owen Abbott

Office for National Statistics

owen.abbott@ons.gov.uk

Room 4200N, ONS, Segensworth Road, Titchfield, Fareham PO15 5RR

The quality of census estimates for different patterns of undercoverage

The 2011 Census estimates will be based upon the population counted by the Census, and the assessment of coverage through the Census Coverage Survey (as it was through the 2001 One Number Census). The basic methodology to estimate the population uses Dual System Estimation in the sample together with a stratified ratio type estimator to generalise to the non-sampled areas.

The resulting quality of the estimates depends on a number of factors, arguably the most important of which is the census response rate – the closer this is to 100% the smaller the adjustments for undercoverage. However, it is not just the headline Census response rate that is important – it is its variability at sub-national levels (e.g. across Local Authorities) and also the variability of census response across small areas (e.g. output areas) within those sub-national levels. Since a ratio estimator is being applied, the variability in census response within the stratum is linked to the variance and therefore the confidence intervals around the census estimates.

This presentation reports on a series of simulation studies that demonstrate the impact of different levels of variability of census response on the width of confidence intervals for sub-national areas. Given that the aim of the 2011 Census is to maximise the quality of the estimates, the results provide some challenges for the 2011 Census field operation.

8D Recent Advances in Statistical Signal Processing (Invited)

Ed Bullmore

University of Cambridge

etb23@cam.ac.uk

Wavelets and scale invariance in analysis of functional neuroimaging time series

Human brain function is often measured as a multivariate time series recorded using magnetocencephalography (MEG) or functional magnetic resonance imaging (fMRI). In general, these time series have power law spectra or long memory properties favouring their analysis in the wavelet domain. Several wavelet based techniques for estimation, hypothesis testing and resampling have been described in relation to human fMRI and MEG. Here I will focus especially on wavelet correlation as a measure of frequency band-specific association between two time series, showing how the scaling of wavelet correlations is theoretically predictable for a pair of long memory processes. I will also discuss how undirected graphs or complex networks can be inferred from wavelet correlation matrices and the topological properties of these human brain functional networks demonstrate scale invariance or a fractal small-world organization.

8D Recent Advances in Statistical Signal Processing (Invited)

Robert Nowak

University of Wisconsin-Madison

nowak@engr.wisc.edu

Compressed Sensing and Wireless Communications

Compressed Sensing (CS) refers to the problem of recovering an unknown p-dimensional vector from n << p observations. General solutions to such underdetermined systems are, of course, impossible. However, if the unknown vector happens to be sparse (i.e., has a small number of non-zero elements), then in many situations exact recovery is possible by solving a convex program. The theory and methodology of CS, which I will review in this talk, is closely-related to the lasso and other sparse regression methods. The main focus of this talk is the application of CS principles to system identification and wireless communications. Reliable wireless communications often requires accurate knowledge of the underlying communication channel, which can be modeled as a convolution operator. The channel identification problem involves probing the unknown channel with a known training waveform and then processing of channel input and output in order to estimate its impulse response function. Wireless communication channels often have sparse impulse response functions since they tend to be the result of a few strongly reflecting objects (e.g., buildings). Classical regression methods do not exploit inherent low-dimensionality of these sparse channels, but sparse regression methods based on convex programming do. Quantitative error bounds for sparse channel recovery are derived by adapting recent advances from the theory of CS. The bounds come within a logarithmic factor of the performance of an ideal channel estimator and reveal significant advantages of the proposed methods over the classical channel estimation schemes.

Dr Gillian Lancaster

Lancaster Univerasity

g.lancaster@lancs.ac.uk

Postgraduate Statistics Centre, Dept of Maths & Stats, Fylde College, Lancaster, LA1 4YF

Joint authors: Dr Claire Glasscoe, Prof Ros Smyth, Prof Jonathan Hill

Matched cohort design to assess parental depression following child's diagnosis of cystic fibrosis

Parents facing the news of cystic fibrosis (CF) in a child are vulnerable to depression and partnership difficulties. The aim of this study was to assess risks for parental depression following the early diagnosis of CF. The longitudinal and observational nature of the study warranted a design that included a control group and was able to assess multiple risk factors. By matching on exposure a prospective, controlled, cross-informant design compared a cohort of 45 parental couples with a child recently diagnosed with CF and a control group of parents matched for age, sex, and position in the family of the affected child. Baseline and follow-up comparisons were made nine months apart.

A stratified analysis, using the Mantel-Haenszel risk-ratio estimator with eight strata, assessed the risk of depression in the sparse dataset. A paired fixed-effect linear regression analysis was conducted to examine whether quality of the couple relationship modified the association between CF/control group membership and depression. Heterogeneity was found within the dataset. Parents with a child with CF □9 months of age at baseline had an elevated risk of depression. The absence of a group effect for depression at follow-up masked this heterogeneity. Evidence was found for an interaction between group membership and the quality of the couple relationship.

With the introduction of newborn screening for CF the majority of children will be diagnosed early and more parents may be at risk of depression. The couple relationship is a clear target for interventions aimed at promoting an adaptive response. This type of study design maximized the potential of a small dataset in an area of study where data collection is difficult and very time-consuming.

Dave Collett

UK Transplant

Dave.Collett@uktransplant.nhs.uk

Fox Den Road, Stoke Gifford, Bristol BS34 8RR

Joint authors: Lisa Mumford

Malignancy in transplant recipients

Organ donation is associated with an increased risk of a number of adverse events, including cancer. Accurate information on the incidence of cancer in recipients of kidneys, livers and cardiothoracic organs has been obtained by linking UK Transplant's records of patient outcome following organ donation with data on cancer registration, assembled by the cancer registries of Great Britain. This has led to this first reported study on cancer incidence in British transplant recipients.

Cancer incidence in transplant recipients is compared to the general population using standardised incidence ratios. Differences between age groups and gender can then be examined, as can trends in the incidence of different types of cancer over time, and for different types of organ. The results show that on an overall basis, transplant recipients have about three times the risk of cancer, with highly elevated risks of skin cancer and lymphoma. However, there is variation in standardised incidence rates across the recipient age groups and the rates are higher in males than females.

Factors affecting the time to first registration of cancer in transplant recipients are investigated using competing risks models to estimate the cumulative incidence of different types of cancer. The impact of immunosuppression on cancer incidence is also investigated and survival following diagnosis of a cancer in a transplant recipient is compared to the general population.

The results of this study provide important information for counselling patients prior to a transplant and for the management of patients following organ transplantation.

Michael Festing

c/o Research Defence Society

michaelfesting@aol.com

10 Central Maltings, Kiln Lane, Manningtree, Essex CO11 1HR

Do toxicologists use the wrong animals in safety testing?

Toxicity testing methods have not changed in several decades and "The inability to better assess and predict product safety leads to failures during clinical development and, occasionally, after marketing." (FDA 2004). This is one of the reasons why drug development is now so expensive.

Toxicologists should re-assess their methods. There is substantial genetic variation in response to drugs both in humans and animals. Failure to take this into account may be decreasing the power of their experiments leading to too many false negative results. Rather than use a single genetically (and therefore phenotypically) heterogeneous "outbred" group of mice or rats in their experiments, they should be using several isogenic strains in a factorial design, without increasing the total number of animals; the main determinant of cost. Not only is this theoretically a better approach, but parallel toxicity studies on gentamycin and chloramphenicol using either a single outbred stock or a factorial design involving isogenic strains clearly shows the superiority of the latter approach for quantitative characters. It is also quite easy to see why the same strategy works for binary characters like tumour incidence.

However, virtually all toxicologists would support the statement by an anonymous toxicologist (2005) that "The variability of toxicity obtained in less well defined animals is a strength in itself, not a problem, when trying to predict safety margin in the non-isogenic human population." None of them seem to realise that increasing the variability of the test population simply leads to lower powered experiments, not to increased generality. Nor do they understand the principles of the factorial experimental design, considering each strain x treatment sub-group as being the "group size", which is then "too small".

So a whole high tech industry employing thousands of scientists appears to be designing their experiments incorrectly, possibly at a substantial cost to society. And there are very few scientists who can argue this case as it requires an inter-disciplinary understanding of toxicology, genetics and statistics. Maybe if statisticians in the pharmaceutical industry understood the problem, they could help to find a solution.

8E Health-related (Contributed)

Ms Phillippa Cumberland

Institute of Child Health, University College London

p.cumberland@ich.ucl.ac.uk

30 Guilford Street, London WC1N 1EH

Joint authors: Dr Mario Cortina-Borja

Using Copulas to model bivariate refractive error data from the 1958 British Birth Cohort

The 1958 British birth cohort comprises everyone born in Britain in one week in 1958 (Power C; IJE 2006;35(1):34-41). At 44/45 years, a random sample of 2,499/9339 (27%) members of the 1958 British birth cohort had autorefraction to measure refractive error. Data on socio-economic status, family characteristics at birth and highest educational achievement were also available.

Refractive error is measured in both eyes. Spherical equivalent (SE) is a summary measure of the strength of lens required to achieve normal vision. Refractive error can be categorised across the continuous range of SE: -13 to -0.75, myopia (short sight); >-0.75 to 1, normal vision; >1 to 9, hypermetropia (long sight). There is potentially an underlying frailty distribution as all infants are born with hypermetropic vision and 'normalisation' takes place as the eye develops. Thus myopia represents a continuation of an underlying process in the development of normal vision.

We were interested in modelling spherical equivalent as a bivariate response. Copulas provide a flexible framework to analyse bivariate distributions with fixed marginals. The marginal distributions of spherical equivalent are left skewed and highly kurtotic. The joint distribution has more variation for hypermetropic subjects than myopic subjects and few outliers. For analysis we took advantage of the R package copula (Yan J; J Statistical Software 2007; 21(4):1-21) and omnibus goodness of fit tests (Scaillet O; J Multivariate Analysis; 98: 533-543).

In this talk we analyse the joint distribution of spherical equivalent and compare different copula models using goodness of fit statistics. We also discuss using different underlying frailty distributions in the copula models and look at the effect of risk factors on spherical equivalent using bivariate regression models with error structures defined by copulas.

Dr Axel Gandy

Imperial College London

a.gandy@imperial.ac.uk

Department of Mathematics, Imperial College London, London SW7 2AZ

Sequential implementation of Monte Carlo Tests with Uniformly Bounded Resampling risk

This talk introduces an open-ended sequential algorithms for computing the p-value of a test using Monte Carlo simulation. The algorithm guarantees that the resampling risk, the probability of a different decision than the one based on the theoretical p-value, is uniformly bounded by an arbitrarily small constant. Previously suggested sequential or nonsequential algorithms, using a bounded sample size, do not have this property. Although the algorithm is open-ended, the expected number of steps is finite, except when the p-value is on the threshold between rejecting and not rejecting. The algorithm is suitable as standard for implementing tests that require (re-)sampling. It can also be used in other situations: to check whether a test is conservative, iteratively to implement double bootstrap tests, and to determine the sample size required for a certain power. A simple example is used to demonstrate the usefulness of the algorithm.

9B Choice experiments (Contributed)

Professor William. J Browne

Department of Clinical Veterinary Science, University of Bristol

William.browne@bristol.ac.uk

Langford house, Lower Langford, Bristol, BS40 5DU

Investigating environmental preferences in laying hens

Preference tests have been influential in animal welfare assessment, although animals rarely make exclusive choices between environments. Understanding the factors that result in variation in choice is important for interpretation. We examined whether various measures taken on birds housed within different environments were significantly associated with their subsequent choice between environments. We tested 60 laying hens in three different environments: wire floor (W), shavings floor (Sh) or shavings floor with additional peat, perch, and nest-box (PPN). The environments were experienced as three sets (W vs Sh; Sh vs PPN; W vs PPN) in a counterbalanced order. During housing periods (five weeks in first environment, five weeks in second environment, for each set) many indicators of physical, physiological and behavioural response were measured. After each set preferences between the two environments comprising that set were assessed with six T-maze choices per bird. Individual birds tended to make 'definite' choices at the end of each set but different birds exhibited different preferences.

To analyse what factors are associated with choice we use a random effect logistic regression modelling framework to account for correlations due to the repeated choices of the individual birds, and group housing. We combine all the experiments in one model and use, as a response, preference for the first experienced environment. We show how in this framework we can test factors that influence general environmental preference and specific environmental preference after adjusting for the particular pair of environments. For example hens may exhibit lower head temperature in environments they prefer or alternatively heavier hens might prefer the wire floor environment over the other two.

Given the large number of measures taken on each hen we also discuss a practical method for reducing the number of measures using a combination of cluster analysis and PCA. By determining how measured responses influence animal choice, we provide a link between two very different methods of welfare assessment.

9B Choice experiments (Contributed)

Dr F P Wheeler

Competition Commission

frederick.wheeler@cc.gsi.gov.uk

Victoria House, Southampton Row, London WC1B 4AD

Can economic choices be modellled without the assumption of utility

The conventional approach in marketing and economics to modelling the regularities in economic choices is the multinomial logit (MNL) model and its derivatives. The models posit the existence of underlying stable preferences and assume that choice is driven by the total utility that each customer associates with the combined attributes of each alternative. Departures from the underlying preference structure are assumed to have an extreme-value distribution.

Since utility is not observable, even in principle, it is not possible to test the validity of the extreme-value assumption. Moreover, it has been pointed out that people may demonstrate preferences that are inconsistent with rational economic choice because individual behaviour is sensitive to context and to the emotions that lie behind choices. Inconsistent behaviour is also likely where prices, or other aspects of the choice on offer, are not transparent.

Nonetheless, there may be regularities in choice behaviour. The question addressed in this paper is whether it is possible to have an analysis that models observable quantities but is not built upon a latent utility structure with its associated assumptions. It is suggested that a Bayesian approach can answer this question.

(The views expressed in this paper are those of the author and are not intended to represent the views, policy or guidance of the Competition Commission.)

Dr Catriona Queen

The Open University

C.Queen@open.ac.uk

Department of Mathematics and Statistics, The Open University, Milton Keynes MK7 6AA

Joint authors: Dr Casper Albers

Those interested in causality, Bayesian networks and forecasting.

The use of intervention for time series modelling is a well established technique for on-line forecasting and decision-making in the context of Bayesian dynamic linear models. Intervention has also been recently used in (non-dynamic) Bayesian networks to investigate causal relationships between variables, and in dynamic Bayesian networks to investigate lagged causal relationships between time series.

The Multiregression Dynamic Model (MDM) is a Bayesian dynamic model and an example of a dynamic Bayesian network. In this paper we demonstrate how intervention in the MDM can aid in the identification of contemporaneous causal relationships between time series, thus going beyond the identification of lagged causal relationships previously addressed in dynamic Bayesian networks. The methods will be applied to the problem of identifying causal relationships between traffic flows in two separate busy motorway networks in the UK.

Swarap De

The Open University

s.de@open.ac.uk

Statistics Group, Faculty of Mathematics, Computing and Technology, The Open University, Walton Hall, Milton Keynes MK7 6AA

Joint authors: Alvaro Faria and Kevin McConway

Bayesian State Space Model for Chernobyl's Radioactive Deposition in Bavaria

A state-space Bayesian model is proposed for the statistical modelling of radioactivity deposition on the ground. A Gaussian hierarchical form of this model is implemented to produce fast estimates of ground contamination levels which combine information from K-model with ground measurements of deposited radioactivity. The proposed model is appropriate for Markov random field processes in space and handles uncertainties associated with predictions from a probabilistic dispersal model (K-model), measurements and spatial interpolation. The analytical formulate allows fast updates and quick calculation of outputs. The model can handle different types of measurements. Such as, gamma dose rates and gamma spectrometry values. The model can be adapted to use in different countries with different measuring resources. Real data of radioactivity deposition from the 1986 Chernobyl accident in Southern Germany are assimilated by the model and the effects of both exponential and spherical isotropic spatial correlation structures are investigated. Also, we are doing learning from data on covariance through Normal-Wishart analysis and compare the results with results from fixed isotropic spatial covariance structure.

9C Bayesian Methods (Contributed)

Samer Kharroubi

University of York

sak503@york.ac.uk

Department of Mathematics, University of York, Heslington, York, YO10 5DD

Joint authors: Prof Tony O'Hagan and Prof John Brazier

A comparison of United States and United Kingdom EQ-5D health states valuations using a nonparametric Bayesian method

Few studies have compared preference values of health states obtained in different countries. This paper applies a nonparametric model to estimate and compare 2 EQ-5D health state utility values using Bayesian methods. The data set is the US and UK EQ-5D valuation studies where a sample of 42 states defined by the EQ-5D was valued by representative samples of the US and UK general population respectively using time trade-off technique. We estimate a utility function applicable across both countries which explicitly accounts for the differences between them, and is estimated using the data from both countries. The paper reports the results of these estimation and comparison and investigates in what respects the US and UK values for EQ-5D health states differ. The paper discusses the implications of these results for future applications of the EQ-5D and further work in this field.

9D High-Dimensional Data & Inferfence (Invited)

Phil Brown

University of Kent, UK

Philip.J.Brown@kent.ac.uk

IMSAS, Cornwallis Bldg, University of Kent, Canterbury, Kent, CT2 7NF

Feature Selection and model choice in a high dimensional setting

We describe a framework for multivariate mixed model analysis which makes use of wavelet basis functions to accommodate spikey local behaviour of the responses. Bayesian versions of variable selection prior distributions are used for identification and to display important features. Robustness priors exploiting the scale mixtures of normals also offer alternatives to more standard `slab and spike' priors. One important aspect is the ability to fuse and `borrow strength' from different data streams, depending on the degree to which they bolster the inference. The methods are illustrated on some mass spectroscopy proteomic data sets.

Jelle Goeman

Leiden University Medical Center

j.j.goeman@lumc.nl

Medical Statistics (S5-P), P.O.Box 9600, 2300 RC Leiden
The Netherlands

Joint authors: S.A. van de Geer
J. C. van Houwelingen

Testing against a high-dimensional alternative

As the dimensionality of the alternative increases, the power of classical tests tends to diminish quite rapidly. This is especially true for high-dimensional data in which there are more parameters than observations. In this paper we discuss a score test on a hyperparameter in an empirical Bayesian model as an alternative to classical tests. It gives a general test statistic which can be used to test a point null hypothesis against a high-dimensional alternative, even when the number of parameters exceeds the number of samples. This test will be shown to have optimal power on average in a neighbourhood of the null, which makes it a proper generalization of the locally most powerful test to multiple dimensions. To illustrate this new locally most powerful test we investigate the case of testing the global null hypothesis in a linear regression model in more detail. The score test is shown to have significantly more power than the F-test whenever under the alternative the large-variance principal components of the design matrix explain substantially more of the variance of the outcome than the small-variance principal components.

9E R: Foundations, Present Applications and Developments (Invited)

Peter Dalgaard

University of Copenhagen

Beyond GLM: The potential for a generic likelihood toolbox

It is a long-standing tradition in statistical computing to build on pre-existing methodology. For instance, the machinery for Generalized Linear Models builds on multiple regression analysis, not just for the fitting algorithm but also for the model specification.

However, evil tongues might express this as "when all you have is a hammer, everything looks like a nail", and in fact there are cases where choosing statistical models on the basis of whether they can be expressed as GLMs is actually harmful rather than helpful.

It is in fact not at all hard to use R to fit more general models by maximum likelihood, as long as they are reasonably well-behaved. This is currently possible using the "mle" function which minimizes an arbitrary function, assumed to be a negative log-likelihood, and creates a fitted model object that can be summarized with asymptotic SEs and analyzed graphically using profile plots.

This paper discusses possibilities for extensions and improvements of this basic approach, for instance to make it easier for the user to generate valid negative log-likelihood function and encode common types of "likelihood arithmetic" (such as mixtures, model combinations, and integration with respect to latent parameters).

9E R: Foundations, Present Applications and Developments (Invited)

Simon Wood

University of Bath

Extended smooth modelling and R

Generalized additive models are GLMs in which the linear predictor is specified partly as a sum of smooth functions of covariates. Relative to GLMs, the principle extra difficulty that such models introduce is the need to choose how smooth the component functions should be. A practical approach to GAMs, exemplified by the mgcv package in R, represents smooths using spline like basis expansions with quadratic penalties guarding against overfit. Smoothing parameters control the amount of smoothing and are selected by AIC, GCV or REML. The general estimation problem that is solved to fit such GAMs is applicable to a much wider family of models, including, for example, functional regression models, varying coefficient models, adaptive smoothing models and models with simple random effects structures.

This talk illustrates how the object orientation built into R (S) has been used to make the gam function in mgcv extensible, and how this extensibility in turn makes it straightforward to develop classes of `generalized smooths' to implement a variety of models well beyond the GAM class.

9E R: Foundations, Present Applications and Developments (Invited)

Antony Unwin

University of Augsburg

R packages and packaging R

R packages are a part of R's success story. However, there are now so many of them that it is difficult to keep track of what is available. Many analyses can be carried out in a variety of different packages and it is not always clear which to choose. Which packages are good or not so good is hard to assess and as Bill Venables put it in a recent mailing to the R-help list: "Most packages are very good, but I regret to say some are pretty inefficient and others downright dangerous."

This paper discusses R packages in general and what might be done to organise the system more effectively. R needs to be packaged in a somewhat different way, if we are to gain the most advantage from the package system.

Goetz M Richter

Soil Science, Rothamsted Research

Goetz.richter@bbsrc.ac.uk;

West Common, Harpenden, AL5 2JQ, UK

Joint authors: A. Gordon Dailey, Andrew Riche, Salvador Gezan

Uncertainties of estimating productivity for biofuel feedstocks at the small field and regional or sub-regional scale

The scientific, socio-economic and environmental assessment of climate change impacts need answers at different scale. Empirical and process-based models are used to estimate the obtainable production of bioenergy crops. The objectives here are to show how crop models of different complexity produce uncertainty of bioenergy production from spatial scale (input variation) and processes (parameter variation). First, empirical, multi-linear regression models were adapted from a single site and several sites across the country, which changed the model variables and parameter error, and almost doubled the error of the model estimates. In succession, we demonstrated how input data for soil and climatic variables affect the sub-regional yield estimates. Firstly, soil data impose limitations on variables and concepts to derive site specific available water capacity implicit to models, pedotransfer functions, stratification and parent material. Secondly, weather data from local stations may introduce a sub-regional error due to landscape (e.g. elevation), as it can be large counties like North Yorkshire, which suggests the use of interpolated weather data. Both input factors can lead to a systematically under- or overestimated yield by 1 t/ha.

In the second part, a process-based grass model is presented with its parameters for phenological and morphological development, and sink-source regulation for the allocation of dry matter. Its multi-annual calibration using single-site growth curves and evaluation against multi-site yields will be presented and compared to the empirical model derived for the same data set. Finally, a global sensitivity analysis is discussed in the context of model and parameter improvement as much as in the context of the optimization of varieties used in bioenergy research.

Marc G Genton

University of Geneva

Marc.Genton@metri.unige.ch

Department of Econometrics, UNIGE
Bd du Pont-d'Arve 40, CH-1211 Geneva 4

Joint authors: Amanda S. Hering

Powering Up with Space-Time Wind Forecasting

The technology to harvest electricity from wind energy is now advanced enough to make entire cities powered by it a reality. High-quality short-term forecasts of wind speed are vital to making this a more reliable energy source. Gneiting et al. (2006) have introduced an accurate and sharp model for forecasting the average wind speed two hours ahead based on both spatial and temporal information; however, this model is split into nonunique regimes based on the wind direction at an off-site location. This work both generalizes and improves upon this model by treating wind direction as a circular variable and including it in the model. It is robust in many experiments, such as predicting at new locations and under rotations of the wind directions. We compare this with the more common approach of modeling wind speeds and directions in the Cartesian space and use a skew-t distribution for the errors. The quality of the predictions from all of these models can be more realistically assessed with a loss measure that depends upon the power curve relating wind speed to power output. This proposed loss measure yields more insight into the true worth of each model's predictions.

Keywords: Circular variables; Power curve; Skew-t distribution; Space-time modeling; Wind direction and speed.

10A Climate & Renewable Energies (Invited)

Chris Glasbey

Biomathematics and Statistics Scotland

chris@bioss.ac.uk

BioSS
King's Buildings
Edinburgh EH9 3JZ

Joint authors: Dave Allcroft

A spatiotemporal auto-regressive moving average model for solar radiation

To investigate the variability in energy output from a network of photo-voltaic cells, solar radiation was recorded at ten sites every ten minutes in the Pentland Hills to the south of Edinburgh. We identify spatio-temporal auto-regressive moving average (STARMA) models as the most appropriate to address this problem. Although previously considered computationally prohibitive to work with, we show that by approximating using toroidal space and fitting by matching autocorrelations, calculations can be substantially reduced. We find a STAR(1) process with a first-order neighbourhood structure and a Matern noise process to provide an adequate fit to the data, and demonstrate its use in simulating realisations of energy output.

Dr Yuzhi Cai

University of Plymouth

ycai@plymouth.ac.uk,

School of Mathematics and Statistics, University of Plymouth
Plymouth, PL4 8AA United Kingdom

Quantile Function Modelling

Markov chain Monte Carlo (MCMC) method has wide applications in many areas, see, for example, Berg (2004) and references therein. Different types of MCMC methods have been proposed to deal with different problems. For example, Green (1995) proposed a reversible jump MCMC method to allow proposals that change the dimensionality of the space; Cai and Stander (2008) and Cai(2007) proposed a MCMC method for quantile self-exciting autoregressive time series models. It is worth mentioning that in all these constructions of the MCMC methods we always assume that the mathematical form of the probability density or distribution function of the model parameters is known. Hence, the corresponding posterior distribution function of the parameters can be obtained explicitly. To the author's knowledge, little work can be found in the literature about MCMC methods for implicit probability density functions. However, in many cases statistical inferences need to be made based on probability density or distribution functions which are only available implicitly. For example, many random variables do not have an explicit distribution function and hence a density function but do have an explicit inverse function of the distribution function, i.e. quantile function. In this paper, we present a MCMC method so that implicit density functions and hence quantile functions can also be dealt with properly in a MCMC framework. This talk will show that statistical modelling through implicit density functions or quantile functions can deal with many statistical problems very well in different areas which may not be dealt with well by using a traditional distribution function approach.

Professor William. J Browne

Department of Clinical Veterinary Science, University of Bristol

William.browne@bristol.ac.uk

Langford house, Lower Langford, Bristol, BS40 5DU

Joint authors: Dr Mousa Golalizadeh, Professor Martin Green and Dr Fiona Steele (DCVS University of Bristol, Nottingham Vet School and GSOE, University of Bristol respectively)

Simple methods to improve MCMC efficiency in random effects models

MCMC methods have continued to grow in popularity as their flexibility in terms of the vast number of models they can fit is realised. The family of MCMC algorithms is large and many applied researchers exposure to MCMC methods is through their use of the default estimation methods provided in software packages such as WinBUGS or MLwiN. Although these packages often try to optimize the steps of the algorithm they use to fit particular models they can still produce algorithms that result in poorly mixing chains. Many statistical methodologists produce model specific methods to improve mixing and create efficient MCMC algorithms, but for this methodology to impact on the applied community it needs to be implemented in available software. One particular way to improve the efficiency of an MCMC algorithm is through model re-parameterisation. Some reparameterisation methods can be easily implemented by modifications to the model code input into WinBUGS or via some forthcoming developments in MLwiN. In this talk we describe briefly three such reparameterisation techniques, hierarchical centering (Gelfand et al. 1995), parameter expansion (Liu et al. 1998) and orthogonalisation of the fixed predictors (Browne et al. submitted) which can be easily implemented in WinBUGS. We will show how these methods perform on a selection of random effect models applied to examples from ecology, veterinary epidemiology and demography.

Andrew Runnalls

Computing Laboratory, University of Kent

A.R.Runnalls@kent.ac.uk

Computing Laboratory, University of Kent, CANTERBURY, Kent CT2 7NF, UK

Towards Provenance Tracking in R

There is increasing interest within information systems in keeping track of the provenance of data objects such as files and database records, i.e. in determining what source data the data object is derived from, and exactly what sequence of operations was applied to the source data to generate the data object. Within the literature on provenance-aware computing (as it is called), it is widely recognised that a pioneer paper was Auditing of Data Analyses, published in 1988 by Becker and Chambers, in which they describe the S AUDIT facility. However, no comparable facility exists in R.

CXXR (http://www.cs.kent.ac.uk/projects/cxxr) is a project by the author to refactor the R interpreter into C++, and a major motivation for this is to facilitate architectural changes in the interpreter allowing the provenance of R data objects to be tracked at various levels of granularity.

The purpose of the proposed paper is to stimulate discussion among statisticians about the sorts of provenance-tracking features they would like to see in R. It will start with an overview of the current state of play in provenance-aware computing, in particular identifying any emerging standards and technologies that developments in R need to take account of. The paper will describe some of the problems that need to be addressed and technical choices that need to be made regarding R, for example questions about serialisation/deserialisation, interfacing with external provenance-tracking tools, or about the granularity with which data should be tracked: e.g. data frame, column of data frame, individual element of a column? Arising from this, the paper will propose that is important to set up an open and flexible underlying architecture, to enable a variety of researchers to try out numerous ideas. Finally the paper will summarise progress within CXXR towards such an open architecture.

The paper is intended to be accessible to statisticians with some familiarity with R or S-plus. Some knowledge of the basic concepts of object-oriented programming will be helpful, but no detailed knowledge of programming will be assumed.

10C Statistics & Sport (Invited)

Tim Swartz

Simon Fraser University

tim@stat.sfu.ca

Dept of Stats, Simon Fraser University , Burnaby, BC, Canada VSA 156

Using sports and the Olympics for statistics teaching, learning and research

This talk describes aspects of a graduate course in Statistics in sport that was first taught at Simon Fraser University in 2004 and will be taught again in 2008. It is argued that the course allows graudate students to gain valuable expertise in (a) statistical modelling, (b) the reading of scientific papers and (c) statistical methodology . Some student projects have resulted in publishable work. We report on some student work in progress that is relevant to synchronized diving at the 2012 London Olymics.

Phil Scarf

University of Salford

p.a.scarf@salford.ac.uk

Centre for OR and Applied Stats, Salford Business School, University of Salford, Salford M% 4WT

Optimum strategy in sport: examples from cricket and track cycling

We model: the distribution of runs scored in a partnership in test cricket using a negative binomial distribution; and the probability of match outcome given the end of third innings position. For the partnership scores in particular, the association between the run-rate and runs scored is modelled. These models then allow us to consider “optimal” batting strategy for a team batting third in a test match and aiming to set a target for the team batting last. On-going work that considers strategy in the match sprint in track cycling is also briefly described.

Dr Nicoletta Rosati

ISEG - Technical University of Lisbon and CEMAPRE

nicoletta@iseg.utl.pt

ISEG/UTL, Department of Mathematics
Rua do Quelhas 6, 1200-781 Lisbon (Portugal)

Joint authors: Montezuma Dumangane (Dr.)
Anna Volossovitch (Dr.)

Departure From Independence and Stationarity in a Handball Match

This paper analyses direct and indirect forms of dependence in the probability of scoring in a handball match, taking into account the mutual influence of both playing teams. Non-identical distribution and non-stationarity, which are commonly observed in sport games, are studied through the specification of time-varying parameters. The model accounts for the binary character of the dependent variable, and for unobserved heterogeneity. The parameter dynamics is specified by a first-order auto-regressive process.

Data from the Handball World Championships 2001-2005 show that the dynamics of handball violate both independence and identical distribution, in some cases having a non-stationary behaviour.

Key-words: Binary choice, dynamic panel data, time-varying parameters, unobserved heterogeneity, dependence, non-stationarity.

Daphne Kounali

Centre of Multilevel Modelling, University of Bristol

Daphne.Kounali@bristol.ac.uk

Centre for Multilevel Modelling
Graduate School of Education, University of Bristol,
2 Priory Road, Bristol BS8 1TX

Joint authors: Tony Robinson, Harvey Goldstein

The probity of Free School Meals eligibility in official records as a measure of education disadvantage.

The use of Free School Meals (FSMs) eligibility recorded by the Pupil Annual Census (PLASC) is widely prevalent in official estimates of economic and education disadvantage. However, the data collected are not FSM-eligibility but FSM claim records and this is explicitly stated in the DFES guidelines to schools for filling their PLASC returns.

In this work we use a Bayesian hierarchical model to characterize the dynamics of poverty at the income thresholds measured by the PLASC FSM claim records where detectability of eligibility is imperfect. The poverty dynamics are modelled as a non-homogenous hidden two-state Markov process, where the observed process is the presence or absence of an FSM claim. This is assumed to be conditionally independent given the hidden process, i.e. the underlying "true" state of falling below the FSM eligibility income thresholds, which evolves according to a first order Markov chain. The proposed model allows the estimation of the transition probabilities of the hidden states as well as the sensitivity of official records to detect those below the intended income thresholds. We also allow the transition probabilities and sensitivity parameters to depend on a number of covariates. Using this model we quantify the size of error in estimating the proportion of pupils who are below the income thresholds implied by FSM-eligibility when using the official claim records. We then proceed to assess the consequent error in the estimates of education disadvantage associated with FSM eligibility.

We use data extracted from PLASC for the whole Hampshire following the cohort of pupils at Reception year during 2001/2 until their KS1 tests. We examine the effect of poverty associated with FSM-eligibility income thresholds on these pupils' performance at KS1 reading tests. (ESRC funding RES – 000-23-0784; PTA-026-27-1600)

James Halse

Department for Children Schools and Families

James.halse@dcsf.gsi.gov.uk

6th Floor, Moorfoot, Sheffield, S1 4PQ

Joint authors: Clare Baker, Michael Greer

Mama Never Told Me (What qualifications she had)

Across the literature on educational development, parental education, in particular mother's education, is often cited key explanatory factor for children's educational progression and attainment. In some studies, however, the level of parental education is collected by proxy from their children. We compare the levels of parental qualifications reported in two studies of young people in England. In the first study, the Longitudinal Study of Young People in England (LSYPE), parents were asked directly about their qualifications whereas in the second study, the Youth Cohort Study (YCS), young people were asked about the qualifications their parents held. The levels of parental qualifications reported in the YCS are significantly higher than those reported in LSYPE. We explore possible reasons for this disparity such as question wording. We measure the associations between parental qualifications and other variables common to the two studies to investigate what biases the apparent over reporting in the YCS leads to.

Frank Dunstan

Cardiff University

dunstanfd@cardiff.ac.uk

Dept of Primary Care and Public Health, Cardiff University, Neuadd Meirionnydd, Heath Park, Cardiff CF14 4YS

Joint authors: John Watkins

Statistical models for identifying cases of child abuse

Identifying child abuse is an extremely important task for paediatricians. Missing cases of abuse is clearly serious; incorrectly accusing someone of child abuse can also have major consequences. Cases can present with a variety of clinical signs, such as a fracture, a burn or serious bruising. Can the pattern of bruising be used to assist in the diagnosis of abuse?

In earlier work we compared bruising in abused children with that in 'controls' who attended outpatients for other reasons. A scoring system was devised which took account of the regions of the body that were bruised, as some are more discriminatory than others. This led to a classification rule with high sensitivity and specificity for discriminating between accidents and abuse.

An important alternative diagnosis is a bleeding disorder; children with such a condition are known to be more liable to bruise. Can the pattern of bruising be used to differentiate them from cases of abuse?

We have collected longitudinal data, weekly for up to 12 weeks, on controls and on children with bleeding disorders, at different developmental stages, with details of the size and location of over 4000 bruises. It is not possible, for practical and ethical reasons, to collect such data on abused children and we also have cross-sectional data on both abused children and controls. We will describe the modelling of the longitudinal data, using repeated measures methods for discrete variables, and discuss how the different types of data can be used to attempt to discriminate between cases of abuse and other causes.

This work is part of a more extensive programme of research aimed at devising methods to assist clinicians in their decision making in cases of suspected abuse.

POSTER PRESENTATIONS

Poster 1

Mr Obisesan Olalekan

Department Of Statistics ,University Of Ibadan,Nigeria

Department Of Statistics,University Of Ibadan, Nigeria.

Joint Authors: Adekanmbi, D.Bolanle

Assessing water pollution using Lognormal Distribution in a Generalised Linear Modelling Framework

Water pollution has generated health and social economic problems in Nigeria and in Africa as a whole. The environment has been polluted by man through the introduction of substances liable to cause hazards to human health and ecological systems. In this study, evaluation of pollutants in Eleyele reservoir in Ibadan, was considered by using the lognormal distribution in a Generalized Linear Model (GLM) framework with the objective of identifying pollutants with the highest pollution risk. This approach is justified since most works on water analysis indicate pathways of pollution only which does not allow sound statistical methodology. The assessment was used to compare pollution levels. The explanatory variables considered in the model are Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD) and Turbidity with the pollution variable as the dependent variable. Data on the composition of the chemicals collected have sixty water samples from the Dam Site, a company dump site, and a receiving stream site. The parameters of the lognormal model were estimated using Maximum Likelihood Method. The Shapiro-Wilk method was used to check the suitability and the goodness-of-fit of the model while the deviance statistic was used for model comparison.

The GLM results confirmed that the DO concentration was insignificant (p-values greater than 0.05) and COD was the most significant (p-value less than 0.05). The low level of DO indicate a threat to live for aquatic organisms. The Shapiro-Wilk Model showed that the distribution of pollutants was not normally distributed; hence the lognormal model is a good fit. The smaller deviance values of the logarithm of variables indicate the suitability of the lognormal model.

Poster 2

Dr O.I Shittu

University of Ibadan

shittu.olanrewju@gmail.com

Department of Statistics , University of Ibadan, Ibadan, Nigeria

Joint authors: Mr T.A Oyeniyi and T.O Olatayo

The Seasonality of Birth in Urban Southwestern Nigeria

The analysis of registered birth has received considerable attention in the statistical literature for the developed countries with less attention paid to such exercise in Africa, particularly in Nigeria. Causes of seasonality have been attributed to cultural factors, temperature and socio economic forces. However, there have been dramatic and puzzling differences in the time periods and in the pattern of seasonal variation between urban and rural areas.

This paper therefore focuses on the study of difference in the degree of seasonality of birth between the rural and urban dwellers in the Southwestern Nigeria in 1992-2005. Harmonic and the spectrum analysis were carried out on available data assuming the period (or cycle length) is known and unknown apriori respectively. The Fisher's exact is used to test the corresponding Fourier frequencies of the observed peaks in the periodograms in order to correctly determine the period of the series. Our results shows that the maxima occur between July and October with between 7% and 21% change in the mean level. Assuming a 40-week gestation period, it can be interfered that conception mostly occurs around December to January when most residents of the urban areas are usually on vacation. The result agrees with the reports from Australia in the USA.

Poster 3

Okunlade Abiola Akintayo

Nigerian Institute of Science Laboratory Technology,Ibadan Nigeria

cosby72001@yahoo.com

P.O.Box 9332, University of Ibadan Post Office Box, Ibadan.

Joint authors: Olayiwola Olaniyi Matthew, Ajayi Crowther University Oyo, Nigeria

The Multivariate Analysis of students' academic performance in Nigerian Universities

In planning, education intervention are often needs to determine what factors are related to success features after the completion of the course of study; identify students at risk and evaluate the impact of any new program on students' performance. In this paper, we shall discuss the use of multivariate analysis of variance {MANOVA} as a statistical tool for addressing all the issues. MANOVA technique is designed to investigate the difference between two or more groups of people with respect to several underlying variables. The key details of the proposed paper is in three folds:

1. To investigate whether the population mean vectors are the same and if not which mean component differs and why,
2 To compare the joint influence of general intelligence and level on the academic performance of students
3 Pair wise comparisons of the mean difference of each level between students that entered through university matriculation examination and direct entry. These goals are adverted by analyzing the students' academic records from their point of entry to their final year.

The purpose of this article is to identify why students who enter the university through UME are always at risk of getting good grade at the end of their course of study. These are proved useful in education strategic planning and implementation.

Multivariate analysis should be the preferred method of operation in educational intervention. As we have seen, it provides other benefits in addition to being the statistically correct procedure. Data are reduced more efficiently and non-predictive variables are eliminated earlier in the analysis process. reliably identified in more easily way (i.e. those students that entered the university through UME to 100L academic records) the technique is used to predict students at risk. A technique that could identify the factors those are predictive of course performance. Using information available from students records we are able to successfully discovered that 80% of the students who eventually have lower grades (pass, 3rd class, second class lower) are from the students that entered the university through UME (i.e. those that their result include 100level records). MANOVA, technique allows for control of overall (experimental) variables and also reduce the dimensionality of the problem (without loosing too much information) thus making possible to compare the general academic performance of students given an insight into the future pattern of students' level of performance. This we believe will assist the educational policy planners and formulators in polytechnics, national college of education and universities.

Poster 4

Alireza Daneshkhah

Strathclyde University

Alireza.daneshkhah@strath.ac.uk

Department Of Management Science,
University Of Strathclyde, Graham Hills Building, 40 George Street, G1 1QE

Joint authors: Prof. Tim Bedford

Emulation of Poincare Return Maps with Gaussian process

In this paper we investigate the use of Gaussian emulators to give an accurate and computationally fast method to approximate return maps, a tool used to study the dynamics of differential equations. One advantage of emulators over other approximation techniques is that they encode deterministic data exactly, so where values of the return map are known these are also outputs of the emulator output, another is that emulators allow us to simultaneously emulate a parameterized family of ordinary differential equations (ODEs) giving a tool to assess the behaviour of perturbed systems. The methods introduced here are illustrated using two well-known dynamical systems: The Rossler equations, and the Billiard system. We show that the method can be used to look at return maps, bifurcation diagram and discuss the further implications for full computation of differential equation outputs.

Poster 5

Dr Anita McGrogan

University of Bath

a.mcgrogan@bath.ac.uk

Department of Pharmacy and Pharmacology, University of Bath, Bath, BA2 7AY

Joint authors: Mrs. Julia Snowball, Professor Corinne S. de Vries

Statin drugs and pregnancy outcomes: a cohort study using the General Practice Research Database

Background: HMG-CoA reductase inhibitors (statins) are used to reduce cholesterol levels to prevent cardiovascular disease. Because of teratogenicity in some animals they are contraindicated in pregnancy. However, limited information exists regarding their effects in human gestation and occasionally, they are used in pregnancy. A few case reports exist suggesting teratogenicity in humans, but thus far evidence is conflicting. Because both exposure to statins and adverse pregnancy outcomes are relatively uncommon, it is a challenge to gather sufficient data for the evaluation of pregnancy outcomes following exposure to statins in utero. Nonetheless it is important because, as a consequence of the rise in obesity and diabetes prevalence, statin use is expected to increase. We carried out a cohort study of statin exposure and pregnancy outcome in women registered with general practitioners (GPs) in the UK using the general practice research database (GPRD).

Methods-Data source: The GPRD contains the anonymised medical and prescription records for about 5% of the UK population. The data are entered on to GPs' computer software during and as part of routine patient care. GPs have been given feedback regarding data quality and completeness and those that are considered 'not up to standard for the purposes of research' are highlighted as such on the database so they can be excluded from studies. Data collection for the GPRD began in 1987 and is ongoing. We identified all women aged between 11 and 49 with a record of pregnancy. Based on records indicating last menstrual period, expected date of delivery, premature delivery, date of delivery etc we estimated the pregnancy period. For those for whom no such information was available we estimated the pregnancy period to be 280 days.

Study design: All women with a record of statin use in the three months leading up to and/or the first trimester of pregnancy were identified and matched to 10 non-statin users of the same age and who were pregnant in the same year. We carried out a cohort study of statin exposure during the first trimester and the association with the following pregnancy outcomes: pregnancy terminations, premature birth, and congenital malformations. Crude and adjusted relative risks (RRadj) with 95% confidence intervals (CI95) were calculated using conditional logistic regression analyses and adjusting for confounders such as smoking and alcohol.

Results: 192 statin users and 1943 matched reference women were identified. The RRadj of recorded pregnancy terminations associated with statin use in the first trimester was 2.48 (CI951.65-3.73). Alcohol abuse was associated with an RRadj of 8.64 (CI952.24-33.29). For 78% of pregnancies, the offspring could be identified on the database. There were no premature births amongst the statin exposed pregnancies. Congenital malformations were recorded for 2% of statin-exposed and for 4% of statin unexposed pregnancies. The RRadj of congenital malformations associated with statin use was 0.66 (CI95 0.15-2.89).

Conclusions: Statin use in early pregnancy is associated with an increased risk of pregnancy terminations. This may be a consequence of the impact statins have on the constitution of the placenta. In addition, given the lower rate of congenital malformations amongst the exposed pregnancies, it is likely this reflects the effects of more intensive prenatal screening amongst the exposed. Work is ongoing to determine what proportion of the terminations was elective.

Poster 6

Mark Edmondson-Jones

MRC Institute of Hearing Research

markej@ihr.mrc.ac.uk

MRC Institute of Hearing Research, Science Rd, University Park, Nottingham. NG7 2RD

Psychometric function estimation in child populations, with stimulus independent lapses

One approach to the analysis of psychophysical data (such as a subject's ability to perceive a given auditory stimulus at a specified level) is via the fitting of psychometric functions. The psychometric function provides a measure of the probability of perception of the given stimulus by the subject and is typically assumed to have a sigmoidal form.

Here we investigate such an approach in relation to data gathered from an adaptive procedure, where stimulus levels are determined by prior performance, where the subjects are asked to identify the source of an 'oddball' stimulus and are forced to make a choice from three alternatives. I.e. the lower asymptote of the psychometric function reflects chance performance. Furthermore, the majority of the subjects studied included a number of young children. This practically limited the number of trials that could be performed. It also necessitated an estimation of stimulus independent lapses at the upper asymptote of the psychometric function due, e.g., to inattention. These factors complicate a conventional logistic regression approach.

The study population comprised 20 subjects aged 6-7, 28 subjects aged 8-9, 25 subjects aged 10-11 and 31 adults. Each subject participated in up to eight training blocks with each block being made up of three interleaved adaptive tracks of 25 trials each.

The principle analysis performed involves the estimation and comparison of hierarchical mixed-effects Bayesian mixture models with binary response (developed using WinBUGS and implemented using the BRugs package in R) in order to investigate the evidence for auditory learning across a number of training blocks and between age groups. Alternative approaches are also discussed.

Poster 7

Dr Vijay M Sarode

Mulund College of Commerce, Mulund(W), Mumbai - 400 080, India.

vijaymsarode@yahoo.com

Ashar Estate, Flat B2-501, Shree Nagar, Wagale Estate, THANE(W), 400 604, INDIA.

Logistic Modeling to determine delivery complications among women in slum in Greater Mumbai

This study uses primary data, collected using cluster sampling of sample size of 433 reproductive women who have given at least one live birth prior to survey on antenatal care indicators, antenatal check-ups, and reproductive health problems during the pregnancy and the complications while delivering a child from Rafi Nagar slum. This paper examines utilization of health services available to these women in slums in Mumbai and also checks whether non utilization of ANC and having reproductive health problems during pregnancy creates complications during child delivery on the basis of standard of living index constructed from household amenities, housing quality and sources of drinking water, electricity and toilet facilities. The findings using logistic regression reveals unimaginable low level of utilization of health services for illiterate women in the study area. Besides to these there is evidence that those respondents did no go for ANC and had reproductive health problems during the pregnancy creates problems during child delivery, particularly to illiterate mothers. This paper suggests that awareness is required at every stages of ANC particular to illiterate women with low SLI category women in a slum.

Poster 8

Poly Chigbu

University of Nigeria

pechigbu@yahoo.com

Department of Statistics, University of Nigeria, Nsukka, Enugu State, NIGERIA

Joint authors: A.V. Oladugba

On Block Structures, Null Analyses of Variance and Expectations of Mean Squares of Quasi-semi-Latin squares

A quasi-semi-Latin square is here defined as a combinatorial object whose entries are ab initio arranged as in the semi-Latin square formation without any regard to any other block structure apart from the one associated with the usual semi-Latin square but which actually has a peculiar blocking system as might eventually be defined as the case may be. In this paper, the block structures of some quasi-semi-Latin squares are considered. The null analyses of variance, covariance matrices and expectations of mean squares of these squares are also generated using their respective block structures.

Poster 9

Dr Michael Williams

GlaxoSmithKline R & D

michael.k.williams@gsk.com

Biostatistics and Programming Development Partners , Building 38 1 227, Greenford Road, Middlesex UB6 0HE

Identification of Events that may Change Risk of Subsequent Adverse Events of Special Interest: A Novel Approach

With the increased focus on presentation of safety data and the use of data collected in clinical trials to assess the significance or otherwise of potential safety signals, a question that has arisen is whether or not we can more effectively assess the risk-benefit balance by more effective use of the data collected as part of a submission process. One aspect of this is how we can use the data to provide more information for patients of the risks involved with a medication, and how they can identify when these risks are likely to occur.

We propose a way of identifying events, which may be minor in nature, but which may lead to a significant change in the risk of subsequent more serious events. The method uses the same principles as the disproportionality indices, as proposed by Evans et al (2001), but instead of identifying drug-event pairs that may be of interest we use it to identify event-event pairs for assessment by use of an Event Association Score (EAS).

A methodology is presented that allows for the results of such an analysis to be ranked for review by medically qualified personnel, which also deals in part with the potential multiplicity issue. We illustrate the method by use of a clinical trail database, and show how the assessment would lead to further, more appropriately chosen analyses.

Poster 10

Carlos Cuevas-Covarrubias

Anahuac University, Mexico.

ccuevas@anahuac.mx

Universidad Anahuac, Escuela de Actuaria, Av. Lomas Anáhuac s/n, Lomas Anáhuac, Huixquilucan Edo. Mex., C.P. 52786 Mexico

Joint authors: Roberto Jasso-Fuentes; Anáhuac University, Mexico.
Nelly Altamirano-Bustamante, M.D.; National Institute of Pediatrics

Curves of growth for Mexican children. An application of quantile regression and kernel smoothing.

Reference tables and charts of growth are very important instruments to assess children's physical development. Both of them describe the expected variations in height and weight through childhood. Mexican pediatricians frequently use international curves of growth that are not representative of the Mexican population. Some of them use national charts that were built fifty years ago. This poster presents an application of nonparametric regression to build these reference tables. Our objective is to obtain curves of growth for nowadays Mexican Children. Our work is based on a random sample of Mexican children with ages from 6 to 12. We start using kernel smoothing to fit a regression curve. Then, we estimate conditional percentiles from the probability distribution of the residuals at different ages; finally, we apply nonparametric regression to build charts of growth (height and weight) based on these conditional percentiles. We also propose a new algorithm to implementation Gaussian Kernels efficiently

Poster 11

Daniel Bergmann

University of Nottingham

Pmxdb1@nottingham.ac.uk

University Park, Nottingham, NG7 2RD

Joint authors: Prof. John King, Prof. Andrew Wood, Dr. Matthew Loose

Granger Causality Methods for Reconstruction of Genetic Networks

Genetic networks have been studied in a mathematical framework since the 1960s by Kauffman and others. Many methods have been used and applied to modelling these and statistical methods have been used due to the stochastic nature of such biological processes.

With the advancement of biological techniques such as microarrays to measure RNA activity, there has been great interest in attempting to recover the underlying networks from whichsuch data arises.

We present the application of Granger causality to this problem in order to detect a causal link between genes and hence determine the structure of the network. Following the paper by Mukhopadhyay and Chatterjee (2007), we first consider the pairwise interaction of genes and test for significance of Granger causality from time series data obtained from each gene. Typically the time series generated from microarray experiments may be short due to both financial and time constraints in obtaining such data. We apply bootstrapping techniques to explore the variability in significance of the Granger causality test, in both time and frequency domains.

These techniques are applied to both artificial networks with typical biological features and also to a real world network, obtained from the developmental phases of Xenopus Laevis.

Poster 12

Tahani Maturi

Durham University

tahani.maturi@durham.ac.uk

Durham University, Science Laboratories,
South Rd, DURHAM DH1 3LE, UK

Joint authors: Dr Pauline Coolen-Schrijner and Prof Frank Coolen

Nonparametric Predictive Inference for ROC curve

The receiver operating characteristic (ROC) curve is a statistical tool for evaluating the accuracy of diagnostic tests. It is widely used in medical trials, radiology, machine learning and data mining. The empirical ROC curve is often used to compare different diagnostic tests in order to select the most accurate test.

We introduce an alternative nonparametric predictive inference (NPI) approach for the ROC curve. NPI uses Hill's assumption A(n) together with the available data to derive lower and upper bounds for the ROC curve for single future observations from the disease and non-disease groups. An example is provided to illustrate our approach.

Poster 13

Dr Mona Kanaan

University of York

mk546@york.ac.uk

Department of Health Sciences, 1st Floor , Seebohm Rowntree Building, University of York , Heslington, York, UK YO10 5DD

Evaluation of the Measles Epidemic in Lebanon from Case Notification data in 2003-2005 and Tracing of Notified Cases in 2005

In recent years, there has been a drive to eliminate measles globally. Although, this has been achieved in many western countries, it is not the case for many developing countries. In this paper, we seek to evaluate the current status of measles in Lebanon via two methods. The first method is based on cases notified to the Epidemiological surveillance unit (ESU) at the Lebanese Ministry of Public Health during the years 2003 and 2005. The second method is based on tracing of notified cases to ESU during 2005. For each data set, we estimate the effective reproduction number (R) for measles in Lebanon. The estimation of R will indicate whether special vaccination campaigns are needed in order to eliminate measles in Lebanon.

Poster 14

Dr Charlotte Bean

Warwick Medical School, University of Warwick

c.l.bean@warwick.ac.uk

University of Warwick, Room B-031, Social Studies Building,
Coventry. CV4 7AL

Data Clustering using Rough Set Theory

This poster presents a knowledge-oriented clustering algorithm that can be applied to data of both single and mixed-attribute type. The algorithm has a simple framework, based on that of hierarchical clustering, and the main clustering tool is a form of indiscernibility relation modified from the field of rough set theory. The clustering technique focuses on extracting maximal knowledge from data, both local and global, with minimal human intervention in order to obtain clusters that are meaningful and free from user-bias. This is achieved by employing well-defined numerical procedures to set key threshold parameters and by making use of a cluster accuracy measure to yield representative clusters, within the boundaries of the given application. The algorithm is unified in its approach to clustering, which ensures consistency in the results when used to cluster the same data by different users, and knowledge can be represented tangibly throughout the clustering process as a series of classification rule sets, thus enhancing interpretability. Numerical techniques control the setting of initial threshold parameters in order to obtain an initial clustering of the data. A defined accuracy measure quantifies the notion of cluster 'meaningfulness'. Throughout the clustering process, clusters are automatically modified using a 'threshold selection rule' and quick supervised clustering of a data set can be achieved using the classification rules obtained from the clustering of a similar data set. This tangible clustering knowledge represented by the rules can further be modified to provide a strategy for automatic decision-making

Poster 15

Martin Schroeder

NCRG, Aston University

shroderm@aston.ac.uk

Neural Computing Research Group, Aston University,
Aston Triangle, Birmingham. B4 7ET.

Joint authors: Dr. Dan Cornford, Dr. Ian T. Nabney

High-dimensional Data Imputation and Visualisation: Application in Geochemistry

Missing data are a common problem in many real, high dimensional data sets and many standard methods for data modelling and data visualisation cannot cope with them. Depending on the nature of the missing data a two-stage processing of the data is often necessary, where one initially models the data to imputes the missing values and then subsequently models the completed data set in the visualisation process. The treatment of the missing values impacts the completed data producing unpredictable consequences on the visualisation, which make them especially critical in real world applications.

In this work we look at the imputation performance of visualisation methods based on density models such as probabilistic PCA, Kernel PCA, Generative Topographic Mapping and Gaussian Process Latent Variable Models. We compare these visualisation-based methods with standard approaches to data imputation including (weighted) mean imputation and iterative multiple regression. Our benchmark data are based on geochemical properties of crude oils from the North Sea and Africa since the overall goal of the project is to evaluate probabilistic models which might be used and understood by non statisticians in this area. We show that the single-stage probabilistic joint imputation-visualisation methods perform better in the presence of missing data than non-probabilistic imputation methods while rendering a two-stage process unnecessary.

Poster 16

Andrea Mercatanti

Bank of Italy

mercatan@libero.it

Via Nazionale 91, 00184 Rome, Italy

Assessing the effect of debit cards on households' spending under the uncounfoundedness assumption

The paper proposes an application of some causal inference methods for the purpose of evaluating the effect of the use of debit cards on households' consumptions. Motivated by the evidence that debit card users overspend in comparison to non-users, the analysis wants to investigate the existence of a causal relationship rather than a mere association. The available dataset allows us to introduce a set of pre-treatment variables so that the uncounfoundedness assumption can be adopted. This gives the advantage of avoiding the introduction of assumptions on the link between the observable and unobservable quantities, and it also improves the precision in comparison to other main methodological options. The analysis results in positive effects on a household's monthly spending; it also provides a comparative application of various causal methods to a real dataset.

Poster 17

Bonnie Cundill

London School of Hygiene and Tropical Medicine

Bonnie.Cundill@lshtm.ac.uk

Infectious Disease Epidemiology Unit, LSHTM, Keppel Street, London WC1E 7HT

Joint authors: Neal Alexander (Dr)

Sample Size Calculations for Common Non-Gaussian Distribution Families Without Normal Approximations

Sample size and power calculations are often done on the basis of Normal approximation, even for data which are not Gaussian and are analysed using generalized linear models (GLMs). For example, some medical statistics textbooks which cover Poisson regression still obtain sample sizes for rates via a normal approximation. We show how this inconsistency can be avoided by using basic GLM theory to calculate sample sizes based on means on the scale of the link function (e.g. log). Formulae are given for the cases of the Poisson, gamma, and the negative binomial, the last of these being a common tool for skewed medical data such as episodes per person, or parasite or entomological counts.

We also examine the magnitude of errors in such Normal approximations. This is done initially via the Berry-Esséen bound for the maximum difference between the cumulative distribution functions (CDFs) of a standard normal distribution and the standardized sample mean of the distribution in question. However, this bound is found to be very lax in the cases of the Poisson and negative binomial, so we concentrate on simulation. For example, for sample sizes of 100, the maximum deviation of the sample mean's CDF from Gaussian is 8% for mean 0.1, decreasing to 3% for mean 1 (with the Berry-Esséen bounds being 25% and 13% respectively). The negative binomial has a second parameter k, with the Poisson corresponding to the limit $k \to \infty$. As expected, the deviations from the Gaussian CDF are larger for negative binomial than Poisson, and inversely related to k. For example, for k=0.1, and the same two values of the mean, the maximum CDF deviations are 9% and 5% respectively, while for k=0.05 they are 10% and 6%. Hence, for moderate sample sizes and plausible parameter values, the normal approximation could give appreciable error in sample sizes, and we suggest using the GLM-based method which is more consistent with appropriate data analysis.

Poster 18

Oliver Zobay

University of Bristol

oliver.zobay@bristol.ac.uk

School of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW

Variational methods in nonparametric Bayesian inference

Over the past few years, variational methods have received growing interest as an alternative to Monte Carlo integration in computational Bayesian inference. These techniques are deterministic and, typically, computationally much faster than MC integration. However, they are based on approximations whose consequences are difficult to assess a priori and which should be investigated carefully under various conditions.

Recently, several variational approaches to inference in nonparametric Bayesian models with Dirichlet process mixtures have been proposed [1]. In this contribution, a systematic investigation of the properties of the variational approximations is presented, and their applicability for various problems, e.g., density estimation, is studied. Application of these variational methods to time-series analysis is also considered.

[1] Kenichi Kurihara, Max Welling, Yee Whye Teh, Collapsed Variational Dirichlet Process Mixture Models, Proceedings IJCAI-07

Poster 19

Dr Alfred Akinsete

Marshall University

akinsete@marshall.edu

Department of Mathematics, One John Marshall Drive, Huntington, WV 25755. USA

Joint authors: Professor Felix Famoye & Professor Carl Lee

The beta-Pareto distribution

In this paper, a four-parameter beta-Pareto distribution is defined and studied. Various properties of the distribution are discussed. The distribution is found to be unimodal and has either a unimodal or a decreasing hazard rate. The expressions for the mean, mean deviation, variance, skewness, kurtosis and entropies are obtained. The relationship between these moments and the parameters are provided. The method of maximum likelihood is proposed for estimating the parameters of the distribution. The distribution is applied to two flood data sets.

Poster 20

Emma Hooper

Office for National Statistics

emma.hooper@ons.gov.uk

Room 1.156, Office for National Statistics, Government Buildings,
Cardiff Road, Newport NP10 8XG

Seasonal adjustment in the presence of bonuses

Messages in data series can be hidden by the presence of systematic seasonal patterns, for example due to the time of year or arrangement of the calendar. The time series analysis technique of seasonal adjustment removes these patterns which helps us to interpret the movements of the series. In official statistics one of two theoretical decompositions of a time series, multiplicative or additive, are commonly found to reasonably fit official data. The UK's Average Weekly Earnings (including bonuses) series presents several challenges for seasonal adjustment. The changing nature of bonus payments in recent years could pose a problem for identifying an appropriate theoretical decomposition and for estimating the seasonal component of this series using the software X-12-ARIMA. It also raises questions about the appropriate interpretation of bonus payments when it comes to seasonal adjustment. This poster (or paper) reviews the approach taken to investigate these issues, and the strengths and weaknesses of the methods tested.

Poster 21 (Winner of competition at YSM2008)

Kavian Thompson, Philip Sayer

Office for National Statistics

kavian.thompson@ons.gov.uk

philip.sayer@ons.gov.uk

Office for National Statistics, Segensworth Road, Titchfield, Hampshire PO15 5RR

Joint authors: Alexa Courtney, Beth Moon

The 2011 Census

A census is a survey of all people and households in the country. It provides essential information that the government needs to develop policies and to plan and run public services such as education. The information it provides is also widely used by academics, business, voluntary organisations and the public.

The census in England and Wales is planned and carried out by the Office for National Statistics.

The next census will take place in 2011 and a detailed planning programme is currently underway to determine what information will be gathered, how it will be processed and how the results will be produced and delivered. This poster presents the process to be carried out between 2010 and 2012 for the 2011 Census and illustrates how researchers contribute to the process and running of the census.

Poster 22 (Winner of competition at RSC2008)

Bryony Hill

University of Warwick

b.j.hill@warwick.ac.uk

Statistics Department, University of Warwick, Coventry CV4 7AL

Joint authors: Eike Thonnes, Wilfrid Kendall

Extraction of Fingerprint Ridges from Pore Data

There are three levels of features in fingerprints: Level 1 - the overall pattern, Level 2 - the bifurcations and ridge endings (minutiae), and Level 3 - the pores and ridge contours. Experiments have shown that Level 3 features hold significant discriminatory information and they are being used increasingly in fingerprint matching.

The aim of my work is to reconstruct the ridge lines of a fingerprint from only the locations of the pores. This is achieved by finding a gradient field which is approximately tangential to the ridge lines.

Previous methods have suffered in areas of low pore density and where image or fingerprint quality is bad. I propose a method using the Log-Euclidean Fréchet mean in order to smooth a sparse field of tensors located at each pore. This technique gives better results as it uses the property that fingerprint ridges are locally parallel

Poster 23

Dr Masayuki Henmi

The Institute of Statistical Mathematics

henmi@ism.ac.jp

4-6-7 Minami-azabu, Minato-ku, Tokyo 106-8569, Japan

Joint authors: Professor John B. Copas

A robust confidence interval against publication bias in random effects meta-analysis

In random effects meta-analysis, one major concern is to calculate a confidence interval for the overall mean effect across studies. A number of methods have been proposed in the literature including the DerSimonian and Laird method, which is commonly used in practice. However, most of these methods can be very sensitive to publication bias. For example, if smaller studies with smaller sample sizes are not more likely to be selected in meta-analysis (this is often observed as skewness of funnel plots), then the coverage probabilities easily decrease as the degree of heterogeneity or the number of studies increases. In this paper, we propose a confidence interval which is less sensitive to publication bias by applying the fixed effects estimate rather than the random effects estimate. There are two reasons for focusing on the fixed effects estimate. One is that it gives smaller weights to smaller studies in combining individual estimates, which leads to less sensitivity to publication bias. The other is that it is easier to make more accurate distributional approximation in constructing a confidence interval. Therefore, our method also improves the coverage probability of the DerSimonian and Laird method, even when there is no publication bias.

Poster 24

Atanu Adhikari

ICFAI University

adhatanu@yahoo.com

Astral Heights, 3rd floor, Road # 1, Banjara Hills, Hyderabad – 500 034, India.

Partitioned prior hyper-parameter in HB analysis: Application and performance.

Application of hierarchical Bayes techniques in researching human behaviour is about couple of decades old. Researchers have used hierarchical Bayes methodology in estimating unit level heterogeneity in parameter estimation. However, while using hierarchical Bayes, these researchers have used HB model considering single prior hyper-parameter in estimation process. Khatri and Rao (1992) showed that one hyper-parameter may be inadequate in individual parameter estimation if the population is heterogeneous. Population level prior hyper-parameter does not eliminate inter-segment variance which may severely affect individual estimate. Superiority of such estimates reduces if the heterogeneity in the population increases as the variability of the true mean value increases (Khatri and Rao, 1992). Since the intra-segment variability (nuisance parameter) decreases in formation of homogeneous groups within the heterogeneous population, the two stage prior distribution is supposed to give better parameter estimates than considering common prior distribution with same hyper-parameter.

This research uses choice based conjoint analysis through a multinomial logit model to estimate the parameter. Hierarchical Bayes method is used to capture unit level heterogeneity in parameter estimation. In this research, the researcher segments the sample in several homogeneous groups and considers mean values within the segment have common prior distribution with certain hyper-parameter which differs from segment to segment. The prior hyper-parameters of several such segments are assumed to have a common distribution with certain parameter at whole population level.

The model is tested and it is found that parameter estimates considering segment level hyper-parameter are significantly better than estimate using one hyper-parameter.

Poster 25

Stephen Walters

University of Sheffield

s.j.walters@sheffield.ac.uk

ScHARR, University of Sheffield, Regent Court, 30 Regent St, Sheffield, S1 4DA

How to analyse data from a cluster randomised trial in primary care: a practical guide

Health technology assessment often requires the evaluation of interventions which are implemented at the level of the health service organisation unit (such as the GP practice). These interventions are implemented for clusters of individuals. Cluster randomised trials (cRCTs) are increasing being used in primary care to evaluate new health technologies.

The majority of statistical analyses, in individually randomised controlled trials (iRCTs), assume that the outcomes on different patients are independent. In cRCTs there is some doubt about the validity of this assumption as the intervention is typically delivered by a health professional (such as a General Practitioner) and a number of patients receive the intervention from each professional. The success of the intervention can depend on the professional delivering it, so that outcomes of patients treated by the same professional may be correlated or "clustered". Hence the evaluation of outcome data from cRCT presents a number of difficulties. The aim of this presentation is to describe statistical methods of adjusting for clustering, in the context of cRCTs.

There are essentially three strategies to analyse the outcome data from cRCTs:
1. Cluster level analysis – analysis is carried out at the cluster level, using aggregate summary data (such as the mean outcome per cluster).
2. Marginal or population-averaged approach, with model coefficients estimated using generalised estimating equations (GEEs).
3. Random-effects (R-E) or cluster specific approach.

In practice both Marginal and R-E models provide valid methods for the analysis of clustered data, although the two approaches lead to different interpretations of the treatment effect.

This presentation will compare and contrast the three approaches, using example data, with binary and continuous outcomes, from a cRCT in primary care designed to evaluate the effectiveness of new Health Visitor led psychological intervention in detecting and treating new mothers with postnatal depression (PND) compared to usual care. The PONDER Trial randomised 100 clusters (GP practices) and collected data on 2659 new mothers in with an 18 month follow-up.

Poster 26

Jouni Hartikainen

Helsinki University of Technology

jmjharti@lce.hut.fi

Helsinki University of Technology
Department of Biomedical Engineering and Computational Science (BECS)
P.O. Box 9203, FI-02015 TKK, FINLAND

Joint authors: M.Sc. Jarno Vanhatalo, Dr. Aki Vehtari

Comparing Poisson and Negative Binomial likelihoods in Disease Mapping via Sparse Gaussian Processes

Gaussian processes are flexible non-parametric models, which are attractive for modelling risk surfaces in the context of disease mapping. The traditional approach is to model the disease occurrences (i.e. deaths or incidences) as a Poisson process with spatially varying rate, which is a product of expected occurrence count and relative risk, for which the latter is given a prior structure for smoothing the risk surface across the spatial domain. Here, the spatial priors are modelled with sparse Gaussian processes, which enable the investigation of larger data sets than full Gaussian processes.

In this work, we compare the Poisson and the Negative Binomial likelihoods with sparse Gaussian process priors in a general disease mapping problem. The Negative Binomial distribution is more robust toward outlying observations and can be used for investigating the overdispersiveness of the data set. Due to analytic intractability the posterior inference is conducted with MCMC as well as analytic approximations, such as Laplace's approximation and Expectation Propagation. Various sparse Gaussian process models are tested with several real world data sets and comparison of the results is given.

Poster 27

Miss Damilola Adekanmbi

Ladoke Akintola University of Technology,Oyo State, Nigeria

dammy_vicky@yahoo.com.au

Department of Pure and Applied Mathematics PMB, 4000, Ladoke Akintola University of Technology, Ogomoso Oyo State, Nigeria

Joint authors: Mr Olalekan Obisesan

Implication of Education on Fecundability of a Sample of Married Women

This study is focussed on checking for the possibility of difference in women's fecundabilities by educational status, thereby assessing the effect of education on women's fecundabilities. Data on cycles to pregnancy based on retrospective reports of a sample of married women categorised into two educational status groups were used for the comparative study. With an assumption that among couples fecundability is distributed as a beta distribution, quantile ratio is used to asses an individual level interpretation of any between-group variations in the beta distribution of the two educational status groups of married women. Likelihood ratio test was also employed in the comparative study. It was discovered that education helps in increasing the conceptive rates of the educated women who are less fertile.

Poster 28

Rebecca Walls

AstraZeneca

Rebecca.Walls@astrazeneca.com

Advanced Science and Technology Laboratory,
AstraZeneca R&D Charnwood, Bakewell Road, Loughborough LE11 5RH

Joint authors: Chris Harbron, AstraZeneca

Statistics in high-content biology

An important strategy in pharmaceutical research is to identify, as early as possible, compounds that are likely to fail at later stages of the drug discovery process, as a consequence of poor efficacy and/or toxicity to humans. One emerging approach is to employ a new technology, "multiparametric high content cell-based assays", which attempt to use in vitro cell models to mimic the complexity of the in vivo disease. Advanced imaging techniques are used to generate extremely large and complex datasets describing the response of a population of cells to a drug in a series of features, such as how the cells change shape or size, with the aim of building predictive models or "fingerprints" from the multiparametric assay data for well characterised compounds that elicit known responses. These fingerprints will then be applied to future compounds in order to predict the biological mechanism of action of the drug and its toxicity.

Traditional multivariate approaches are difficult to apply directly to such problems as the information captured for each multivariate feature is not a static data point; instead we have measurements taken over a dose range, resulting in a dynamic response to the compound for each of the features, yielding datasets with a three-dimensional cube-like structure (compounds by doses by features). In this presentation we evaluate and compare the properties of a range of statistical approaches for generating robust and reliable predictive models for future application from these three dimensional data.

Poster 29

Graham Horgan

Biomathematics & Statistics Scotland

g.horgan@bioss.ac.uk

Rowett Institute, Bucksburn, Aberdeen, AB21 9SB

Missing spots and missing data in high-dimensional proteomics.

A feature of electrophoresis gels used to study protein expression in biological samples is that many spots can be missing, i.e. not detected, in each gel. This can be due to low or zero expression, or to being obscured by nearby spots, or to simple error. There is information in the distribution of the spots and their positions that should allow us to look for evidence of these effects. We present some data and comment on these matters, and also investigate what the distributions tell us about the spatial correlation of spot intensities, and how best to detect treatment differences in expression.

Poster 30

Dr Fahimah Al-Awadhi

Kuwait University

falawadi@kuc01.kuniv.edu.kw

P.O.Box 66619 Bayan 43757 Kuwait

Joint authors: Dr. M. Soltani

An Embedded Markov Chain for Stationary Processes and its Applications

A Markov chain is associated to a finite order one-sided moving average of a discrete time stationary Gaussian process. A method is developed to specify thresholds $0=L0 <L1< ... < Lm < Lm+1=\infty$ for given on target significant levels $\pi 0, \ldots, \pi m$; $\sum i=0m \pi i =1$; in the sense that in the long run the probability that the moving average process lies in Li,Li+1, will be πi, i=0,…,m. Special inputs, AR(1) and MA(1) are treated in details. This study extends the work of Soltani et al. (2007) where the inputs were assumed to be i.i.d.; and a single threshold was considered.

Poster 31

Marco Geraci

University of Manchester

marco.geraci@manchester.ac.uk;

North West Cancer Intelligence Service
Manchester Office, Christie Hospital NHS Trust
Kinnaird Road, Withington, Manchester M20 4QL

Joint authors: Prof. Jillian M Birch; Prof. Tim OB Eden; Dr Anthony Moran; Dr Robert D Alston

Analysis of cancer survival by region in teenagers and young adults in England

Various studies have showed that tumours in young people aged 13 to 24 years are a major source of morbidity and mortality in this age range in the UK, exceeded only by deaths from transport accidents as a cause of death. There is a growing, international recognition that teenage and young adult cancer patients have particular physical, social and educational needs in addition to the need for appropriate disease-specific treatment.

National incidence, survival and mortality analyses provide important information for service planning and for identifying those diagnostic groups that present greatest challenges. Between 1979 and 2003, approximately 1600 new cases of neoplasms have been diagnosed every year in England alone and the rate has increased by 1.5% per year.

We analysed survival of patients aged 13 to 24 years diagnosed with cancer in England between 1979 and 2001, followed up to 31 December 2003. The number of deaths for each observation is assumed to be Poisson distributed and the five-year relative survival rate is estimated using a generalized linear model (Dickman et al, StatMed, 2004). Heterogeneity between Government Office Regions was tested for by using a likelihood ratio test statistic.

Lymphomas, bone tumours, soft tissue sarcomas, germ cell tumours and carcinomas showed significant geographical variability in survival after taking into account followed-up time, time period, age, and socioeconomic deprivation. Between 1993 and 2001, the five-year relative survival rates ranged from 50% for bone sarcomas and leukaemias to almost 100% for thyroid carcinomas, overall. The geographical patterns showed by survival rates were also observed when comparing national incidence and mortality data using incidence to mortality ratios.

These results suggest there may be inequalities in service delivery around the country and provide important baseline data to help the development of specialised service provision for this group of patients.

Poster 32

Fiona Warren

Dept. of Health Sciences, University of Leicester

fcw2@leicester.ac.uk

University of Leicester, Dept of Health Sciences,
2nd Floor Adrian Building, University Road, Leicester LE1 7RH

Joint authors: Keith Abrams, Alex Sutton, Tim Bongartz, Eric Matteson

Development of evidence synthesis methods using hierarchical models to investigate influences of class effects and dose-response: application to anti-TNF drugs for rheumatoid arthritis.

The aim for this research is to develop and compare meta-analysis models for evidence synthesis of data where the primary outcome is an adverse event due to a clinical intervention. In the case that a drug therapy causes an adverse event, it may be of interest to investigate the adverse event profiles for multiple drugs within a class, with development of mixed treatment comparison methods. A further element of interest is the concept of dose-response modelling, which is related to duration of use. Meta-analysis methods using Bayesian hierarchical models will be used to develop adverse event profiles for individual drugs, bearing in mind these issues.

An interesting clinical example to which such methods will be applied is that of anti-tumour necrosis factor (anti-TNF) drugs for rheumatoid arthritis. There have been concerns regarding an increased risk of malignancies for anti-TNF users*; signals from available data are, however, difficult to interpret due to sparsity of events. Using this example, we begin with a simple model combining data for all drugs in this class, and extend this model with additional levels in the hierarchy to represent anti-TNF drugs individually and then to model anti-TNF drugs as an overall class with related but non-identical adverse event profiles for malignancy. This model will be further adapted to incorporate data for dosage and external data, for example from observational studies. Comparisons between models will be made using the deviance information criterion (DIC).

*Bongartz T, Sutton AJ, Sweeting MJ, Buchan I, Matteson EL & Montori V (2006). Anti-TNF antibody therapy in rheumatoid arthritis and the risk of serious infections and malignancies: systematic review and meta-analysis of rare harmful effects in randomized controlled trials. JAMA, 295(19): 2275-2285.

Poster 33

Maria Vazquez

University of Warwick

M.d-l-A.Vazquez-montes@warwick.ac.uk

15 Quaker Court, Banner street, London, EC1Y 8QA.

A marginal-structural estimator of causal effects

We consider the problem of estimating the causal effect of two treatments in an observational study. From the literature, the Horvitz-Thompson estimator and Rosenbaum's model-based direct adjustment on the propensity score method are available under a counterfactual approach. However, the Horvitz-Thompson estimator is not invariant under translations in the potential outcomes, and Rosenbaum's method ignores the model of the propensity score in the last stage of the estimation. Assuming a missing data point of view, the regression imputation method can be applied as well. We propose a marginal-structural (MS) estimator, which corrects the invariance property found in the Horvitz-Thompson. The MS estimator can be seen as a special case of Rosenbaum's but including the propensity score model; under certain conditions it can be more robust than the regression imputation estimator. Descriptions, comparisons, an application example, and some simulation results will be included in the poster.

Poster 34

Dr John Haywood

Victoria University of Wellington, New Zealand

John.Haywood@vuw.ac.nz

School of Mathematics, Statistics and Computer Science, Victoria University of Wellington, PO Box 600, Wellington, New Zealand

Joint authors: Dr John Randal

Modelling multiple structural breaks in seasonal data. Did 9/11 affect NZ visitor arrivals?

We demonstrate the poor performance, with trending seasonal data, of existing methods for endogenously dating multiple structural breaks. Motivated by iterative nonparametric techniques, we present a new approach for estimating parametric structural break models that performs well. We suggest that iterative estimation methods are a simple but important feature of this approach when modelling seasonal data. The methodology is illustrated by simulation and then used for an analysis of monthly short term visitor arrival time series to New Zealand, to assess the effect of the 9/11 terrorist attacks. While some historical events had a marked structural effect on trends in those arrivals, we show that 9/11 did not.

Poster 35

Adil Yousif

Qatar University

aeyousif@qu.edu.qa

Dept. of Math, College of Arts and Science, P. O. Box 2713, Qatar University, Doha, Qatar

Measures of Performance Evaluation and Ranking

Assessment of school performance after applying certain method(s) or making some changes in classroom practices and ranking of several schools are always a concern for educators. In statistics education there are several measures used to assess the progress of a school or grade-level performance and ranking schools in a district or a whole country or even a whole region. Examples of these measures are: Effect Size, Confidence Interval, Correlation Coefficient, Value-added, Significance of the t-test. This study was aiming to look at each one of these measures and make a comparison of their degrees of efficiency.

In this paper it is suggested that the use of the t-statistics value or the p-value together with the signs of performance difference to both assess the progress of each school or grade-level over several years as well as ranking schools or groups can be an alternative. The effect of the sample size of each group involved in the assessment or ranking on each measure will be discussed.

An example based on a real performance data for several schools is used, and an interpretation of the p-value for assessing and ranking of the schools in descending order based on their performance. Results of other measures, for the same sample, will be displayed for a comparison.

Poster 36

Marika Vezzoli

University of Brescia

vezzoli@eco.unibs.it

Contrada Santa Chiara, 50, 25122 Brescia - Italy

Joint authors: Roberto Savona

Multidimensional Distance to Collapse Point and Sovereign Default Prediction

This paper focuses on predictability of sovereign debt crisis proposing a two-step procedure centered on the idea of a multidimensional distance-to-collapse point. The first step is nonparametric and devoted to construct a generalized early warning system that signals a potential crisis every time a group of indicators exceeds specific thresholds. The second is parametric and tries to contextualize the country default within a theoretical-based process depending on the distance from the thresholds estimated in the first step. Such regression approach helps to understand how non-parametric distances could be mixed together in order to reduce the n-dimensional measures of fundamentals, then obtaining a "generalized" distance to normality which summarizes all relevant information to better predict the likelihood of default for a given country across time.

From a purely technical viewpoint, our work is closed to the recent paper by Manasse and Roubini (2007) which has also inspired the searching in a new statistical algorithm, called CRAGGING (CRoss-validation AGGregatING) and formally proposed in Vezzoli and Stone (2007), that proves to be able to remove some important limitations of traditional data mining approaches, namely the Classification And Regression Trees (CART). While Manasse and Roubini (2007) straightforwardly implement the binary recursive tree approach, we go further by using the generalization of the original CART introduced in Vezzoli and Stone (2007), then obtaining a data mining approach tailored to the specific data structure we dealt with.

The empirical analyses provide convincing evidence on the ability of our approach in predicting a potential crisis. More precisely, in the first step we are able to construct homogeneous groups conditional on signals and since the number of the groups is obtained through an optimization procedure, the approach delivers a potential sovereign risk rating system with corresponding empirical default probabilities conditional on signals by construction. Again, the regression approach run in the second step acts as a useful tool to show how non-parametric distances could be mixed together in order to generalize and parameterize the multidimensional distance to normality. We find that over 22 potential predictors, 7 variables are sufficient to achieve significant share of correctly classified observations, all pertaining to traditional categories commonly used to assess the debt-service capacity of a sovereign. When considering thresholds of 0.05, 0.1, 0.25 and 0.5, the corresponding percentage of correctly classified defaults are 0.8182, 0.8030, 0.6818 and 0.5152.

Poster 37

Dr Philip E Cheng

Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

pcheng@stat.sinica.edu.tw

128 Academia Road, Sec 2, Taipei, Taiwan ROC

Joint authors: Dr Michelle Liou, Mr Jiun W Liou, Dr John AD Aston

Elements of Linear Information Models

An information theoretical approach to analysing contingency tables is examined. Relative entropy identities are used to yield orthogonal information decompositions and linear information models in contrast to the hierarchical log-linear models. It includes developing simple schemes for selecting and testing information models plus maximum likelihood estimation of the usual odds ratio parameters of the selected models. Empirical studies with high-way contingency tables illustrate further advantages in selecting information models for a wider application.

Poster 38

Ms Diane Berry

MRC Centre of Epidemiology for Child Health, UCL Institute of Child Health

d.berry@ich.ucl.ac.uk

30 Guildford Street, LONDON, WC1N 1EH

Joint authors: Dr M Cortina Borja, Dr E Hyppönen,

Effect of vitamin D status on seasonal patterns of metabolic markers

Vitamin D is produced by sunlight induced synthesis in the skin after exposure to UV-B radiation. It can be obtained from diet, however sun exposure is the key source. Vitamin D status (measured by serum 25(OH)D concentration) shows strong seasonal patterns and seasonal variations in metabolic markers are commonly used to hypothesise a link with vitamin D. To our knowledge the effect of adjustment for 25(OH)D concentrations on the seasonal patterns in metabolic markers has not previously been investigated.

Using data from the 1958 British birth cohort collected from August 2002 to March 2004, we set out to establish seasonal variations in three metabolic markers (tPA, D-dimer and vWf) and the effect of variations in 25(OH)D on the observed patterns. We modelled yearly changes using harmonic functions adjusted by confounding factors with laboratory assay batch as a random effect. The models were then adjusted by 25(OH)D in two ways: as a continuous measure of 25(OH)D, and through fitted values from the seasonal models for 25(OH)D. The optimal models for the inflammatory markers were selected in order to assess their association with 25(OH)D and were re-tested for seasonal patterns.

We present the final models that were used to predict the mean levels of markers from the partial regression coefficients of the seasonal pattern. Our results imply a role of vitamin D status in establishing the seasonal patterns of tPA and show a significant association between this marker and 25(OH)D. This finding may suggest that improved vitamin D status helps to maintain anti-thrombotic homeostasis; a mechanism which could be linked to the proposed beneficial effects on cardiovascular health.

Poster 39

Yu-Kang Tu

University of Leeds

y.k.tu@leeds.ac.uk

Biostatistics unit, Centre for Epidemiology and Biostatistics, Worsley Building, Clarendon Way, Leeds LS2 9JT

Testing the relationship between the percentage change and baseline value

How to test the relationship between change and initial value has been a controversial issue in the statistical literature, and this controversy is probably due to a misunderstanding of a statistical phenomenon, regression to the mean. It has been suggested that methods, such as Oldham's method, that attempt to test the equality of variances between baseline and follow-up values are more appropriate than methods, such as Blomqvist's formula, that attempt to adjust for errors in the measurements (Tu & Gilthorpe 2007 Statistics in Medicine). However, another related problem regarding testing the relationship between percentage change and baseline value still remains unresolved. Suppose x is the baseline value, y the post-treatment value, and the percentage change is defined as (x – y)/x. Testing the relation between percentage change and initial value using correlation or regression analysis suffers the same criticism as testing the relation between change and baseline did due to mathematical coupling. Mathematical coupling occurs when one variable directly or indirectly contains the whole or part of another, and the two variables are then analysed using correlation or regression. As a result, the statistical procedure of testing the null hypothesis – that the coefficient of correlation or the slope of regression is zero – might no longer be appropriate, and the results need to be interpreted cautiously. However, it is not clear how to obtain the correct null hypothesis and how to conduct statistical testing against it.

The aim of this presentation is therefore to show that when the following assumptions are satisfied, the correct null hypothesis for testing the percentage change and baseline value can be obtained by using the formula first given by Karl Pearson one century ago: neither the mean of x nor y is zero, and the ratio of the standard deviation to the mean is less than one. We will use two examples to illustrate how to obtain the correct null hypothesis for significance testing. Our results show that the incorrect application of correlation or regression analysis to the testing of the relationship between the percentage change and baseline value can potentially give rise to misleading results, and that our proposed new approach yield similar results compared to computer simulations.

Poster 40

Alexis Boukouvalas

Aston University

boukouva@aston.ac.uk

Neural Computing Research Group, Aston University
Aston Triangle, Birmingham, B4 7ET

Joint authors: Dr D. Maniyar, Dr D. Cornford and Dr A. Singer

Gaussian process emulation of stochastic models: developments and application to rabies modelling

Emulators are commonly used for statistical analysis of complex simulator models where the computational complexity of the simulator makes direct analysis infeasible. Emulators are typically implemented using Gaussian processes employing a finite set of design points at which the simulator is evaluated. It is typically assumed the simulator is a deterministic mapping justifying exact interpolation. However stochastic models are becoming increasingly commonplace to reflect either intrinsic physical process variability or uncertainty in our knowledge or implementation of the underlying process. In this work we briefly review existing work on stochastic emulation, which is a relatively new field, and typically requires the simulator to be run multiple times at a given design point. We propose a method where repeated observations at a given design point are not required, but can be used where this makes inference more efficient. We employ two coupled Gaussian processes which are used to emulate the first two moments of the simulator. Inference is performed in an iterative fashion using an expectation-maximisation like algorithm. Our method develops the most likely heteroscedastic Gaussian process regression method published in the machine learning literature. Our contribution extends this method to: allow both repeated simulator runs and single realisations; take into account uncertainty introduced by finite sample sizes; and, applies the method within a wider emulation framework.

We demonstrate the flexibility and robustness of our method on a complex stochastic simulator of rabies in a two vector species disease model. Our exploratory analysis of the model reveals heteroscedasticity in the response with certain factors crucially affecting model behaviour. We show how screening and linear dimensionality reduction methods can help us better understand the importance of different factors and their interactions in determining model output. We conclude with a discussion of our findings and suggestions for future work.

INDEX OF PRESENTING AUTHORS