Serial no.	Decision	Code
10	Accepted: Invited	
11	Accepted: Oral	
13	Accepted: Oral	
15	Accepted: Oral	
16	Accepted: Oral	
17	Accepted: Oral	
18	Accepted: Poster	
20	Accepted: Poster	
21	Accepted: Poster	
22	Accepted: Oral	
23	Accepted: Oral	
24	Accepted: Oral	
25	Accepted: Oral	
26	Accepted: Poster	
27	Accepted: Oral	
29	Accepted: Oral	
30	Accepted: Oral	
31	Accepted: Oral	
32	Accepted: Oral	
33	Accepted: Oral	
34	Accepted: Oral	
35	Accepted: Oral	
36	Accepted: Invited	
37	Accepted: Poster	
38	Accepted: Oral	
39	Accepted: Poster	
40	Accepted: Oral	
41	Accepted: Oral	
43	Accepted: Oral	
44	Accepted: Oral	
46	Accepted: Oral	
47	Accepted: Oral	
48	Accepted: Oral	
49	Accepted: Invited	
50	Accepted: Oral	
51	Accepted: Oral	
52	Accepted: Oral	
53	Accepted: Invited	
54	Accepted: Poster	
55	Accepted: Oral	
56	Accepted: Invited	
57	Accepted: Oral	
58	Accepted: Oral	
59	Accepted: Oral	
60	Accepted: Oral	

61 Accepted: Oral
62 Accepted: Oral
63 Accepted: Oral
64 Accepted: Oral
65 Accepted: Oral
66 Accepted: Oral
68 Accepted: Oral
69 Accepted: Oral
71 Accepted: Oral
72 Accepted: Oral
73 Accepted: Oral
74 Accepted: Oral
75 Accepted: Poster
76 Accepted: Oral
77 Accepted: Oral
78 Accepted: Poster
79 Accepted: Poster
80 Accepted: Oral
81 Accepted: Invited
82 Accepted: Invited
83 Accepted: Poster
84 Accepted: Oral
85 Accepted: Invited
86 Accepted: Oral
87 Accepted: Invited
88 Accepted: Oral
89 Accepted: Oral
90 Accepted: Oral
91 Accepted: Invited
92 Accepted: Oral
93 Accepted: Oral
94 Accepted: Oral
95 Accepted: Oral
96 Accepted: Oral
97 Accepted: Oral
98 Accepted: Oral
99 Accepted: Oral
100 Accepted: Oral
101 Accepted: Oral
102 Accepted: Oral
103 Accepted: Oral
104 Accepted: Oral
106 Accepted: Oral
107 Accepted: Poster
108 Accepted: Oral
109 Accepted: Oral

110 Accepted: Oral
111 Accepted: Invited
112 Accepted: Oral
113 Accepted: Oral
114 Accepted: Invited
115 Accepted: Oral
117 Accepted: Oral
118 Accepted: Invited
119 Accepted: Oral
120 Accepted: Oral
121 Accepted: Oral
122 Accepted: Poster
124 Accepted: Oral
126 Accepted: Oral
127 Accepted: Poster
128 Accepted: Oral
129 Accepted: Oral
130 Accepted: Oral
131 Accepted: Oral
132 Accepted: Oral
133 Accepted: Oral
134 Accepted: Poster
135 Accepted: Oral
136 Accepted: Oral
137 Accepted: Poster
138 Accepted: Oral
139 Accepted: Invited
140 Accepted: Invited
141 Accepted: Invited
142 Accepted: Oral
143 Accepted: Poster
144 Accepted: Oral
145 Accepted: Oral
146 Accepted: Poster
147 Accepted: Poster
148 Accepted: Oral
149 Accepted: Oral
150 Accepted: Oral
151 Accepted: Oral
152 Accepted: Oral
153 Accepted: Oral
154 Accepted: Poster
155 Accepted: Oral
156 Accepted: Oral
157 Accepted: Poster
158 Accepted: Oral

160 Accepted: Oral
161 Accepted: Oral
162 Accepted: Oral
163 Accepted: Oral
164 Accepted: Oral
165 Accepted: Oral
167 Accepted: Poster
168 Accepted: Oral
170 Accepted: Oral
172 Accepted: Oral
173 Accepted: Oral
174 Accepted: Invited
175 Accepted: Oral
176 Accepted: Oral
177 Accepted: Oral
178 Accepted: Oral
179 Accepted: Oral
180 Accepted: Oral
181 Accepted: Oral
182 Accepted: Oral
183 Accepted: Oral
184 Accepted: Oral
185 Accepted: Oral
186 Accepted: Invited
187 Accepted: Oral
188 Accepted: Oral
189 Accepted: Poster
190 Accepted: Oral
191 Accepted: Oral
192 Accepted: Oral
193 Accepted: Oral
194 Accepted: Oral
195 Accepted: Poster
196 Accepted: Poster
197 Accepted: Oral
198 Accepted: Oral
200 Accepted: Oral
201 Accepted: Oral
202 Accepted: Oral
203 Accepted: Oral
204 Accepted: Oral
205 Accepted: Oral
206 Accepted: Oral
207 Accepted: Oral
208 Accepted: Oral
209 Accepted: Oral

210 Accepted: Oral
212 Accepted: Oral
213 Accepted: Oral
214 Accepted: Oral
216 Accepted: Poster
219 Accepted: Oral
220 Accepted: Poster
221 Accepted: Invited
223 Accepted: Oral
224 Accepted: Oral
225 Accepted: Oral
227 Accepted: Poster
228 Accepted: Oral
229 Accepted: Oral
232 Accepted: Oral
233 Accepted: Oral
234 Accepted: Oral
235 Accepted: Poster
236 Accepted: Oral
237 Accepted: Oral
238 Accepted: Oral
241 Accepted: Oral
242 Accepted: Oral
243 Accepted: Invited
244 Accepted: Oral
246 Accepted: Oral
248 Accepted: keynote
249 Accepted: keynote
250 Accepted: keynote
251 Accepted: keynote
252 Accepted: Invited
253 Accepted: keynote
254 Accepted: keynote
255 Accepted: keynote
256 Accepted: keynote
257 Accepted: keynote
258 Accepted: keynote
259 Accepted: keynote
260 Accepted: keynote
261 Accepted: Poster
262 Accepted: Poster
265 Accepted

10

Transnational statistics - motivation, methods and means

Mr Stephen Horn
Independent, Canberra, Australia

Abstract

Official statistics is not qualified by authority to collect, although understood conventionally as coming from national agencies through appropriate statute. With international agencies including significant statistical presence the reference has extended to activities where authority arises from consensual agreement among (national) parties as much as formal treaty obligations. The standing of the collecting agency, and the integrity of the issuing agency provide the user with guarantees concerning credibility of output. Quality however is hostage to policies of the respective boards, and willingness of national agencies to provide data requested.

This presentation will sketch a new order of official statistics to accompany speculation on forms of transnational democracy emerging from recent developments that go beyond national legislatures to manage in isolation.

Following Picketty's schema* I will outline a state of play - with emerging practices and demand indicators; set out some prerequisites for a mature production system; and focus on methods issues peculiar to this new order.

I conclude by connecting such speculation with the entirely concrete category of global statistics as expounded for instance by Rademacher, using statistics itself as a bridge.

*ref. Table 8, p306, *Thomas Picketty, Time for Socialism - dispatches from a world on fire, 2016-2021, Yale University Press*

Rademacher, W, 2019

11

Quasi-Score Matching Estimation for Spatial Autoregressive Model with Random Weights Matrix and Regressors

Dr Xuan Liang, Dr Tao Zou
The Australian National University, Canberra, Australia

Abstract

Due to the rapid development of social networking sites, the spatial autoregressive (SAR) model has played an important role in social network studies. However, the commonly used quasi-maximum likelihood estimation (QMLE) for the SAR model is not computationally scalable as the network size is large. In addition, when establishing the asymptotic distribution of the parameter estimators of the SAR model, both weights matrix and regressors are assumed to be nonstochastic in classical spatial econometrics, which is perhaps not realistic in real applications. Motivated by the machine learning literature, this paper proposes quasi-score matching estimation for the SAR model. This new estimation approach is still likelihood-based, but significantly reduces the computational complexity of the QMLE. The asymptotic properties of parameter estimators under the random weights matrix and regressors are established, which provides a new theoretical framework for the asymptotic inference of the SAR type models. The usefulness of the quasi-score matching estimation and its asymptotic inference are illustrated via simulation studies and real data analysis.

13

R-VGAL: A Sequential Variational Bayes Algorithm for Generalised Linear Mixed Models

Miss Bao Anh Vu, Dr David Gunawan, Associate Professor Andrew Zammit-Mangion
University of Wollongong, Wollongong, Australia

Abstract

Models with random effects, known as generalised linear mixed models (GLMMs), are useful for analysing clustered data. Parameter inference with these models is difficult because of the presence of cluster-specific random effects, which must be integrated out when evaluating the likelihood function. Here, we propose a sequential variational Bayes algorithm, called Recursive Variational Gaussian Approximation for Latent variable models (R-VGAL), for estimating parameters in GLMMs. The R-VGAL algorithm operates on the data sequentially, requires only a single pass through the data, and can provide updates as new data are collected without the need of re-processing the previous data. At each iteration, the R-VGAL algorithm requires the gradient and Hessian of a "partial" log-likelihood function evaluated at the new observation, which are generally not available in closed form for GLMMs. To circumvent this, we propose using an importance-sampling-based approach for estimating the gradient and Hessian using Fisher's and Louis' identities. We find that R-VGAL can be unstable when traversing the first few data points, but that this issue can be mitigated by using a variant of variational tempering in the initial steps of the algorithm. Through illustrations on both simulated and real datasets, we show that R-VGAL provides good approximations to the exact posterior distributions, that it can be made robust through tempering, and that it is much faster than Markov chain Monte Carlo.

15

Does living in a 20-minute neighbourhood influence food purchasing behaviours?

Ms Laura Oostenbach[1], Dr Alissa Burnett ORCID iD[1], A/Prof Karen Lamb ORCID iD[2], Dr Lukar Thornton ORCID iD[3]

[1]Deakin University, Melbourne, Australia. [2]University of Melbourne, Melbourne, Australia. [3]University of Antwerp, Antwerp, Belgium

Abstract

Introduction

Many cities globally have adopted the planning concept of the 20-minute neighbourhood, assuming that living within 20 minutes of key services promotes healthier living. Although a range of lifestyle benefits have been posited, evidence about whether these neighbourhoods promote healthy local living is lacking. Given emphasis on diet and physical activity behaviours, we examined whether living in a 20-minute neighbourhood (20MN) impacts the distance and travel mode to the primary supermarket.

Methods

Neighbourhood services in Victoria and South Australia were geocoded, with 20MNs identified as areas with 20-minute access to multiple key services. An online food behaviour survey was completed by 769 adults living in 20MNs and non-20MNs in Melbourne and Adelaide in 2018-2019. Outcomes were (log-transformed) distance from home and mode of transport to the primary supermarket. Spatial autoregressive models were fitted to compare distances between those in 20MNs and non-20MNs, including spatially lagged errors using the inverse distance matrix between participant home addresses.

Results

The median distance travelled to the primary supermarket was longer in non-20MNs (Melbourne-3.4 km, Adelaide-3.0 km) than 20MNs (Melbourne-1.4 km, Adelaide-1.4 km). Modelling indicated those living in 20MNs travelled 51% (95% CI: 42-60%) and 57% (95% CI: 49-64%) shorter distances compared to those in non-20MNs in Melbourne and Adelaide, respectively. Car was the most common mode of transport, although the percentage was much higher among those from non-20MNs than 20MNs (Melbourne: 94% vs. 61%, Adelaide: 98% vs. 78%). More participants in 20MNs reported walking (Melbourne: 30% vs. 1%, Adelaide: 19% vs. 1%).

Conclusions

This study provides evidence to support the claim that 20MNs promote more localised living, with those living in 20MNs travelling shorter distances to their primary supermarket. Furthermore, although car use was the dominant form of transport in both 20MNs and non-20MNs, a much higher proportion from 20MNs chose to walk.

16

The Block-Correlated Pseudo Marginal Sampler for State Space Models

Dr David Gunawan[1], Dr Pratiti Chatterjee[2], Prof Robert Kohn[2]
[1]University of Wollongong, NSW, Australia. [2]UNSW, NSW, Australia

Abstract

Particle Marginal Metropolis-Hastings (PMMH) is a general approach to Bayesian inference when the likelihood is intractable, but can be estimated unbiasedly. Our article develops an efficient PMMH method, which we call the multiple PMMH (MPM) algorithm, for estimating the parameters of complex state space models. Several important innovations are proposed. First, the multiple particle filters are run in parallel andthe trimmed means of their unbiased likelihood estimates are used. Second, a novel block version of PMMH that works with multiple particle filters is proposed. Third, the article develops an efficient auxiliary disturbance particle filter, which is necessary when the bootstrap filter is inefficient, but the state transition density cannot be expressed in closed form. Fourth, a novel fast sorting algorithm is developed to preserve the correlation between the logs of the likelihood estimates at the current and proposed parameter values. These features enable our sampler to scale up better to higher dimensional state vectors than previous approaches. The performance of the sampler is investigated empirically by applying it to non-linear Dynamic Stochastic General Equilibrium models with relatively high state dimensions and with intractable state transition densities and to multivariate stochastic volatility in the mean models. Although our focus is on applying the method to state space models, the approach will be useful in a wide range of applications such as large panel data models and stochastic differential equation models with mixed effects.

17

Neural Point Estimation for Fast Optimal Likelihood-Free Inference

Mr Matthew Sainsbury-Dale ORCID iD[1], Prof Andrew Zammit-Mangion ORCID iD[1], Prof Raphaël Huser[2]

[1]University of Wollongong, Wollongong, Australia. [2]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Abstract

Neural point estimators are neural networks that map data to parameter point estimates. They are substantially faster than classical methods, likelihood-free and, due to their amortised nature, amenable to fast bootstrap-based uncertainty quantification. In this talk, we provide an overview of this relatively new and powerful inferential tool, from its foundations in classical decision theory to its implementation with user-friendly open-source software. We illustrate the framework on Gaussian processes and max-stable models for spatial extremes, and we show that it substantially outperforms the composite likelihood methods typically employed for making inference with the latter class of models.

18

Statistical analyses of ordinal outcomes in randomised controlled trials: a scoping review

Chris Selman ORCID iD[1,2], Prof Katherine Lee[2,1], Kristin Ferguson[1], Dr Clare Whitehead[1,3], Dr Brett Manley[1,3,2], Dr Robert Mahar[1,2]
[1]University of Melbourne, Parkville, Australia. [2]Murdoch Children's Research Institute, Parkville, Australia. [3]Royal Women's Hospital, Parkville, Australia

Abstract

Abstract submission

Download file

20

Scalable Posterior Sampling from Gaussian Mixture Models via Randomly Weighted Expectation-Conditional-Maximisation

Mr. Santiago Marin, Dr. Anton Westveld, Dr. Bronwyn Loong
The Australian National University, Canberra, Australia

Abstract

Sampling from the joint posterior distribution of Gaussian mixture models (GMMs) via standard Markov chain Monte Carlo (MCMC) imposes a number of computational challenges, which have prevented a broader full Bayesian implementation of these models. By definition, MCMC draws are correlated, may get trapped in areas of high posterior density — leading to mixing limitations between posterior modes — and require a large number of expensive linear algebra operations. Thus, we propose a method to sample, in a scalable fashion, from an approximated joint posterior distribution of GMMs. We build on recent weighted Bayesian bootstrap (WBB) ideas, and combine them with a tempered Expectation-Conditional-Maximisation (ECM) algorithm to compute maximum a posteriori (MAP) estimates on many independently randomised objective posterior functions. Given the non-convex nature of these objective functions, the inclusion of a tempering profile reduces the risk of landing in sub-optimal modes. Our proposed method generates approximate posterior draws that are independent, explores the entire posterior distribution, enables uncertainty quantification and, by making use of modern numerical optimisation algorithms, reduces the number of expensive linear algebra operations. We demonstrate the performance of our method and compare it with competing approaches through extensive simulations, in addition to a real-world data set.

21

Intimate Partner Violence and Mental Health Effects: A population-based study among married women in Sri Lanka.

Mrs Lakma Gunarathne ORCID iD, Associate Professor Jahar Bhowmik ORCID iD, Professor Maja Nedeljkovic ORCID iD, Dr. Pragalathan Apputhurai ORCID iD
Swinburne University of Technology, Melbourne, Australia

Abstract submission

Download file

22

Forecasting high-dimensional functional time series: Application to subnational age-specific mortality

Mr Cristian Jimenez-Varon ORCID iD[1], Prof Ying Sun ORCID iD[1], Prof Hanlin Shang ORCID iD[2]
[1]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. [2]Macquarie University, Sydney, Australia

Abstract

We consider modeling and forecasting high-dimensional functional time series (HDFTS), which can be cross-sectionally correlated and temporally dependent. We present a novel two-way functional median polish decomposition, which is robust against outliers, to decompose HDFTS into deterministic and time-varying components. A functional time-series forecasting method, based on dynamic functional principal component analysis, is implemented to produce forecasts for the time-varying components. By combining the forecasts of the time-varying components with the deterministic components, we jointly obtain forecast curves for multiple populations. Illustrated by the sex- and age-specific mortality rates in the US, France, and Japan, which contain 51 states, 95 departments, and 47 prefectures, respectively, the proposed model delivers more accurate point and interval forecasts in forecasting multi-population mortality than several benchmark methods.

23

Spatial prediction of solar-induced chlorophyll fluorescence (SIF) from multivariate OCO-2 data

Josh Jacobson ORCID iD, Noel Cressie ORCID iD, Andrew Zammit-Mangion ORCID iD
University of Wollongong, Wollongong, Australia

Abstract

Satellite remote sensing has become a key tool for observing and understanding a plethora of environmental processes. For complete representation of these processes, scientists need to create smoothed and gap-filled (i.e., “Level 3”) products from sparse, big, and often multivariate remote sensing data. However, these Level 3 products are constructed one variable at a time and are rarely accompanied with a quantification of uncertainty. Here, we show how multivariate spatial statistics can be used to construct Level 3 products from multivariate remote sensing data by borrowing strength across variables for improved predictions and formal uncertainty quantification. Our spatial-statistical methodology uses a combination of spatial basis functions to capture the spatial trend and a version of (co)kriging that accounts for nonstationary measurement errors, resulting in a Level 3 data product with accompanying prediction standard errors. We illustrate our approach on solar-induced chlorophyll fluorescence (SIF), part of the natural process of photosynthesis. NASA’s Orbiting Carbon Observatory-2 (OCO-2) satellite offers SIF retrievals made simultaneously with its retrievals of column-averaged atmospheric carbon dioxide (XCO_2) concentrations, which we consider to be noisy and incomplete measurements of a bivariate latent process. We compare kriging and cokriging predictions with a spatial-trend-surface analysis using cross-validation. Our fine-resolution SIF product can be easily aggregated to coarse resolutions along with a coherent quantification of uncertainty. Hence, our results could also allow statistical comparisons to be made with other SIF products at varying resolutions.

24

PUMP: Estimating power when adjusting for multiple outcomes in multi-level experiments

Dr. Kristen Hunter ORCID iD[1], Dr. Luke Miratrix ORCID iD[2], Dr. Kristin Porter[3]
[1]UNSW, Sydney, Australia. [2]Harvard Graduate School of Education, Cambridge, USA. [3]K.E. Porter Consulting LLC, Berkeley, USA

Abstract

For randomized controlled trials (RCTs) with a single intervention's impact being measured on multiple outcomes, researchers often apply a multiple testing procedure (such as Bonferroni or Benjamini-Hochberg) to adjust p-values. Such an adjustment reduces the likelihood of spurious findings, but also changes the statistical power, sometimes substantially, which reduces the probability of detecting effects when they do exist. However, this consideration is frequently ignored in typical power analyses, as existing tools do not easily accommodate the use of multiple testing procedures. We introduce the PUMP R package as a tool for analysts to estimate statistical power, minimum detectable effect size, and sample size requirements for multi-level RCTs with multiple outcomes. PUMP uses a simulation-based approach to flexibly estimate power for a wide variety of designs, number of outcomes, multiple testing procedures, and other user choices. By assuming mixed effects linear models, we can draw directly from the joint distribution of test statistics across outcomes in order to estimate power via simulation of the test statistics. One of PUMP's main innovations is accommodating multiple outcomes, which are accounted for in two ways. First, power estimates from PUMP properly account for the adjustment in p-values from applying a multiple testing procedure. Second, as researchers change their focus from one outcome to multiple outcomes, different definitions of statistical power emerge. PUMP allows researchers to consider a variety of definitions of power in order to choose the most appropriate types of power for the goals of their study. The package supports a variety of commonly-used frequentist multi-level RCT designs and linear mixed effects models. In addition to the main functionality of estimating power, minimum detectable effect size, and sample size requirements, the package allows the user to easily explore sensitivity of these quantities to changes in underlying assumptions.

25

VAR-SOPHE - Selecting variables that maximise variability between cluster patterns of different datasets

Ms Lucy Conran ORCID iD[1], Professor Inge Koch[2], A/Professor Berwin Turlach[1]
[1]University of Western Australia, Perth, Australia. [2]RMIT, Melbourne, Australia

Abstract

We study statistical approaches for the estimation of variability within subjects, between subjects of the same groups and between subjects of different groups. Our data are multivariate flow cytometry measurements, typically 40-100 thousand blood cells in 5-20 variables each.

We describe our variable selection method, VAR-SOPHE. VAR-SOPHE ranks variable subsets by their ability to maximise variability between datasets. We first bin the data and then use a modified version of the clustering method SOPHE of Jing et al (2021) to create cluster patterns. We compare these cluster patterns and select the best variable subsets for determining difference.

We apply the VAR-SOPHE approach to flow cytometry measurements of subjects with and without diseases. We illustrate these results with suitable visualisations of the range of variability produced by different subsets of variables.

Reference.

J. Jing, I. Koch, and K. Naito (2012). Polynomial Histograms for Multivariate Density and Mode Estimation. Scandinavian Journal of Statistics 39, 75-96.

Abstract submission

Download file

26

A comparison of strategies for selecting auxiliary variables for multiple imputation

Dr Rheanna Mainzer ORCID iD[1,2], Dr Cattram Nguyen[1,2], Prof John Carlin[1,2], AProf Margarita Moreno-Betancur[1,2], Prof Ian White[3], Prof Katherine Lee[1,2]
[1]Murdoch Children's Research Institute, Parkville, Australia. [2]The University of Melbourne, Parkville, Australia. [3]University College London, High Holborn, United Kingdom

Abstract

Abstract submission

Download file

27

Sama Radial Indicator (SRI) for Measuring Latent Variables

Dr. Samanthi Konarasinghe Wiriththamulle Gamage ORCID iD, Dr Edward Mariyani-Squire
Western Sydney University, Parramatta, NSW 2150, Australia

Abstract

Attached below

Abstract submission

Download file

29

Citi Bike: an Authentic Assessment for Statistics & Data Science

Dr Matt Moores ORCID iD, Dr Carole Birrell
University of Wollongong, Wollongong, Australia

Abstract

The benefits of integrating real-world datasets into the teaching of statistics and data science undergraduate degrees are widely recognized: it can increase student motivation towards studying statistics as they engage with the data in context; it promotes statistical thinking and active learning. However, finding suitable sources of data for this purpose can be difficult and time-consuming. The bicycle sharing service Citi Bike in New York City makes its data publicly available for download as comma-separated value (CSV) files. This is an excellent resource for teachers of statistics: a rich set of data with individual bike trip characteristics organised by date.

We have used the Citi Bike dataset in our first-year Introduction to Statistics subject, which has been offered as a general elective since 2018. The characteristics of these data are well suited for providing authentic data experiences, including: importing the data into the statistical software; transforming and computing additional variables; identifying missing or unreliable observations; splitting the data; and merging datasets. Active and collaborative learning takes place as students gain hands-on experience in analysing data from industry.

The students work in groups to analyse the data, with each team member drawing a simple random sample of trips, producing numerical and graphical summaries for a different day of the week, then combining their results to make comparisons. A combined statistical report is submitted for assessment. We discuss how we scaffold towards this assessment, as well as how it serves the learning outcomes of our mathematics and data science degrees, including our Work-Integrated Learning (WIL) Curriculum Classification Framework.

30

Decaying correlation parameter values obtained from previously analysed cluster randomised trials

A/Prof Jessica Kasza[1], Dr Rhys Bowden[1], Dr Yongdong Ouyang[2,3], Prof Monica Taljaard[2,3], Prof Andrew Forbes[1]

[1]Monash University, Melbourne, Australia. [2]Ottawa Hospital Research Institute, Ottawa, Canada. [3]University of Ottawa, Ottawa, Canada

Abstract

A frequently-applied assumption in the analysis of data from cluster randomised trials is that the outcomes from all participants within a cluster are equally correlated. That is, the intracluster correlation, which describes the degree of dependence between outcomes from participants in the same cluster, is the same for each pair of participants in a cluster. However, recent work has discussed the importance of allowing for this correlation to decay as the time between the measurement of participants in a cluster increases. Incorrect omission of such a decay has implications for the design and analysis of cluster randomised trials: for example, confidence intervals for estimated treatment effects can be too narrow or too wide, depending on the characteristics of the design. When planning studies, researchers often rely on previously reported analyses of trials to inform their choice of intracluster correlation. While some estimates of decaying correlations are available, most reported analyses of clustered data do not incorporate a decay. Researchers planning a trial under a decaying correlation may then face the challenge of specifying plausible combinations of decaying correlation parameters that are compatible with the intracluster correlation value obtained under no decay.

In this talk we show that it is possible to use intracluster correlation values obtained from models that incorrectly omit a decay to inform plausible choices of decaying correlations. Our focus is on intracluster correlation estimates for continuous outcomes obtained by fitting linear mixed models. We describe how plausible values for decaying correlations (i.e. those correlations associated with discrete time decay correlation structures) may be obtained given these estimated intracluster correlations. We will demonstrate an online app that allows users to obtain these consistent values. This can be used at the trial planning stage to assess the sensitivity of sample size and power calculations to decaying correlation structures.

31

A Double Fixed Rank Kriging Approach to Spatial Regression Models with Covariate Measurement Error

Mr. Xu Ning ORCID iD, Dr. Francis Hui, Professor Alan Welsh
ANU, Canberra, Australia

Abstract

In many applications of spatial regression modeling, the spatially-indexed covariates are measured with error, and it is known that ignoring this measurement error can lead to attenuation of the estimated regression coefficients. Classical measurement error techniques may not be appropriate in the spatial setting, due to the lack of validation data and the presence of (residual) spatial correlation among the responses. In this article, we propose a double fixed rank kriging (FRK) approach to obtain bias-corrected estimates of and inference on coefficients in spatial regression models, where the covariates are spatially indexed and subject to measurement error. Assuming they vary smoothly in space, the proposed method first fits an FRK model regressing the covariates against spatial basis functions to obtain predictions of the error-free covariates. These are then passed into a second FRK model, where the response is regressed against the predicted covariates plus another set of spatial basis functions to account for spatial correlation. A simulation study and an application to presence-absence records of Carolina wren from the North American Breeding Bird Survey demonstrate that the proposed double FRK approach can be effective in adjusting for measurement error in spatially correlated data.

32

Exact sample sizes for clinical trials

Professor Chris Lloyd
University of Melbourne, Melbourne, Australia

Abstract

I briefly describe the difficulties in providing required sample sizes for clinical trials that guarantee type 1 and type 2 error control. Required sample sizes depend on the test statistic assumed and in this study we use the so-called E-test, which is known to have extremely favourable size properties and higher power and alternatives. To compute exact powers for this test in real time is not currently feasible so a corpus of pre-computed exact powers (and sizes) was created, covering sample sizes up to 500. When there are no solutions within the corpus, a novel extrapolation technique is used. Exact size can be computed after the sample sizes have been extracted, however for the E-test the exact size is virtually always very close to the nominal target. All the code has been converted into an R-package exact.n which is available on CRAN and illustrated.

33

Copula-based hierarchical spatial statistical models

Mr Alan Pearse, Dr David Gunawan, Distinguished Professor Noel Cressie
School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, Australia

Abstract

In hierarchical spatial statistical modelling, noisy and incomplete spatial data are used to perform optimal inference on an underlying latent spatial process. A commonly used latent-process model assumes it is the sum of a mean process plus a zero-mean Gaussian error process with a stationary covariance function of spatial lags. A commonly used data model assumes the measurement errors are additive and follow a Gaussian white-noise process. In some cases (e.g., for count data), a spatial generalised linear mixed model (SGLMM) construction can be used instead, where the data model is a member of the exponential family of distributions, and a function of its parameters is a latent Gaussian process model. While Gaussian spatial processes are commonly used (for reasons of familiarity, tractability, and additivity), spatial statistical modellers are looking for increased flexibility to handle any type of observed data and latent spatial process, which can be discrete or continuous. Spatial copula models allow this flexibility, although they have not been developed within a hierarchical statistical framework that distinguishes between the data and the process. In this paper, we present a fully Bayesian, copula-based, hierarchical spatial statistical model and hence expand the range of models available for skewed, bounded, and other non-Gaussian spatial data. Any admissible copula model, such as the Gaussian copula, t-copula, and vine copula can be used within our hierarchical spatial statistical model. These multivariate copula models can capture complex, non-Gaussian forms of spatial dependence in the latent process. SGLMMs are shown to be contained in our class of copula-based spatial models. For our new class of copula-based hierarchical spatial models, we develop a Bayesian-inference framework from a carefully constructed Markov chain Monte Carlo (MCMC) sampler. Simulation studies and an application of the model to satellite-based atmospheric methane data are presented to establish its capabilities.

34

Synthetic data generation through the lens of a differential privacy framework

Dr Joseph Chien ORCID iD
ABS, Canberra, Australia

Abstract

National statistical offices (NSOs) have been using synthetic data as a solution which balances the need to provide access to quality information while maintaining confidentiality. For example, both Statistics New Zealand’s and Statistics Canada have released synthetic census-based data. While there have been many disclosure risk measures proposed for the synthetic data in the literature, there has been limited research on developing a particular differential privacy (DP) mechanism for a synthetic data generation. This joint research explores how we can apply a DP framework in a fully conditional specification synthetic data generation mechanism. By quantifying a mechanism in terms of DP parameters such as the privacy loss budget, this research will assist in informing NSOs' trade-off between utility and risk of releasing a synthetic dataset.

35

Exact Testing for Heteroscedasticity in a Two-Way Layout in Variety Frost Trials when Incorporating a Covariate

Miss Angelika Pilkington[1], Dr Brenton Clarke ORCID iD[1], Prof Dean Diepeveen[1,2]
[1]Murdoch University, Murdoch, Australia. [2]Department of Primary Industries and Regional Development, Western Australia, South Perth, WA, Australia

Abstract

Two-way layouts are common in grain industry research where it is often the case that there are one or more covariates. It is widely recognized that when estimating fixed effect parameters, one should also examine for possible extra error variance structure. An exact test for heteroscedasticity, when there is a covariate, is illustrated for a data set from frost trials in Western Australia. While the general algebra for the test is known, albeit in past literature, there are computational aspects of implementing the test for the two-way when there are covariates.

In this scenario the test is shown to have greater power than the industry standard, and because of its exact size, is preferable to use of the Restricted Maximum Likelihood Ratio Test (REMLRT) based on the approximate asymptotic distribution in this instance. Formulation of the exact test considered here involves creation of appropriate contrasts in the experimental design. This is illustrated using specific choices of observations corresponding to an index set in the linear model for the two-way layout. Also an algorithm supplied complements the test. Comparisons of size and power then ensue. The test has natural extensions when there are unbalanced data, and more than one covariate may be present. Results can be extended to Balanced Incomplete Block Designs.

36

Maximising uncertainty to create certainty: estimating Australia’s domestic business supply chains using entropy maximisation

Anthony Russo
Australian Bureau of Statistics, Adelaide, Australia

Abstract

Over recent years, the world has witnessed an increase in economic disruptions, ranging from natural disasters and trade embargoes to the COVID-19 pandemic and cyber attacks. Their impacts on national and international business supply chains present nations with sovereign risks and have demonstrated the vulnerabilities of modern integrated economies. Across the globe, governments are increasingly interested in measuring resilience in the business supply chain network, forecasting the impacts of emerging or potential disruptions, developing effective mitigation strategies, and facilitating economic recovery. The Australian Bureau of Statistics (ABS) has been undertaking an experimental project to evaluate the feasibility of estimating the domestic business supply chain network with a future view to enabling the estimation of the risks and likely impacts of economic shocks through the Australian economy. This is difficult to do with existing statistics that are not designed to capture the trading relationships between businesses and hence the dynamics of network interactions. Following methods implemented at the Dutch Central Bureau of Statistics and in the emerging body of literature on network reconstruction, the ABS has utilised supply-use tables from the national accounts and GST data on business sales and purchases to construct a probabilistic prototype supply chain network based on the method of entropy maximisation. Network reconstruction presents new opportunities to quantify complex economic and social systems. As a comprehensive map of the economy, a reconstructed network would enable the ABS to fill critical data gaps and provide new statistical tools in policymaking, macro-economics & supply chain research.

37

Random sampling is a mathematical necessity beyond debate or opinion for valid statistical inferences

Dr. John Xie ORCID iD
Charles Sturt University, Wagga Wagga, Australia

Abstract

This poster aims at reminding researchers of the unignorable fundamental concern of random sampling issue in statistical inferences. Random sampling refers to the process of clearly defining a target population and then a probabilistic sample(s) can be literally taken from the population. More specifically, a target population is the collection of all sampling units relevant to some well-defined research questions and, by some objective chance mechanism, each sampling unit has the equal or known probability to be selected into a sample which is referred to as 'a random sample'. Therefore, each sample unit can be weighted by the inverse of the selection probability to get unbiased estimates of the parameters that fully define the population. Statistical inferences aim at making justifiable conclusions about a population based on sample data. The random sampling thus becomes a matter of mathematical necessity rather than a matter of debate or opinion for any valid statistical inferential analysis. However, it is a matter of fact that practical and/or ethical factors make it literally impossible in most instances to obtain random samples for real life research. This unavoidable fundamental flaw alone invalidates the generalizability of research findings beyond the sample data on the ground of statistical inferences. Sadly, most textbook authors unduly dismiss (or simply ignore) the concerns of random sampling issue with statistical fables (e.g., 'your sample of observations can be imagined to come and may be regarded as a random sample'); hence most statistical analysis practices simply assume that the random sampling flaw in statistical inferences is negligible.

Abstract submission

Download file

38

Unsupervised Machine Learning for Anomaly Detection in Big Data

Dr Adam Leinweber ORCID iD
Australian Bureau of Statistics, Adelaide, Australia

Abstract

Single Touch Payroll (STP) is an Australian Government initiative that aims to reduce employers' reporting burdens to government agencies. The Australian Taxation Office (ATO) receives payroll information from employers and provides selected employer job level data items from the STP system to the ABS in order to produce statistics. The resultant dataset is large in scale and is updated with high frequency. In the interest of reducing human effort in data validation, an efficient semi-automated method for identifying anomalous entries is being assessed to inform human decision-making as part of a broader validation and editing process. Two unsupervised machine learning algorithms have been trained in the interest of flagging anomalous behaviour using previous reporting history as a baseline for normalcy. These two anomaly detection algorithms were chosen for their efficient performance, and explainability. The first is a local outlier factor, which compares a given point to the local density of points in the same space; and the other is an isolation forest, which operates by partitioning the space until a given point is isolated. These algorithms have been trained on an industry-by-industry basis, as each industry poses different anomalous behaviour. The trained algorithms can successfully and efficiently identify anomalies in large datasets and a prototype is in the process of being evaluated for use as part of the data validation pipeline.

39

Reliability Models Considering Non-identical Component Strength of Power-Muth Distribution

Prashant Kumar Sonker, Dr. Mukesh Kumar ORCID iD
Banaras Hindu University, Varanasi, India

Abstract

This article explores the extension of the stress strength reliability model of a system and the multi-component systems when the components of the system are considered to be non-identical. These components are separated into two categories. Each component of the system has some strength and the common random stress applied to it. The component strength of both the categories follows the power Muth (PM) distribution and the stress applied to the components also follows the PM distribution. The estimation of stress strength reliability and multi-component stress-strength reliability is carried out using well-known ML and MPS estimation methods. Based on varying parameters, the reliability of the models is discussed. All the statistical calculations are done by using Monte Carlo simulation. Real data applicability of the extended model is also performed in the article.

Abstract submission

Download file

40

Survey Design and Estimating Equations when Combining Big Data with Probability Samples

Dr Ryan Covey ORCID iD[1], Lucca Buonamano ORCID iD[2]
[1]Australian Bureau of Statistics, Melbourne, Australia. [2]Australian Bureau of Statistics, Sydney, Australia

Abstract

The use of big data in official statistics and the applied sciences is accelerating, but statistics computed using only big data often suffer from substantial selection bias. This leads to inaccurate estimation and invalid statistical inference. We rectify the issue for a broad class of linear and nonlinear statistics by producing estimating equations that combine big data with a probability sample. Under weak assumptions about an unknown superpopulation, we show that our integrated estimator is consistent and asymptotically unbiased with an asymptotic normal distribution. Variance estimators with respect to both the sampling design alone and jointly with the superpopulation are obtained at once using a single, unified theoretical approach. A surprising corollary is that strategies minimising the design variance almost minimise the joint variance when the population and sample sizes are large. The integrated estimator is shown to be more efficient than its survey-only counterpart if dependence between sample membership indicators is small and the finite population is large. We illustrate our method for quantiles, the Gini index, linear regression coefficients and maximum likelihood estimators where the sampling design is stratified simple random sampling without replacement. Our results are illustrated in a simulation of individual Australian incomes. Finally, we explore our integrated estimator using a linked ABS dataset of actual individual Australian incomes comprising the 2015-16 Survey of Income and Housing, 2015-16 income tax returns and the 2016 Census.

41

Best Subset Selection for Linear Dimension Reduction models using Continuous Optimization

Prof Benoit Liquet[1], Dr Sarat Moka[2], Prof Samuel Muller[1]
[1]School of Mathematical and Physical sciences, Macquarie University, Sydney, Australia.
[2]UNSW, sydney, Australia

Abstract

Choosing the most important variables in supervised and unsupervised learning is a difficult task, especially when dealing with high-dimensional data where the number of variables far exceeds the number of observations. In this study, we focus on two popular multivariate statistical methods - principal component analysis (PCA) and partial least squares (PLS) - both of which are linear dimensionality reduction techniques used in a variety of fields such as genomics, biology, environmental science, and engineering. Both PCA and PLS generate new variables, known as principal components, that are combinations of the original variables. However, interpreting these components can be challenging when working with large numbers of variables. To address this issue, we propose a method that incorporates the best subset selection approach into the PCA and PLS frameworks using a continuous optimization algorithm. Our empirical results demonstrate the effectiveness of our method in identifying the most relevant variables. We illustrate the use of our algorithm on two real datasets - one analyzed using PCA and the other using PLS.

43

How does the disease activity of lupus change over time? Variance splitting

Dr Ning Li ORCID iD
Monash University, Melbourne, Australia

Abstract

Disease activity is central to lupus research. Understanding disease activity in lupus helps to guide treatment decisions, evaluate the effectiveness of new treatments, and ensure appropriate patient care. The disease activity of lupus usually varies over time, with mixed periods of flares and remissions of varying lengths.

After a data regulation process, we built a random-effects model of the yearly change in disease activity of lupus on the lagged disease activity and various covariates in the panel data framework and obtained results that were largely compatible with other studies.

However, regression diagnosis questioned the model's adequacy. Why was the model invalid, and what was the cause of the problem? Was any critical variable missed out in the data collection? Was the relationship under study nonlinear? Were the individual-specific effects correlated with the covariates in the model?

Expert opinion excluded the first possibility, and a residual-versus-fitted value plot excluded the second. We found that the primary relevant information was indeed collected in the variables of the Asia Pacific Lupus Collaboration (APLC) cohort but not adequately extracted from the data. To better reveal the effect of a variable on disease activity evolvement, we recognized the possibility of different impacts from different sources by partitioning the total variance in an independent variable into orthogonal parts - one reflecting the between-patient difference and one reflecting the within-patient variability.

In this talk, we shall demonstrate that decomposing the within-patient variation (i.e., variation in the variable within each patient over time) from the between-patient variation (i.e., variation in the variable between different patients) can better account for the likely unequal impacts on disease activity change from different sources, improve the model accuracy, and lead to a new clinic finding.

Abstract submission

Download file

44

Gauging statistical literacy as students transition through university

Stephanie Budgett ORCID iD[1], Charlotte Jones-Todd ORCID iD[1], Amy Renelle ORCID iD[2]
[1]The University of Auckland, Auckland, New Zealand. [2]The University of Waikato, Hamilton, New Zealand

Abstract

Statistical literacy is essential in order to participate as an informed citizen. Students transition through, and beyond, university relying on disciplinary knowledge and statistical literacy skills to make sense of today's data-driven world. From deducing meaning from data visualisations in the media to interpreting an R_0 value, students will regularly encounter statistical concepts beyond their academic courses. As students progress through a hierarchical suite of applied statistics courses, we examine increasingly more complex statistical content. However, the development of their statistical literacy skills is not, necessarily, considered.

We interviewed three experienced statistics instructors asking what skills, they believed, were key for students. Common themes included (1) the ability to recognise and consider statistical information embedded in everyday situations, (2) the capacity to retain and transfer disciplinary knowledge to new contexts, and (3) the capability to clearly and confidently communicate to varied audiences.

This informed a small exploratory study and the development of several statistical tasks that were presented to students at different stages of their statistics education journey. Responses to these tasks enabled us to characterise students' statistical literacy skills and highlight the knowledge base that students, at different stages, tended to draw upon when faced with the same types of statistical scenarios. We found that, regardless of the stage of their study, students could generally navigate statistical scenarios well when set in routine or familiar contexts. However, few students demonstrated the versatility to extract and manipulate statistical information when presented in unexpected or non-traditional formats.

Our findings indicate a need to support students in applying their statistical literacy skills in novel and unfamiliar situations. We believe this is particularly important in the modern world where, given the proliferation of fake news, a statistically literate citizenship is vital.

46

The Potential of Consumer Data for Estimation of Household Expenditure Statistics

Mr Eugene Schon[1], Ms Lujuan Chen[2]
[1]Australian Bureau of Statistics, Melbourne, Australia. [2]Australian Bureau of Statistics, Canberra, Australia

Abstract

Big data represents an opportunity to generate real-time information products, complementing the depth of detail provided by validated surveys and censuses. This research examines the potential of 2019-2020 real time consumer data and assesses its utility in supporting the estimation of household expenditure statistics coherent with official ABS statistics under the Household Income and Expenditure Data program.

Methodology was developed to calculate weekly household expenditure estimates from a transaction level consumer dataset. Specifically, a graph-theory method was developed to form households from individual records. Text analysis was utilised to classify households into four broad categories, based on the presence of one or more than one adult and the presence or absence of children. Aggregated financial transactions were mapped to ABS expenditure classifications. The Iterative Proportional Update Algorithm was applied to weight the sample to ensure representativeness with respect to person and household level population benchmarks. Finally, weekly household expenditure was estimated and compared with household expenditure estimates extrapolated from the ABS 2015-2016 Household Expenditure Survey.

The results demonstrated that mean weekly household expenditure calculated from the consumer data aligned well with the Household Expenditure Survey at the national and state level except for Northern Territory. However, the consumer data estimates differed considerably from Household Expenditure Survey results at the 2-digit Household Expenditure Classification level.

47

Time series analysis and forecasting with Bayesian Dynamic Generalized Additive Models

Dr Nicholas Clark ORCID iD
School of Veterinary Science, University of Queensland, Gatton, Australia

Abstract

Time series analysis and probabilistic forecasting are standard goals in applied statistics. But most time series tools available to applied researchers are restricted to traditional forecasting models such as ARIMA, GARCH or Exponential Smoothing variants. These models cannot readily handle features that dominate many real-world datasets (particularly those in ecology), including overdispersion, clustering, missingness, discreteness and nonlinear effects. Using the flexible and powerful probabilistic programming ecosystem Stan, we can build more appropriate time series models to meet this complexity head on. In this talk I will introduce Bayesian Dynamic Generalized Additive Models (DGAMs) and illustrate their advantages for analyzing and forecasting real-world time series. I will discuss {mvgam}, an open-source R package that can fit DGAMs with nonlinear effects, hierarchical effects and dynamic processes to data from a wide variety of observation distributions. These models are especially useful for analysing multiple series, as they can estimate hierarchical smooth functions while learning complex temporal associations with latent vector autoregressive processes or dimension-reduced dynamic factor processes. Because the package uses Hamiltonian Monte Carlo inference through Stan, it is straightforward to create Stan code and all necessary data structures so that additional stochastic elements can be added to suit the user's bespoke needs. Other key features of {mvgam} are functions to critique models using rolling window forecasts and posterior predictive checks, online data augmentation via a recursive particle filter and graphical tools to visualise probabilistic uncertainties for smooth functions and predictions. I hope to show how models that describe real-world complexity, both through nonlinear covariate functions and multi-series dependence, are useful to ask targeted questions about drivers of change.

48

PanIC: am information criterion approach for generic model selection problems

Dr Hien Nguyen
University of Queensland, St Lucia, Australia

Abstract

Model selection is a ubiquitous problem that arises in the application of many statistical and machine learning methods. In the likelihood and related settings, it is typical to use the method of information criteria (IC) to choose the most parsimonious among competing models by penalizing the likelihood-based objective function. Theorems guaranteeing the consistency of IC can often be difficult to verify and are often specific and bespoke. We present a set of results that guarantee consistency for a class of IC, which we call PanIC (from the Greek root 'pan', meaning 'of everything'), with easily verifiable regularity conditions. The PanIC are applicable in any loss-based learning problem and are not exclusive to likelihood problems. We illustrate the verification of regularity conditions for model selection problems regarding finite mixture models, least absolute deviation and support vector regression, and principal component analysis, and we demonstrate the effectiveness of the PanIC for such problems via numerical simulations.

49

Explanatory Models Analysis: Why, When and How?

Przemyslaw Biecek ORCID iD
Warsaw University of Technology, Warszawa, Poland. University of Warsaw, Warsaw, Poland

Abstract

Anscombe’s quartet is a great illustration of the need to enrich models with data visualisation. Through a simple example, it shows that data visualisation is very useful, or even essential, in the statistical analysis of data.

Today, after 50 years, we are witnessing a real revolution in the number and variety of predictive models that are being developed for the most diverse applications. These models are increasingly complex with a number of coefficients that effectively make them impossible to understand. Problems with understanding a model occur with as few as 10 coefficients, but today’s DNN models often have thousands, millions or billions of coefficients.

In this talk, I will present the reasons why it is useful to visualise predictive models not only to understand them, but also to improve them. Then I will discuss situations when this is needed, most often enforced by emerging regulations such as the AI-act being discussed in the European Union. Then I will present selected tools that can be used to visualise predictive models, which are increasingly being developed under the umbrella of eXplainable Artificial Intelligence (XAI). These techniques can be used to visualise a single model, a pair of models (e.g. during champion-challenger analysis) or a larger set of models (e.g. during Rashomon set analysis).

50

Increasing sample size asymptotic for two-way crossed mixed effect model

Dr Ziyang Lyu
UNSW, Sydney, Australia

Abstract

Asymptotic results for the maximum likelihood and restricted maximum likelihood (REML) estimators of the parameters in the two-way crossed mixed effect model for clustered data when the total of row, column and cell go to infinity. A set of mild conditions is given under which the estimators are shown to be asymptotically normal with an elegantly structured covariance matrix. There are no restrictions on the rate at which the cluster size tends to infinity but it turns out to be essential to specify the regression function in an appropriate way.

51

Moving from academia to industry as a statistical consultant: The work is the same but not quite

Dr Marijke Welvaert
Australian Red Cross Lifeblood, Sydney, Australia

Abstract submission

Download file

52

It's time to build: a small area estimation methodology for time-to-event data

Mr Nelson Chua[1,2], Mr Benjamin Long[3]
[1]Australian Bureau of Statistics, Canberra, Australia. [2]Australian National University, Canberra, Australia. [3]Australian Bureau of Statistics, Sydney, Australia

Abstract

Housing availability is a current, topical social issue in Australia. Consequently, the Australian Bureau of Statistics was funded through the National Housing and Homelessness Agreement to produce small area estimates of dwelling completions. Historically, these statistics have only been produced at a broad geographical level due to the limited amount of survey data available in small areas.

We propose a model-based approach to small area estimation that incorporates time-to-event analysis, allowing for estimates to be produced alongside associated uncertainty measures. This method produces sensible estimates of dwelling completions at a small area level that are consistent with existing estimates at a broad geographical level. We envision that this methodology is applicable in contexts beyond official statistics.

53

Conditional particle filters with bridge backward sampling

Dr Santeri Karppinen[1], Professor Sumeetpal S. Singh[2], Professor Matti Vihola[1]
[1]University of Jyväskylä, Jyväskylä, Finland. [2]University of Wollongong, Wollongong, Australia

Abstract

Conditional particle filters (CPFs) with backward/ancestor sampling are powerful methods for sampling from the posterior distribution of the latent states of a dynamic model such as a hidden Markov model. However, the performance of these methods deteriorates with models involving weakly informative observations and/or slowly mixing dynamics. Both of these complications arise when sampling finely time-discretised continuous-time path integral models, but can occur with hidden Markov models too. Multinomial resampling, which is commonly employed with CPFs, resamples excessively for weakly informative observations and thereby introduces extra variance. Furthermore, slowly mixing dynamics render the backward/ancestor sampling steps ineffective, leading to degeneracy issues. We detail two conditional resampling strategies suitable for the weakly informative regime: the so-called `killing' resampling and the systematic resampling with mean partial order. To avoid the degeneracy issues, we introduce a generalisation of the CPF with backward sampling that involves auxiliary `bridging' CPF steps that are parameterised by a blocking sequence. We present practical tuning strategies for choosing an appropriate blocking. Our experiments demonstrate that the CPF with a suitable resampling and the developed `bridge backward sampling' can lead to substantial efficiency gains in the weakly informative and slow mixing regime.
Joint work with Santeri Karppinen (Jyväskylä) and Matti Vihola (Jyväskylä)

54

Impacts of COVID-19 Pandemic on Tourism Trends in Sri Lanka and Australia

Mr Udaya Banda Konarasinghe Konarasinghe Mudiyanselage ORCID iD
Rajarata University of Sri Lanka, Anuradhapura, Sri Lanka

Abstract

Abstract submission

Download file

55

Generalised Method of Moments Estimation for Exponential-Family Random Graph Models

Dr Pavel Krivitsky ORCID iD, Mr Andrew Murphy
University of New South Wales, Sydney, Australia

Abstract

Exponential-Family Random Graph Models (ERGMs), a popular framework for network data, are specified through their sufficient statistic—a vector of network features that embody the social forces affecting its formation. This sufficiency allows an ERGM to be estimated without observing the entire network. For example, given an egocentric sample, in which actors in the network report the attributes of themselves and their immediate connections, it is possible to estimate such features as density, degree distribution, and homophily, from which an MLE for the corresponding ERGM can be obtained.

The statistics that can be observed or estimated are not always the same as those believed to affect the network formation, however. For example, we may observe only censored degree ("list up to 5 friends"), which can cause us to underestimate the network density and bias the MLE. We propose a theoretical framework and efficient computational techniques for generalised method of moment estimation for ERGMs, allowing estimation with arbitrary identifiable combinations of observed and sufficient statistics. This is demonstrated this on simulated and real-world censored egocentric data.

56

Multimodal User Testing: Producing comprehensive, task-focused guidelines for chart designMultimodal Graphical Testing: A 360 degree

Dr Emily Robinson ORCID iD[1], Dr Reka Howard ORCID iD[2], Dr Susan Vanderplas ORCID iD[2]
[1]California Polytechnic University, San Luis Obispo, USA. [2]University of Nebraska - Lincoln, Lincoln, USA

Abstract

For at least the last 100 years, researchers have been testing statistical graphics and arguing about which chart designs are better. Many of these studies produce conflicting recommendations: should we use pie charts to display data about the relative proportions of a whole, or are stacked bar charts better?

Much of the time, user testing statistical graphics takes a back seat to aesthetic preferences and gut feelings, but even when we test graphics, we often only use one methodology that is focused on a specific use case.

For instance, visual inference is often used to determine whether someone can detect an effect, but it does not allow us to easily examine whether users can extrapolate from the data shown, or can draw logical conclusions from a chart.

In this presentation, I'll discuss ongoing research examining chart design choices using multiple testing methods.

Each of these methods has been designed to measure the usability of charts for a specific task: perception, estimation, and forecasting.

We'll consider the benefits and drawbacks to this type of user testing and discuss the nuances of design decisions on chart usability.

57

Engagement in and perception of flipped learning amongst specialist statistics students

Dr Eilidh Jack ORCID iD[1], Dr Elinor Jones ORCID iD[2], Dr Mitchum Bock ORCID iD[1]
[1]University of Glasgow, Glasgow, United Kingdom. [2]University College London, London, United Kingdom

Abstract

Flipped learning has been increasing in popularity as an active learning strategy in tertiary education. Although flipped learning has been found to improve students' communication skills, promote independent learners, and change learning habits, it can often be difficult to implement effectively, particularly in large classes. This talk will discuss the implementation of flipped learning in two Russell Group UK universities teaching large classes of over 200 students with an aim to understand the effectiveness of flipped learning in an evidence-based way. The courses were aimed at specialist statistics students but were delivered at different levels, namely a first-year course (in a four-year degree programme) and a second-year course (in a three-year degree programme). Each course employed a range of activities for students to complete prior to lectures in order to prepare them for the face-to-face lectures which developed a deeper understanding of each topic.

This talk will discuss student engagement with flipped learning material of differing formats (videos/reading material/quizzes) using Moodle analytics data and make comparisons between the universities. We will also discuss student perceptions of flipped learning at each university, investigate whether this is consistent with the analytics data and explore patterns between engagement and measures of achievement, such as course grades.

Initial findings show similar patterns in engagement with flipped material between the institutions, with engagement dropping over the weeks of a course and large numbers of students failing to engage with the material at all. This contrasts with the survey findings which suggest that students found the flipped tasks helpful to complete and always or almost always completed these.

We will finish with a discussion around the challenges of effective implementation of active learning strategies and how to address the differing student preferences of learning practices, particularly within large classes for specialist students.

58

Formal diagnostics for modelling spatial processes in field trials

Miss Monique Jordan, Dr. Alison Smith ORCID iD, Prof. Brian Cullis ORCID iD
University of Wollongong, Wollongong, Australia

Abstract

Each year in Australia field trials are conducted to compare the yield performance of crop varieties across different environments. The trials typically comprise rectangular arrays of plots indexed by rows and columns. Current methods of analyses for individual trials follow those of Gilmour et al. (1997) where a separable (row x column) autoregressive process of order one (often denoted AR1 x AR1) is used as a baseline model for modelling smooth local trend and a graphic of the sample variogram is used as a diagnostic for detecting non-stationarity such as in the form of extraneous variation. This diagnostic is informal and is open to interpretation leading to large disparities in the final models fitted by different practitioners. We investigate formal diagnostics for spatial modelling in field trials to provide a more vigorous framework for such analyses.

59

A Survey of Simulation Studies Evaluating Statistical Methods in Ecology and Evolutionary Biology Research

Ms Coralie Williams ORCID iD[1], Dr Malgorzata Lagisz ORCID iD[1], Mr Kyle Morrison ORCID iD[1], Mr Lorenzo Ricolfi ORCID iD[1], Dr Yefeng Yang ORCID iD[1], Professor David Warton ORCID iD[2,1], Professor Shinichi Nakagawa ORCID iD[1]

[1]Ecology and Evolution Research Centre and School of Biological Earth and Environmental Sciences, The University of New South Wales, Sydney, Australia. [2]School of Mathematics and Statistics, The University of New South Wales, Sydney, Australia

Abstract

Simulation studies hold a central role in evaluating statistical methodologies. Conducting a simulation study is an in-silico experiment, where the underlying process of generating the data is known, and the statistical performance can be assessed under diverse scenarios. Here, we present the findings from a survey of 100 research articles in ecology and evolutionary biology journals that use at least one simulation study to evaluate statistical methods. The objectives of our survey are twofold: (1) to characterise simulation studies in ecology and evolutionary biology, and (2) to assess and quantify the reporting practices of these studies. A framework called ADEMP (Aims, Data-generating mechanisms, Methods, Estimand/statistical target, Performance measures) was introduced by Morris et al. (2019, Stat Med, 38, p2074), and provides key steps to plan, code, analyse and report simulation studies. Using the ADEMP framework as a basis, we put together a list of reporting items to characterise and assess the reporting practices of simulation studies. Our results provide an overview of the current state of simulation studies in ecology and evolutionary biology, quantifying the common and rare characteristics of simulation studies but also identifying areas for improvement to ensure transparency and reproducibility. By promoting clear and consistent reporting practices, researchers can reduce the potential misapplication of simulations. With the transparent evaluation of statistical methods via simulations, researchers can better understand the statistical properties of analysis approaches and the situations in which they are best applied. This study emphasises the importance of embracing standardised approaches in evaluating statistical methods, ultimately strengthening the validity and impact of research findings from simulation studies.

60

Confidences in Hypotheses: Measures that Complement Hypothesis Testing Outcomes

Dr Graham Bornholt ORCID iD
Griffith University (Ret.), Gold Coast, Australia

Abstract submission

Download file

61

Student Opinions on Presentation of Information to Reduce Statistics Anxiety

Dr Anna Riach[1], Dr Amanda Shaker[2], Ms Ellen Marshall[3]
[1]University of Leeds, Leeds, United Kingdom. [2]La Trobe University, Melbourne, Australia.
[3]Sheffield Hallam Universty, Sheffield, United Kingdom

Abstract

Negative feelings and fear of studying statistics, often referred to as statistics anxiety, can affect grades and be a barrier to effective teaching of statistics to non-specialists at higher education (Onwuegbuzie, 2000). We have previously found that presenting students with content that raises awareness of statistics anxiety in a two-hour workshop, reduces statistics anxiety. However, we know there are students who are too anxious to attend such workshops (Marshall et al., 2022). Attempts to deliver the workshop as part of a timetabled class within a subject have the advantage of reaching all students who are anxious about statistics. The disadvantage is that students who have positive feelings towards statistics may see little benefit in a whole hour of contact time being used for such a workshop. In this multi-institutional study, lecturers presented the statistics anxiety awareness content one slide per lecture or practical throughout a subject, in order to reach those that need it while not exasperating those that do not. To gain an objective opinion of how useful the intervention was, the effect of the slides was compared with other features of the teaching. These other features, such as the use of humour, collecting own data, and attitudes of teaching staff, are included since the literature reports them to be useful in reducing statistics anxiety (Chew et al, 2014).

Chew, P.K. and Dillon, D.B., 2014. Statistics anxiety update: Refining the construct and recommendations for a new research agenda. Perspectives on Psychological Science, 9(2), pp.196-208.

Marshall, E., Riach, A., Shaker, A.J. and Haigney, D. (2022). Are students too anxious for statistics anxiety workshops? MSOR Connections, 20(2), pp. 94-105. https://doi.org/10.21100/msor.v20i2.1287

Onwuegbuzie, A.J., 2000. Statistics anxiety and the role of self-perceptions. The Journal of Educational Research, 93(5), pp.323-330. https://doi.org/10.1080/1356251032000052447

62

Causal modelling to estimate the impact of Total Fire Ban declarations

Dr Elena Tartaglia ORCID iD[1,2], Dr Matt Plucinski[3], Dr Carolyn Huston[1]
[1]CSIRO, Melbourne, Australia. [2]Department of Energy, Environment and Climate Action, Melbourne, Australia. [3]CSIRO, Canberra, Australia

Abstract

Bushfires occur naturally in the Australian landscape, and have the potential to destroy assets, property and lives. Each summer, restrictions are put in place to reduce the risk of damaging bushfires. One such restriction is the declaration of the Total Fire Ban (TFB) for a region which prohibits open air fires to be lit. Government regulators and those impacted by fire restrictions, such as farmers and power companies are very interested in understanding how effective these interventions are at reducing bushfire incidence and impacts. Our work aims to provide a quantitative understanding of the impact that TFB declarations have on bushfire incidence and impacts. Since it is not feasible or ethical to run an experiment to measure this impact, causal inference methods for observational data to estimate the effect of a TFB declaration are applicable. The data is a comprehensive fire incident dataset containing all bushfire records across Victoria from July 2003 to June 2020 which has been compiled by CSIRO in collaboration with Victorian fire authorities. In this talk, we will present our progress in this study.

63

Probability-Based Panel Performance Compared to Other Survey Modes: More Australian Evidence

Dr Benjamin Phillips[1,2], Dr Dina Neiger[1,2], Andrew Ward[1], Darren Pennay[1,2], Sam Slamowicz[1], Sandra Ropero[1]

[1]The Social Research Centre, Melbourne, Australia. [2]The Australian National University, Canberra, Australia

Abstract

The Australian Comparative Study of Survey Methods (ACSSM) fielded in December 2022 systematically trialled nine sampling frames and survey modes: web mode using a probability-based online panel, Life in Australia™ (n = 582), a video-assisted live interviewing arm recruited from Life in Australia™ (n = 600), mobile RDD CATI (high effort, n = 803), RDD SMS push-to-web (n = 596), and four nonprobability access panels (n ≥ 850 per panel). The ACSSM follows an earlier Australian comparative study (Lavrakas et al. 2022) fielded in 2015, which compared various probability and nonprobability modes, and its replication on Life in Australia™ in 2017.

The questionnaire used items for which high-quality benchmarks were available across a range of domains including health, substance use, disability, caring, psychological distress, labour force status and voting. We deliberately included items likely subject to mode effects.

We compare relative strengths and weaknesses of these methods and compare the performance of the trialed survey methods with respect to the external high-quality benchmarks. Being in part a repetition of an earlier study, we are able to examine how performance of methods changed over time. This paper contributes to the multinational evidence base on the performance of contemporary and emerging methods for general population surveys.

64

A little bit of knowledge is a wonderful thing. Helping researchers bridge the gap between statistical theory and practice

Dr Alex Shaw ORCID iD, Dr Kathrin Schemann, Mr Jim Matthews, Mr Chris Howden
Sydney Informatics Hub, The University of Sydney, Sydney, Australia

Abstract

"I'm no good at statistics", or some variation on the theme, is uttered by many new clients during initial sessions with our university statistical consulting service. Such a "crisis of confidence" can arise when researchers encounter difficulty connecting the statistical theory they have been taught with its real-world application in analysis of their research data. We set out to design researcher training that would empower researchers to conduct their own analyses and further researcher engagement with our consulting service. Over the past three years we have designed and delivered a modular statistical training system consisting of twelve 90-minute workshops. To date, these workshops collectively have been delivered more than 90 times.

The central design principle of our workshops are researcher-oriented workflows. These are essentially “how-to” guides that describe a series of practical steps we recommend they undertake in their analysis. A "general research workflow" proceeding from hypothesis generation to publication is introduced in our “Research Essentials” workshop. In this foundational workshop, researchers learn how each stage of their research project entwines with specific statistical considerations (e.g. study design, exploratory data analysis, building interpretable linear models). Subsequent workshops focus on these considerations individually and introduce workflows that address them effectively. We further maximise the accessibility of our workshops by i) keeping assumed knowledge to a minimum so that researchers of widely varying statistical experience can benefit ii) remaining “software agnostic”, so that the workflows can be applied using software of the researcher’s choice and iii) keeping them concise and introductory so that busy researchers can gain exposure to a range of relevant statistical methods applicable in diverse research areas.

By delivering accessible yet authoritative introductory workshops on various statistical topics within an integrated, modular framework we build researchers confidence to engage with statistics and use our consulting service more effectively.

65

Restricted maximum likelihood estimation in generalized linear mixed models

Dr. Luca Maestrini, Dr. Francis Hui, Prof. Alan Welsh
The Australian National University, Canberra, Australia

Abstract

Restricted maximum likelihood (REML) is generally preferred to maximum likelihood for inference on linear mixed model parameters as the maximum likelihood estimator of the variance components can be biased downwards. The same issue is present in maximum likelihood estimation for generalised linear mixed models (GLMMs); however, the concept of REML as formulated for the class of linear mixed models does not naturally extend to more general models. Several attempts to develop REML-type approaches for GLMMs in analogy to the criterion for linear mixed models have been proposed over the last four decades. We arrange, discuss and compare various implementations of REML for GLMMs considering some major classes of approaches based on approximate linearisation, integrated likelihoods, modified profile likelihoods and direct bias correction in the score function.

66

Finding cost-efficient incomplete stepped wedge designs using an iterative approach

Mr Ehsan Rezaei-Darzi ORCID iD, Associate Professor Jessica Kasza ORCID iD, Professor Andrew B Forbes ORCID iD, Dr Kelsey L Grantham ORCID iD
School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia

Abstract

Stepped wedge trials, a particular type of longitudinal cluster randomised trial in which clusters switch from the control to the intervention condition in a staggered manner, can be costly and burdensome. Recent work has investigated the iterative removal of cluster-period cells from the design that contribute relatively small amounts of information, producing a sequence of incomplete stepped wedge designs. Many of these designs retain high power to detect effects of interest. We now investigate the cost-efficiency, seeking to identify incomplete designs that retain high power while limiting trial costs. We provide a framework that incorporates the costs per cluster and per individual, and of restarting a cluster following a pause in data collection.

We consider linear mixed models for continuous outcomes, with constant cluster-period sizes, categorical period effects, and assume repeated cross-sectional sampling and a discrete-time decay within-cluster correlation structure. For each of the incomplete designs that we obtain from the iterative procedure, we then assess the total cost, variance of the treatment effect estimator and study power. We show that incomplete designs with approximately half the number of measurements as a complete design tend to have minimally reduced precision and study power for a substantially lower total cost. However, when considering designs where clusters have a pause in data collection, total costs may not necessarily decrease with progressively reduced designs.

Our methods enable trialists to examine the trade-off between the total trial cost and the power of a design. Incomplete designs requiring only half as many measurements may be preferable to complete designs - reducing both data collection burden and study costs, while having the potential to maintain high levels of power. “Staircase”-type designs, where clusters only contribute measurements immediately before and after the treatment switch, are indicated as particularly cost-efficient variants of the stepped wedge design.

Abstract submission

Download file

68

A tutorial on the analysis of some well-known social data sets using R

Dr Wessel Moolman ORCID iD
Walter Sisulu University, Mthatha, South Africa

Abstract

Each of the data sets shown below originated from some aspect of social life.

1 Danish do-it-yourself.

2 Facilities in East Jerusalem.

3 Household expenditures.

4 Chest, waist, and hips measurements.

5 Ice Cream consumption.

6 Housing satisfaction.

7 Car changing patterns.

8 Fathers and sons' occupations.

9 Measurement of protein in ground meat.

For each data set the analysis will be discussed in the form

1 Problem and data layout.

2 Description of method of analysis to be used.

3 Presentation of results and conclusions.

69

Solutions to sparsity in small area level survey statistics: Mapping the prevalence of cancer risk factors in Australia

Mr James Hogg ORCID iD
Queensland University of Technology, Brisbane, Australia

Abstract

Studies estimate that 32% of all cancers in Australia could be prevented if the population avoided exposure to known modifiable risk factors, such as smoking . However, in Australia there are no small area-level data on cancer risk factors with high resolution and complete coverage, which hinders the development of targeted prevention strategies.

We developed a Bayesian two-stage small area estimation methodology designed to estimate prevalence rates from sparse data with incomplete geographic coverage to overcome some of the limitations of the 2017-2018 National Health Survey. Our model incorporated spatial smoothing and hierarchical modelling techniques to combine survey only covariates with a vast array of small area level auxiliary information, including census, remoteness, and socioeconomic data.

The modelling generated prevalence estimates for 2223 small areas across Australia for five cancer risk factor groups, including smoking, alcohol, physical activity, diet and weight. To validate our prevalence estimates, we used internal Bayesian benchmarking and borrowed strength from previously published cancer risk estimates provided by the Social Health Atlases of Australia. The estimates presented in this talk will be included in the next major release of the internationally renowned Australian Cancer Atlas.

71

Estimation of daily smoking prevalence for disaggregated statistical areas in Australia

Dr Sumonkanti Das ORCID iD, A/Prof Bernard Baffour ORCID iD, A/Prof Alice Richardson ORCID iD
Australian National University, Canberra, Australia

Abstract

Official statistics on health outcomes for small domains are highly prized by policymakers and researchers, for measuring and monitoring progress of communities towards healthy lifestyles. Countries like Australia use their National Health Survey (NHS) to monitor adult health behaviours such as daily smoking. However, the NHS cannot be used to estimate accurate daily smoking prevalence at the disaggregated statistical area (SA) level due to lack of information. This National Health and Medical Research Council funded study aims to estimate the prevalence of daily smoking at SA levels 3 and 4.

Direct estimates of daily smoking and their smoothed standard errors for the domains at SA3 level have been used as input for developing multilevel models, which are expressed in a hierarchical Bayesian framework and fitted by Markov Chain Monte Carlo (MCMC) simulation. The developed models provide consistent estimates at the most detailed level domains by borrowing cross-sectional and spatial strengths. The detailed level domain predictions are then aggregated to obtain estimates at higher aggregation levels. Prevalences and their standard errors are plotted in a bivariate choropleth map, which permits simultaneous exploration of the spatial distribution and its accuracy.

A socio-economic index for areas (SEIFA) is investigated as a contextual variable. On its own SEIFA does not describe the socio-economic circumstances of the Indigenous population very well, so a number of contextual variables extracted from 2016 Census data have also been examined in the model development. Performance of the model-based estimates is compared with the corresponding design-based direct estimates. Significant inequalities within and between the SA3s are investigated to look for patterns that suggest policy actions that can be applied at aggregated or disaggregated levels. These findings can help health researchers and policymakers to deliver programs that enable the most vulnerable to meet their health goals in a timely way.

72

Temporal assessment of heatwaves using Cox proportional hazards

Dr Jason West ORCID iD
Bureau of Meteorology, Brisbane, Australia

Abstract

We measure differences in temporal clustering of abnormally high temperatures in urban and rural areas using a Cox proportional hazards approach. The use of Cox regression as a statistical framework to infer characteristics in the distribution function of climate variables offers a powerful method for modelling the full probability distribution rather than limiting analysis to statistical parameters only. This approach ensures that assessing the probabilities of extremes are as informative as the central tendency and naturally accounts for non-stationarities resulting from external influences such as climate change. We apply this concept to observed weather variables to assess the full effect of temperature clustering by location and time. We use observed maximum daily temperatures across 28 urban and rural areas in Australia over the period 1956–2022, along with monthly observations for the Oceanic Niño Index (ONI) sea surface temperature anomalies and the Indian Ocean Dipole (IOD) sea surface temperature index as climate covariates. We find that location and time, as well as the climate covariates ONI and IOD, are significant indicators of heatwave frequency and duration. Despite the urban heat island effect, urban locations are less prone to extremes in heatwave frequency and duration than rural locations. Furthermore, the frequency and duration of temperature clustering has increased in the post-1980 epoch relative to the pre-1980 epoch, with the effects most acutely felt in rural areas. These results have important implications for heatwave forecasting and public health management in regions prone to temperature clustering related to shifts in climate covariate forecasts.

73

Probabilistic Forecast Reconciliation For Emergency Services Demand

Professor Rob Hyndman ORCID iD[1], Associate Professor Bahman Rostami-Tabar ORCID iD[2]
[1]Monash University, Clayton, Australia. [2]Cardiff University, Cardiff, United Kingdom

Abstract

Accurate forecasts of ambulance demand are crucial inputs when planning and deploying staff and fleet. Such demand forecasts are required at national, regional and sub-regional levels, and must take account of the nature of incidents and their priorities. These forecasts are often generated independently by different teams within the organization. As a result, forecasts at different levels may be inconsistent, resulting in conflicting decisions and a lack of coherent coordination in the service. To address this issue, we exploit the hierarchical and grouped structure of the demand time series, and apply forecast reconciliation methods to generate both point and probabilistic forecasts that are coherent and use all the available data at all levels of disaggregation.

The methods are applied to daily incident data from the Welsh Ambulance Service NHS Trust, from October 2015 to July 2019, disaggregated by nature of incident, priority, managing health board, and control area.

We use an ensemble of forecasting models, and show that the resulting forecasts are better than any individual forecasting model. We validate the forecasting approach using time-series cross-validation.

74

Work-place in a safe place

Dr Jono Tuke ORCID iD
University of Adelaide, Adelaide, Australia

Abstract

Statistical consulting involves numerous skills beyond the ability to analyse data including communication, time and project management, and teamwork. While these "soft skills" are emphasised in many University Graduate Attributes, they are not easily taught or assessed in a traditional coursework setting. In 2023, we introduced an Honours and Postgraduate course at The University of Adelaide on Statistical Consulting, with the aim of addressing this gap.

Nothing prepares a statistician for statistical consulting like the actual experience of statistical consulting. To achieve this, without giving undue pressure on the students, we used role-playing. The students communicated with a false collaborator played by the lecturer.

In this talk, I will show how we set up and assessed the role-playing component of the course and discuss the results, including feedback from students. I will discuss what worked and what didn’t, and discuss some suggestions for other educators considering using this approach or teaching statistical consulting.

75

Factor analytic mixed models for multi-phase multi-environment trial data.

Mr David Hughes ORCID iD[1], Mr William Fairlie ORCID iD[2,3], Professor Marijka Batterham ORCID iD[1], Dr Alison Smith[1], Professor Brian Cullis ORCID iD[1]
[1]University of Wollongong, Wollongong, Australia. [2]The University of Adelaide, Glen Osmond, Australia. [3]Australian Grain Technologies, Roseworthy, Australia

Abstract

The Hagberg-Perten falling number (FN) test is the industry standard to measure starch degradation cause by late maturity α-amylase enzyme activity in flour. The measurement of FN is a so-called multi-phase trial involving two phases, namely a field phase and a laboratory phase.

In this talk, we extend the concepts presented in Smith et al. (2006) and Smith et al. (2015) to encompass multi-environment trial data. We present the analysis of a multi-phase multi-environment trial dataset in which the trait of interest is FN. This dataset spans 6 years from 2014-2019 and contains more than 230 environments and 124 genotypes.

Using the Design Tableau approach of Smith & Cullis (2019), an appropriate mixed model is specified which accommodates the block structure for each phase as well as allowing for additional sources of variation and correlation. The extent of the genotype by environment interaction in the Australian wheat growing regions is explored using a factor analytic linear mixed model.

Smith, A. B., Lim, P., & Cullis, B. R. (2006). The design and analysis of multi-phase plant breeding experiments. The Journal of Agricultural Science, 144(5), 393-409.

Smith, A. B., Butler, D. G., Cavanagh, C. R., & Cullis, B. R. (2015). Multi-phase variety trials using both composite and individual replicate samples: a model-based design approach. The Journal of Agricultural Science, 153(6), 1017-1029.

Smith, A. B., & Cullis, B. R. (2019). Design Tableau: an aid to specifying the Linear Mixed Model for a comparative experiment. NIASRA Working Paper.

76

Homogeneity and Sparsity Pursuit Using Robust Adaptive Fused Lasso

Dr Le Chang
Australian National University, Canberra, Australia

Abstract

The fused lasso regression is a popular method that identifies homogeneous groups and sparsity patterns in regression coefficients based on either the presumed order or a more general graph structure of the covariates. However, the traditional fused lasso may yield misleading outcomes in the presence of outliers. In this work, we propose an extension of the fused lasso, named the robust adaptive fused lasso (RAFL), which pursues homogeneity and sparsity patterns in regression coefficients while accounting for potential outliers within the data. By utilizing Huber's loss or Tukey's biweight loss, RAFL can resist to outliers in the responses or in both the responses and the covariates. Furthermore, a novel optimization algorithm that employs the alternating direction method of multipliers (ADMM), embedded with an accelerated proximal gradient (APG) algorithm, is developed to solve RAFL efficiently. Our simulation study shows that RAFL offers substantial improvements in terms of both the grouping accuracy and the prediction accuracy compared to the fused lasso, particularly when dealing with contaminated data. Additionally, a real analysis of cookie data demonstrates the effectiveness of RAFL.

77

Optimal methods for outlier detection when benchmarking in clinical registries: a simulation study

Ms Jessy Hansen ORCID iD, A/Prof Arul Earnest ORCID iD, Dr Ahmadreza Pourghaderi ORCID iD, Prof Susannah Ahern
Monash University, Melbourne, Australia

Abstract

Background: Governments are increasingly involved in public reporting of health care provider outcomes using benchmarked clinical registry data, creating potential reputational risks for poor performers. Despite this, little research has assessed the best statistical methods of benchmarking within registries, and much remains unknown about the performance of outlier detection techniques within this context. Robust evaluations of benchmarking and outlier classification under a range of clinical registry settings are required to identify the optimal methods and provide guidance to registries to ensure accurate outlier detection.

Methods: A parametric simulation study was conducted with the registry parameters of outcome prevalence, dispersion, outlier definition, proportion of outliers and risk adjustment fit varied in a partial factorial framework. Outliers flagged from a combination of four models (unadjusted, ordinary, fixed effects and random effect logistic regression) and four classification techniques (95 and 99.8% confidence intervals and control limits) were compared to simulated ‘true’ underperformers using the performance measures of sensitivity, specificity, and positive and negative predictive values.

Results: The accuracy of outlier detection increased with prevalence, with little evidence of interaction with the other parameters of interest. Sensitivity was poor at low outcome prevalence for all techniques, but there was greater consistency in model performance when using confidence intervals to classify outliers. Few outliers were flagged from methods using random effect model estimates, resulting in very poor sensitivity for all scenarios. The positive predictive value was generally quite low; the highest average across simulations of the base case registry scenario was 19.2% for fixed effects with 99.8% funnel plot control limits.

Conclusions: More research is needed to evaluate method performance under further varied registry parameters to determine optimal methods of outlier detection. This is critical to ensure that health service outcomes are accurately communicated to governments, clinical and consumer stakeholders including the public.

78

Simultaneous autoregressive models with measurement errors and missing data

Mr. Anjana Wijayawardhana ORCID iD, Dr. Thomas Suesse ORCID iD, Dr. David Gunawan ORCID iD
University of Wollongong, Wollongong, Australia

Abstract

Efficient estimation methods of simultaneous autoregressive (SAR) models are well developed. They have been extended to handle missing data and measurement errors in the response variable separately. This paper proposes two efficient likelihood-based estimation methods: direct marginal maximum likelihood (ML) and expectation maximisation algorithms for estimating SAR models with both measurement errors and missing data. The spatial error model (SEM) and the spatial autoregressive model (SAM), two popular SAR model types, are considered. The mechanism for missing data is limited to missing at random (MAR). Several computational approaches are developed to reduce the computational complexity of the proposed estimation methods. The performance of the proposed methods is investigated empirically using simulated and real datasets.

79

Video-Assisted Live Interviewing in Comparison to Other Survey Methods in Australia

Dr Dina Neiger[1,2], Dr Benjamin Phillips[1,2], Sam Slamowicz[1], Grant Lester[1], Sam Luddon[1], Emma Farrell[3], Kirsten Gerlach[3], Philip Carmo[3]
[1]The Social Research Centre, Melbourne, Australia. [2]Australian National University, Canberra, Australia. [3]Australian Bureau of Statistics, Canberra, Australia

Abstract

Social Research Centre in collaboration with the Australia Bureau of Statistics (ABS), systematically trialled video-assisted live interviewing (VALI; n = 600) in parallel with other survey modes and sampling frames in a study fielded December 2022. This included web mode using a probability-based online panel, Life in Australia™ (n = 582), mobile RDD CATI (n = 803), RDD SMS push-to-web (n = 596), and four nonprobability access panels (n ≥ 850 per panel). The questionnaire used items with available high-quality benchmarks across a range of domains including health, substance use, disability, caring, psychological distress, and labour force status. We deliberately included items likely subject to mode effects.

The VALI trial built on ABS learnings from implementing VALI during the pandemic for national household surveys that could no longer use face-to-face mode.

VALI respondents were recruited from Life in Australia™. As the sample was drawn from a panel, we provide detailed information on drivers of non-response at various stages using panel profile and participation history: the expression of interest in VALI interviewing, scheduling, and completing an interview.

We compare error in VALI to the other survey methods trialled with respect to the benchmarks. The broad range of modes included in the study help to situate VALI performance with respect to both self- and interviewer-administered modes and probability and nonprobability frames.

We also discuss lessons learned with respect to implementation and includes findings from qualitative debriefs with VALI respondents and interviewers.

This paper contributes to the emerging evidence base on VALI methods from Europe and the U.S. It provides detailed information on drivers of nonresponse to VALI interviews and errors in VALI in comparison to other methods with respect to external benchmarks.

80

Innovative Teaching and Assessment Strategies in STEM Courses: Analytical Study of their Impact on Student Engagement, Experience, and Success

Dr Sumaira Qureshi[1], Dr Yibeltal Alem[1], Dr Joanne Thandrayen[2]
[1]University of Canberra, Canbera, Australia. [2]The Australian National University, Canbera, Australia

Abstract

In this study, we explored the impact of innovative teaching and assessment strategies on student engagement, experience, and success in an Introductory Statistics and related STEM courses, in a mixed delivery setting. We undertook a comprehensive analysis of student performance and feedback from 2019 to 2022 to establish an association between these strategies and the aforementioned outcomes. This study provides insights into how the implementation of well-designed learning activities and assessments can lead to enhanced student success. Furthermore, the study highlights the potential of innovative teaching and assessment strategies to enhance learning experiences before, during, and after major disruptions such as the Covid-19 pandemic.

81

Using machine learning to determine a ‘best’ address from administrative data for New Zealand’s 2023 Census

Katie Simpson, Stephen Merry, Nathaniel Matheson-Dunning, Sam Cleland
Stats NZ, Christchurch, New Zealand

Abstract

New Zealand’s 2023 Census will be a combined model by design, with administrative data sources used to supplement census responses. One key improvement from the 2018 Census is the use of a machine learning model to determine an individual’s “best” admin address from the range of available sources. This has shown an improvement over previous rules-based approaches, overall and across various sub-populations.

This new method provides clear value as we look to increase our use of admin data across official statistics. In the context of the 2023 Census, admin addresses are critical as they determine where we can add non-responding members of the admin resident population into the census file. The scores from the address model contribute directly to our methods for including admin records into dwellings (both responding and non-responding). By deriving more reliable addresses, we can have more confidence we are including these admin records in the correct location.

We will discuss the data and methods considered, with a focus on the ways this is being implemented to improve quality of the 2023 Census.

82

Statistics Education – Building Research Collaborations

A/Prof Ayse Bilgin ORCID iD
Macquarie University, North Ryde, Australia. International Association for Statistical Education (IASE), Hague, Netherlands

Abstract

Statistics is important, as educators, we know that. It is important to add to body of statistics education to convince our colleagues to do research in statistics education, and to convince our students to do more than just one statistics unit during their university studies. We need to understand what makes students to choose statistics to study. Collaborating with international colleagues who are as keen as you, might open up opportunities to find ways of improving statistics education and get international recognition for your efforts and research. What are the possibilities? How can you become a member of national and/or international statistics education community? The most obvious questions one might ask are “what do we mean by research collaboration”, “how can we find people/researchers to collaborate with”, “which communities can be useful to engage with?”, and “where do you start?”. The possibilities for collaborations could start within your own institution (which is very obvious but mostly neglected) and extend to national and international collaborations. Building collaborations is not hard, they are easy but you need to know where to start and how to grow existing collaborations! Research collaborations enable publications, recognition, reputation building that help researchers to contribute to the body of knowledge. Additional benefit of publications produced by research collaborations is that you can use them in your teaching (research-teaching nexus), grant applications (to support your claim for expertise), promotion applications and new job applications. In this talk, I will show the possibilities for your future statistics education research based on my lived experiences. I would like to show you which doors to open, which ways to go.

83

GLM for partially pooled categorical predictors with a case study in biosecurity

Dr Christopher Baker ORCID iD[1,2,3], Professor Howard Bondell[1,2], Evelyn Mannix[1,2,3], Dr Elena Tartaglia[4], Tom Waring[1,2], Professor Andrew Robinsons[3]

[1]School of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia. [2]Melbourne Centre for Data Science, The University of Melbourne, Melbourne, Australia. [3]Centre of Excellence for Biosecurity Risk Analysis, Melbourne, Australia. [4]Data61, CSIRO, Melbourne, Australia

Abstract

National governments use border information to efficiently manage the biosecurity risk presented by travel and commerce. In the Australian border biosecurity system, data about cargo entries are collected from records of directions: that is, the records of actions taken by the biosecurity regulator. An entry is a collection of import lines where each line is a single type of item or commodity. Analysis is simple when the data are recorded in line mode: the directions are recorded individually for each line. The challenge comes when data are recorded in container mode, because the same direction is recorded against each line in the entry, meaning that we don't know which line(s) within the entry are non-compliant. We develop a statistical model to use container mode data to help inform biosecurity risk of items. We use asymptotic analysis to estimate the value of container mode data compared to line mode data, do a simulation study to verify that we can accurately estimate parameters in a large dataset, and we apply our methods to a real dataset, for which important information about the risk of non-compliance is recovered using the new model.

84

Ultra-High Dimensional Regularised Feature Screening in Regression

Mr. Ibrahim Joudah ORCID iD, Prof. Samuel Muller ORCID iD, Dr. Houying Zhu ORCID iD
Macquarie University, Sydney, Australia

Abstract submission

Download file

85

Uncertainty quantification and communication for the earth sciences.

Associate Professer Edward Cripps ORCID iD
University of Western Australia, Perth, Australia

Abstract

Advances in technology and the availability of data-acquisition devices have increasingly centralised the role of the data analytics in the earth sciences, which in turn inform data driven decision making across science, industry and government. Still, empirical based decision making continues to be made under conditions of uncertainty: data is messy; statistical model selection/estimation is complex; underlying physics that discretised numeric methods attempt to resolve are mis-specified. This recognition implies that, when the consequences of a decision is substantial, robust uncertainty quantification ought to accompany the fusion of domain knowledge and empirical evidence. This talk is based on a series of recent papers, providing an overview on: recent applications/methods developed with earth scientists and industry partners for probabilistic models of temporal and spatio-temporal meteorological, oceanographical and geophysical processes; experiences on conveying to non-statistical colleagues the meaning of uncertainty and its consequences for decision making; deployment of software for private (industry) and public use.

86

Applications of distributionally robust optimization to index tracking

Leyang Zhao, Professor Guoyin Li ORCID iD, Professor Spiridon Penev ORCID iD
UNSW Sydney, Sydney, Australia

Abstract

We review our recent results on applying ideas of distributionally robust optimization to index tracking problems.

Index tracking is a popular form of asset management. The tracking error is expressed as an expectation of a function of the difference between the returns of the index and of the portfolio. We assume that there is a model uncertainty in the distribution of the assets, hence a distributionally robust approach is appropriate. We use Bregman divergence in measuring the deviation between the actual and the nominal distribution of the components of the index and arrive at a semi-analytical form of the solution of the robust index-tracking problem.

Further, in the case of the Kullback-Leibler (KL) divergence (which is a limiting form of the Bregman divergence), we discuss how ideas of the exponential cone representability of the KL divergence, a regularized distributionally robust index tracking problem can be reformulated as a nonlinear conic programming problem and how this problem can be further simplified when the regularization is convex so that an effective numerical solution can be delivered. More generally, if the regularization can be written as a difference of convex functions, a solution can still be obtained by solving a sequence of conic linear programming problems.

The proposed approaches are applied to real financial data and to simulated data sets to demonstrate the superiority of the methods in comparison to some non-robust methods.

87

Centre of Statistics and Data Science Education - a new era of the Royal Statistics Society Centre for Statistics Education (UK)

Dr Elinor Jones
Department of Statistical Science, University College London, London, United Kingdom

Abstract

Prof Rhys Jones talk will outline and discuss international collaborations in relation to the newly established Centre for Statistics and Data Science Education (CSDSE) at the University of Surrey, UK. This exciting new centre will build on the experience of the former Royal Statistical Society Centre for Statistics Education (RSSCSE) in promoting the improvement in statistical and data science education. It will promote data literacy for all, help to develop subject-specific data skills in partnership with other bodies, and argue for specialist provision of a data science curriculum targeted at the 16-19 age range.

88

Variable selection in frailty mixture cure models using adaptive LASSO

Dr Richard Tawiah ORCID iD[1], Prof Howard Bondell[1], Prof Shu-Kay Angus Ng[2]
[1]University of Melbourne, Melbourne, Australia. [2]Griffith University, Brisbane, Australia

Abstract

Abstract submission

Download file

89

The saturated pairwise interaction Gibbs point process as a joint species distribution model

Dr. Ian Flint[1], Professor Nick Golding[2], Professor Peter Vesk[1], Assoc Prof Yan Wang[3], Professor Aihua Xia[1]

[1]the University of Melbourne, Melbourne, Australia. [2]Curtin University, Perth, Australia. [3]RMIT University, Melbourne, Australia

Abstract

In an effort to effectively model observed patterns in the spatial configuration of individuals of multiple species in nature, we introduce the saturated pairwise interaction Gibbs point process. Its main strength lies in its ability to model both attraction and repulsion within and between species, over different scales. As such, it is particularly well-suited to the study of associations in complex ecosystems. Based on the existing literature, we provide an easy to implement fitting procedure as well as a technique to make inference for the model parameters. We also prove that under certain hypotheses the point process is locally stable, which allows us to use the well-known `coupling from the past' algorithm to draw samples from the model.Different numerical experiments show the robustness of the model. We study three different ecological datasets, demonstrating in each one that our model helps disentangle competing ecological effects on species' distribution.

90

Lessons from post-publication statistical reviews of linear regression in health research

Ms Lee Jones ORCID iD[1,2], Prof Adrian Barnett ORCID iD[1], Assoc Prof Dimitrios Vagenas[1]
[1]QUT, Brisbane, Australia. [2]QIMR Bergofer, Brisbane, Australia

Abstract

This study explores current statistical practice and identifies common statistical misconceptions and errors researchers make when using linear regression. Statistical practices were assessed in the health and biomedical field by randomly sampling 100 published papers from PLOS ONE in 2019. Forty statisticians were recruited to review the papers, with papers randomly allocated to statisticians ensuring that two independent statisticians rated each paper.

The results show that the average researcher tends to over-rely on p-values and significance rather than the contextual importance and robustness of conclusions drawn, with an estimated 69% of authors of papers not discussing the scientific importance of parameter estimates and only 23% directly interpreting the size of regression coefficients. Only 37% of authors reported checking any of the linear regression assumptions; the most frequently assessed assumption was normality, with most authors incorrectly checking the outcome ("Y") variable rather than model residuals.

Recommendations for improving this interpretation gap include teaching statistics holistically, where most statistics can be seen in a regression framework rather than a series of unconnected and rote-learned tests. Practical recommendations from this study include greater reporting transparency, with journals providing researchers with template papers reporting common statistical methods. To help them assess statistical methods, reviewers should receive basic statistical training and potentially access automated tools which guide statistical feedback.

Abstract submission

Download file

91

Inference of Volatility Model and Risk Premia from Multiple Information Sources - An Approximate Bayesian Cut Posterior Approach

Ole Maneesoonthorn ORCID iD, Gael Martin, David Frazier, Ruben Loaiza-Maya
Monash University, Melbourne, Australia

Abstract

This paper utilizes three sources of information, daily returns, high frequency data and market option prices, to conduct inference about stochastic volatility models. The inferential method of choice is the approximate Bayesian computation (ABC) method, which allows us to construct posterior distributions of the model unknowns from data summaries without assuming a large dimensional measurement model from the three information sources. We employ ABC cut posteriors to dissect the information sources in posterior inference and show that it significantly reduces the computational burden compared to conventional posterior sampling. The benefit of utilizing multiple information sources in inference is explored in the context of predictive performance of financial returns and option prices.

92

Predictions for Anomalous Objects with Property Pricing Applications.

A/Prof Andriy Olenko ORCID iD
La Trobe University, Melbourne, Australia

Abstract

This talk presents a prediction model for out-of-sample correction under the assumption of "scalable" features. First, it introduces and discusses the motivation and technical details of the model. Then, the methodology is illustrated with the property market data. The practical analysis is divided into several stages, beginning with the application of a genetic algorithm with several fitness functions to select the most important variables and build the general model using a training set of properties. In the next stage of the analysis, several approaches to determine anomalous properties are discussed, including density-based clustering, isolation forest, and leverage points. Finally, the proposed out-of-sample correction methodology is applied to the models and anomalous properties chosen at the previous stages. Several alternative methods are analysed and compared. The approach provides a framework for out-of-sample correction prediction, which can be applied to other models trained on data with "scalable" features.

93

Cokrig-and-Regress with Bootstrap for Spatially Misaligned Data

Mr. Zhi Yang Tho, Dr. Francis Hui, Prof. Alan Welsh, Dr. Tao Zou
The Australian National University, ACT, Australia

Abstract

Understanding the relationship between the response and covariates in the analysis of spatial data is crucial in various fields. This article considers the regression problem for spatially misaligned data where the response and covariates are observed at different spatial locations, by studying a cokrig-and-regress (CNR) method to estimate the regression coefficients that are associated with potentially highly non-linear functions of the covariates. The CNR estimator is obtained by replacing the unobserved covariates with their empirical best linear unbiased prediction based on the observed but misaligned covariates under the multivariate normal assumption, which is also known as cokriging. A generalized kronecker product covariance matrix is used to model the cross correlations between the covariates and a flexible parametric bootstrap approach that accounts for such cross correlations and the additional uncertainty from the prediction of unobserved covariates is proposed to give alternative point estimates and to perform inferences on the regression coefficients. Simulation studies demonstrate that the bootstrap approach greatly outperforms the naive variance estimator that ignores the variability of the covariates' prediction in terms of variance estimation of the CNR estimator and empirical coverages of confidence intervals. The proposed CNR method with the bootstrap approach is applied to the air pollution and meteorological data in China to study the complex associations between $PM_{2.5}$ concentration and meteorological covariates under the spatial misalignment problem, which shows that the confidence intervals based on the parametric bootstrap method are always wider than the confidence intervals constructed using the naive variance estimator.

94

Warped 3-D spatial processes for inferring seabed conditions for geotechnical engineering applications

Dr Michael Bertolacci[1], A. Prof. Andrew Zammit Mangion[1], Mr Juan Valderrama Giraldo[2], Dr Michael O'Neill[2], Prof. Fraser Bransby[2], Prof. Phillip Watson[2]

[1]University of Wollongong, Wollongong, Australia. [2]University of Western Australia, Perth, Australia

Abstract

The design of offshore structures must account for the geotechnical properties of subsea soils. However, as data collection in an offshore setting is difficult and costly, geotechnical measurements are often taken at only a small number of locations at a site. This leads to a need for methods to predict the properties of the subsea soil at unsampled locations using spatially-sparse data. The methods currently used in offshore engineering for doing so do not capture important features of the data, do not always quantify uncertainty in a coherent way, and are not rigorously tested using real data. To address these issues, we propose a hierarchical spatial modelling framework for predicting the 2-D and 3-D geotechnical properties of subsea soil from sparse data. Our hierarchy describes several components of variability inherent in subsea soil, including a region-wide vertical mean profile modelled using B-splines, and a deviation process modelled using a nonstationary 3-D spatial Gaussian process. Process nonstationarity is captured by warping in three dimensions and by allowing the process variance to change with depth. We consider a range of models of increasing complexity, and estimate model parameters using Bayesian methods. We apply our method to measurements of cone tip pressure from piezocone cone penetrometer tests (PCPTs) taken at six different sites off the west coast of Australia. Our results show that capturing nonstationarity in the vertical profile is critical for prediction accuracy and valid uncertainty quantification, while horizontal dependencies in the deviation process are important only at some sites. Our method performs well on all considered sites and metrics, demonstrating its value to offshore geotechnical engineering.

95

Model averaged tail area confidence intervals in nested linear regression models

Assoc. Prof. Paul Kabaila ORCID iD, Ms Ayesha Perera
La Trobe University, Melbourne, Australia

Abstract

The performance, in terms of coverage probability and expected length, of the model averaged tail area confidence interval, proposed by Turek and Fletcher in 2012, depends greatly on the data-based model weights used in its construction. We generalise the computationally convenient exact formulas for the coverage probability and expected length of this confidence interval, published by Kabaila, Welsh and Abeysekera in 2016 for two nested linear regression models, to the case of *two or more* nested linear regression models. This enables the numerical assessment of the performance of the model averaged tail area confidence interval for any given data-based model weights in the context of three (or possibly more, depending on computational constraints) nested linear regression models. We are then able to provide a very informative further exploration of the influence of these model weights on the performance of this confidence interval.

96

Embedding communication into the statistics curriculum

Colette Mair, Kieran Brown, Emma Beekman, Esther Olowe, Callum Macaulay, Jingxuan Fu, Alice Parodi
University of Glasgow, Glasgow, United Kingdom

Abstract

Key qualifications often listed for data analytics/statistician roles are strong communication, interpersonal skills, teamwork, data management, interpretation, and rationalization of data in addition to industries asking applicants for awareness and knowledge of the field.

We bring together results from three surveys aimed at final-year undergraduate statistics students to understand their views on teaching excellence, graduate employability, and course evaluations to understand how students view, and how they can better participate in, the statistics curriculum to achieve post-graduation success.

Results indicate that students valued learning material relevant to exams. Research seminars, peer feedback and team projects were viewed as not important. Most students did not report class representatives as an important aspect of their student experience and deemed further reading not to be an important aspect of an excellent lecture and did not indicate a preference for interactive learning over didactic lectures. When asked what useful skills they developed through their programme all students identified programming skills and statistical analyses. When asked what characteristics they believe would help them in employment, students responded with a combination of communication skills, writing skills, insights into the industry, willingness to learn new things, organisation, and computing skills. Feedback is an essential component of learning development. Students agreed that course evaluations are necessary and useful in building relationships with lecturers and while staff implement the feedback they receive, students currently do not see it, and their learning may not benefit from being part of this process.

These results highlight a need to better embed communication throughout the statistics curriculum. We propose the use of course evaluations as a feedback dialogue tool to encourage and enhance communication and relationships between staff and students and develop self-regulated learning, in addition to a peer-assisted learning framework as a means for students to collaboratively develop professional skills.

97

Mathematics and Gender: Analysis of the ChooseMaths student surveys

Professor Inge Koch
RMIT, Melbourne, Australia

Abstract

To get some insights into nature versus nurture, we consider surveys of about 8000 Year 5-9 school students that teachers of AMSI's ChooseMaths program collected from 2016 to 2019 in the 120 schools across Australia that had participated in the ChooseMaths program. The surveys' aims consisted in examining the attitudes of students to mathematics across the genders and school years, and in evaluating the effectiveness of well-designed `treatment' regarding students', and in particular girls', change of attitude to mathematics and the building of confidence. The findings of the analysis are encouraging and show that attitudes to mathematics and confidence in one's mathematical ability can improve with the right `treatment'. Although girls start at a lower confidence level in their Year 5, their gains are larger than those of boys. The analyses further highlight that the changes in confidence and attitude due to treatment are much larger than the differences between the genders.

98

Spatial Dissimilarity Models with Application to Antarctic Species Turnover Analysis

Research Fellow Xiaotian Zheng[1,2], Associate Professor Andrew Zammit Mangion[1,2], Distinguished Professor Noel Cressie[1,2]

[1]National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong, Australia. [2]Securing Antarctica's Environmental Future, University of Wollongong, Wollongong, Australia

Abstract

Monitoring change in species composition, often referred to as species turnover, is an informative way of measuring biodiversity. Generalised dissimilarity modelling is commonly used in ecology to understand species turnover through site-pairwise dissimilarities in species composition, which depends on environmental predictors in monotonic relations under a generalised linear model. However, this approach using predictors to explain spatial variation is unable to accommodate complex spatial dependence, especially when the ecological process underlying dissimilarity is spatially dependent. To this end, we develop a dissimilarity-analysis framework based on spatial generalised linear mixed models. The framework extends the classical model by taking into account more structured dependence and thus introduces spatial dependence among dissimilarities. Monotonic relations are modelled using nonparametric regressions based on shape-constrained Bernstein polynomials. Our modelling approach offers avenues to incorporate expert knowledge into the monotonic relations. Finally, we integrate spatial statistical downscaling into the model to allow for predictors that are only available at coarse resolutions. Our approach, which is based on conditional-probability modelling, naturally accounts for the uncertainty that may arise from the downscaling procedure. This compares favourably with the traditional protocol in which downscaling is used as a pre-processing step when preparing predictor data. We investigate model properties analytically and through simulation studies, and we illustrate our methodology with an analysis of species turnover in Antarctica.

99

Engaging with online distance learners to promote retention and success

Colette Mair, Craig Alexander, David McArthur, Eilidh Jack
University of Glasgow, Glasgow, United Kingdom

Abstract

The University of Glasgow offers fully online master's programmes in Data Analytics. The programme attracts a global audience of students who are in full-time relevant employment from a range of backgrounds, educational experiences, and levels of knowledge in programming, mathematics, and statistics. Online students tend to be older and more likely to have caring responsibilities when compared to traditional on-campus undergraduate and postgraduate students.

To better understand this cohort of students, we used the Online Learning Readiness Scale to determine students' learning characteristics and asked students what they liked and disliked about online learning and what aspects of online learning contributed to their performance. Most strikingly, online distance learners indicated a strong motivation to learn, likely through their employment, and confidence in self-directed learning. The freedom and flexibility of online learning made their learning possible; our assessment structure enabled them to organise their schedules, and the technology-enhanced learning material was their main connection to the programme. On the other hand, students identified hesitancy in seeking help or posting to online forums and there was a general feeling of lack of interaction with peers and staff. Academically, frequent live sessions, offering one-to-one support and lecture-style videos where students can see the course lecturer can help to build relationships between learners and staff. Prompting students to introduce themselves at the beginning of the programme via a student forum, encouraging students to respond to each other’s posts, and facilitating peer-to-peer study sessions can encourage relationships between learners. Personally, pastoral care is an essential component and we have found that early and frequent contact between staff, students and advisers is imperative to ensure student retention and success.

To effectively engage with online distance learners, we not only need to understand their motivations to learn but provide mechanisms to support students academically and personally.

100

Accessibility for statistics teaching and learning materials: opportunities and challenges

David McArthur, Craig Alexander ORCID iD, Colette Mair
University of Glasgow, Glasgow, United Kingdom

Abstract

Within the mathematical sciences, many course notes are compiled using LaTex or Markdown, which often produce output as a PDF. The PDF format is popular because of its universal properties, with most software, devices and operating systems supporting this format. However, PDFs are designed to be a digital version or paper-based notes, and as such have fixed colours, line-width, line-spacing and limited navigation. PDF style documents also do not permit the addition of interactive elements, such as short quizzes, to enhance student learning.

In this presentation, we will discuss some of the challenges such course notes present in terms of accessibility and provide approaches to make these materials more accessible to learners. Our approaches are based on using an HTML based format for notes. This approach has involved taking into consideration features which are thoughtful, and concentrate on simplicity first and foremost, with multiple options for users.

The HTML notes are self-contained, making them just as portable and universal as PDF files, as they can be opened in any browser. The design is simple and distraction-free and includes an accessibility pane providing a variety of formatting options. The HTML file is responsive and can adapt to a variety of screen sizes allowing learners the flexibility to read notes on their phone or tablet. Within our HTML file, we have proposed our own pre-compiled, accessible version of LaTeX-as-SVG using MathJax, allowing for the use of screen readers for mathematical equations.

We will also discuss considerations for data visualisations from an accessible perspective, considering the addition of alternative text added to plots for screen readers and considerations for colour blind learners. We will conclude with a discussion on ongoing work on how to incorporate features for visually impaired learners, looking at translating LaTeX to Braille and printing plots using a Braille embosser.

101

Randomization-Based Intro Stats course using three Rs: New Tricks, Old Dogs

Dr. Shu-Min Liao
Amherst College, MA, USA

Abstract

In response to the global concerns on undergraduate students' well-being and mental health, a new statistics course named "Happy Intro Stats" (HIS) was created and offered at our U.S.-based liberal art college in Fall 2022. This is an interactive and fully inclusive introductory statistics course designed to address the importance of self-care on mental health and help students understand inequities in mental health status and access via statistical investigations. To ensure that this course is "fully inclusive" and accessible to students with different academic background and preparation, we impose no prerequisites and expect no prior coding experience when designing the course. Furthermore, we center the core of this introductory course on the ideas and logic of statistical inference, rather than probability theories and normality-based methods, and extend what George Cobb (2007) called "three Rs of inference" – Randomize, Repeat, and Reject – for permutation tests to bootstrap confidence intervals as "Resample, Repeat, and Range." We then apply the proposed three-R structures to create different modules for various inference procedures using the mosaic package in R. Like Cobb (2007), we believe that randomization-based methods are not only easier to implement these days, but also better pedagogical tools to help students learn the core ideas and logic of statistical inference – the main learning goal of most introductory statistics courses. A comparative study is conducted between this HIS course and our existing intro stats course which requires calculus as a prerequisite and teaches normality-based inference methods using traditional methods. Both courses under comparison were taught ty the same instructor in the same semester.

Reference: Cobb, G. W. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum? Technology Innovations in Statistics Education, 1(1). http://dx.doi.org/10.5070/T511000028

102

P-CUBE—A MULTI-STEP PRECISION PATHWAY FOR PREDICTING ALLOGRAFT SURVIVAL IN HETEROGENEOUS COHORTS OF KIDNEY TRANSPLANT RECIPIENTS

Dr. Yunwei Zhang[1,2], Professor Jean Yang[1], Professor Samuel Muller[2], Professor Germaine Wong[1]
[1]The University of Sydney, Sydney, Australia. [2]Macquarie University, Sydney, Australia

Abstract

Cohort heterogeneity exists in many clinical and omics datasets. Discovering subgroup specific risk factors is crucial for personalised medicine and for providing accurate survival prediction. While numerous unsupervised clustering methods have been applied to identify potential subgroups in the population data, the detection of the associated risk factors and the corresponding survival modelling are conducted in separate steps. Therefore, subgroup specific prediction is not guaranteed to be superior than fitting the full model.

To address this gap, in this talk we present Personalized Prediction Pathway (P-cube), which is a two-step workflow that uses survival model predictability to guide subgroup identification using both unsupervised and supervised learning strategies. Two key innovative ideas in P-cube involve (1) using model predictability to guide subgroup identification; (2) using recipient features, the first layer of the data, for subgroup identification while using all three layers of data, recipients, donors and compatibilities variables, in the prediction.

We will show how applying P-cube to the Australian and New Zealand dialysis and transplant registry data (ANZDATA, 2008 – 2017) can identify specific risk factors within this heterogenous population and thus offers a personalized approach to survival prediction. We will conclude that P-cube has the capacity and flexibility to be applied to various data modalities such as proteomics and metabolomics.

103

Spatial modeling and future projection of extreme precipitation extents

Dr Peng Zhong[1], Dr Manuela Brunner[2,3], Dr Raphaël Huser[4], Dr Thomas Opitz[5]
[1]University of New South Wales, Sydeny, Australia. [2]ETH Zurich, Zurich, Switzerland. [3]Institute for Snow and Avalanche Research SLF, Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Davos, Switzerland. [4]King Abdullah University of Science and Technology, Jeddah, Saudi Arabia. [5]INRAE, Avignon, France

Abstract

Extreme precipitation events with large spatial extents may have more severe impacts than localized events as they can lead to widespread flooding. It is debated how climate change may affect the spatial extent of precipitation extremes, whose investigation often directly relies on simulations from climate models. Here, we use a different strategy to investigate how future changes in spatial extents of precipitation extremes differ across climate zones and seasons in two river basins (Danube and Mississippi). We rely on observed precipitation extremes while exploiting a physics-based mean temperature covariate, which enables us to project future precipitation extents. We include the covariate into newly developed time-varying r-Pareto processes using a suitably chosen spatial aggregation functional r. This model captures temporal non-stationarity in the spatial dependence structure of precipitation extremes by linking it to the temperature covariate, which we derive from observations for model calibration and from bias-corrected climate simulations (CMIP6) for projections. For both river basins, our results show a negative correlation between the spatial extent and the temperature covariate for most of the rain season and an increasing trend in the margins, indicating a decrease in spatial precipitation extent in a warming climate during rain seasons as precipitation intensity increases locally.

104

Estimation of Zero-Inflated Generalized Poisson Model with Missing Data

Mr George Ringo Manapat[1], Dr Iris Ivy Gauran ORCID iD[2], Dr Joseph Ryan Lansangan[1]
[1]University of the Philippines Diliman, Quezon City, Philippines. [2]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Abstract

Count data are frequently encountered in various fields of research, including health studies. However, the commonly used Poisson model requires equidispersion, a typically violated assumption among many datasets because of heterogeneity due to zero inflation and overdispersion. Moreover, missing data can also be a critical issue in practical applications. Driven by these research gaps, we propose a novel estimation procedure for analyzing overdispersed and zero-inflated count data with missing values within a broad class of zero-inflated count models. Some theoretical properties of this class are discussed, and we demonstrate through numerical studies that our proposed procedure is comparable to existing practices. We illustrate the applicability of the estimation procedure using health data.

106
Understanding confidence intervals in context

Associate Professor Sue Finch, Professor Ian Gordon
University of Melbourne, Parkville, Australia

Abstract

A small body of research has investigated student and researcher understanding of confidence intervals. This has mostly focused on the interpretation of the confidence coefficient, the meaning of the width of the confidence interval, and the lexical ambiguity of the term “confidence”.

Most of this research has used fixed response option questions or simple hypothetical scenarios to assess students’ or researchers’ knowledge.

We present an analysis of over 150 graduate students’ written explanations based on an assessment task. Students were shown a graphical presentation of confidence intervals in a meaningful context, where they consider analysis of the relevant data and respond to a common misunderstanding of the interpretation of results presented.

A correct understanding of the meaning of a confidence interval or a simple examination of the data provided to the students would allow correction of the misunderstanding. However, many students failed to perceive the misunderstanding, or to provide a coherent response.

In contrast, on fixed response option questions, most students chose a correct interpretation of a confidence interval. We discuss the implications for teaching the applied interpretation of this fundamental statistical concept.

107

Crossmaps: A principled approach to ex-post data harmonisation and dataset integration

Ms Cynthia Huang ORCID iD
Monash University, Melbourne, Australia

Abstract

Researchers and practitioners spend considerable time and effort designing and producing merged datasets. Unfortunately, dataset designs and merged datasets can be very difficult to audit, modify or reuse. Even with knowledge of the programming language, decoding long custom data wrangling scripts can be prohibitively complex and time consuming. Ex-post harmonisation scripts are especially difficult because careful recoding, aggregation or redistribution of values is required to merge related datasets such as labour statistics collected under evolving occupation nomenclature.

Moreover, coordination and verification issues arise when data are produced and used by different parties such as in interdisciplinary teams. The design of a harmonised dataset might require domain or statistical expertise from team members without sufficient time or programming skills to implement the data transformations. Data wrangling may be delegated to other team members or research assistants, and may even require different software to the data analysis. Downstream data users often rely on ad hoc validation to check a merged dataset matches the intended design. Without adequate checks, errors can go undetected unless they produce obviously suspect results downstream.

In this talk, I present a new approach to describing, implementing and validating ex-post harmonisation and data transformations. I will also demonstrate how to implement this approach, and verify existing data integration workflows, using the *xmap* package in the R language. The proposed approach separates categorical recoding and nomenclature transformation operations from procedural code by encoding the transformations as machine-readable ‘crossmap’ objects. The crossmap structure extends standard crosswalk or lookup table approaches using graph and matrix representations of data transformations. It encodes relations between nomenclatures as directed, bipartite weighted graphs. Graph properties of a crossmap can be checked to ensure consistency with the harmonisation design, circumventing the need for ad hoc validation of the transformed data.

108

Improved National Macadamia Yield Forecasting Using an Ensemble Approach

Dr Xingjuan Li, Dr David Mayer
Queensland Department of Agriculture and Fisheries, Brisbane, Australia

Abstract

Macadamia yield forecasting is vital for procurement, price negotiation and logistics. However, obtaining accurate forecasts has proven to be difficult. Yield forecasting is challenging because of limited data available to develop the models. Besides, there are many uncertain factors (e.g. climate impact and market price) and interacting factors (e.g. local weather and soil characteristics) affecting the accuracy of modelling. Previously, macadamia forecasts were formulated using statistical models, such as general linear models. However, these approaches rely on feature selections, which is subjective. Machine learning models are end-to-end learning approaches that can automatically learn from the data. However, these approaches tend to have bias/variance trade-off issues. In this work, we used data collected from 2000 to 2022, including price, green normalized difference vegetation index, vapour-pressure-deficit, temperature, rain, evaporation rate, and soil-water-index. We showed that training multiple independent models and averaging their results on a test dataset improve forecasting accuracy and reduce generalization errors. Specifically, we predicted national macadamia production within 4% error in the past three years using an ensemble approach. These error rates meet the nominated project criteria of ‘within 10% of actual production’. Forecasting models include general linear models, least absolute shrinkage and selection operator, principle components regression, support vector regression, Bayes linear regression, extreme gradient boosting, and long short-term memory. In 2020, the forecast was 49,750 tonnes and the crop came in at 50,300 tonnes, a 1.1% error. In 2021, the forecast was 53,000 tonnes, with an actual tonnage of 55,200 (4.0% error). In 2022, the forecast was 58,900 tonnes, compared with actual production of 56,800 tonnes (a 3.7% error). For the 2022 crop, individual model performance on a regional level showed that statistical models performed better than the machine-learning models. The results for this year will be known in December when 2023 crops are finalised.

Abstract submission

Download file

109

Work integrated learning for the common good

A/Prof Ayse Bilgin ORCID iD, Dr Frank Valckenborgh ORCID iD
Macquarie University, North Ryde, Australia

Abstract

The literature has many articles about work integrated learning (WIL) but not many of them relate to statistics and/or mathematics graduates. Some papers discuss helping students to improve their employability skills addressing soft skills like team work, communication, and leadership, while others talk about the importance of working with authentic real data on real-world problems. The importance of a capstone unit for a statistics major has been talked about for more than two decades, while the importance of a capstone unit for mathematics was only addressed in the literature more recently. In either case, finding problems from industry that can realistically be solved within a semester by third-year students is not so easy. The process has many steps. The most important hurdle after finding suitable partners and the problems they propose relates to the supervision of the students during the WIL activities. In a professional degree, such as nursing or teaching, WIL has been the norm since these professions first appeared, but for disciplines like statistics and mathematics that has hardly been the case. Many people know what a nurse is doing and there are plenty of nurses who can help/supervise student nurses during their WIL activities. How many people could confidently say what a statistician and/or mathematician is doing? The best answer for this question is that a statistician and/or mathematician works for the common good unless they choose to make money (which is also possible). In this paper we will present examples of student projects and their positive impact on the common good.

110

Designing and analysing partially clustered trials with continuous outcomes

Kylie Lange ORCID iD[1,2], Dr Thomas Sullivan[2,1], Assoc Prof Jessica Kasza ORCID iD[3], Dr Lisa Yelland[2,1]

[1]The University of Adelaide, Adelaide, Australia. [2]South Australian Health and Medical Research Institute, Adelaide, Australia. [3]Monash University, Melbourne, Australia

Abstract

Methods for designing and analysing cluster randomised trials are well established. However, many clinical trials involve partially clustered data, where only some observations belong to a cluster (e.g. neonatal trials including singletons and twins). Recently, we defined four types of partially clustered trial designs characterised by whether clustering occurs pre- or post-randomisation and the method of randomisation for clustered observations. However, the performance of analysis methods for such trials have received limited attention and sample size formulas are only available for generalised estimating equations (GEEs). We aimed to assess (1) the performance of mixed effects models versus GEEs for analysis of partially clustered trials, and (2) whether existing sample size formulas based on GEEs provide appropriate power for analysis via mixed models.

A simulation study was conducted to evaluate the performance of mixed models versus GEEs for estimating the effects of treatment on a continuous outcome. We considered a maximum cluster size of 2 and simulated datasets with the sample size required to achieve 80% power according to GEE-based formulae for cluster, individual and balanced randomisation. Simulation parameters were chosen to reflect the range of scenarios observed in practice. Datasets were analysed using mixed effects models and GEEs with an independence or exchangeable working correlation structure.

GEEs generally performed well with some exceptions when the ICC was high. Performance of the mixed model was typically comparable to GEEs, though non-convergence (maximum 11%) and under-coverage (minimum 75%) occurred in more extreme settings (e.g. few pairs, high ICCs). Calculating sample size using the exchangeable correlation GEE resulted in approximately 80-85% power for the mixed model across all scenarios.

In conclusion, in many partially clustered trial settings both GEEs and mixed effects models perform well. Designing these trials using existing GEE based sample size formulas may be appropriate for analysis via mixed models.

111

A Spatial Examination Of The Genetic Drivers Of Retinal Thickness

Victoria Jackson ORCID iD[1,2], Yue Wu[3], Roberto Bonelli[1,4], Brendan Ansell[1,2], Samaneh Farashi[1,2], Liam Scott[1,2], Yuka Kihari[3], Julia Owen[3], Aaron Lee[3], Melanie Bahlo[1,2]
[1]Population Health and Immunity Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Australia. [2]Department of Medical Biology, University of Melbourne, Parkville, Australia. [3]Department of Ophthalmology, School of Medicine, University of Washington, Seattle, USA. [4]The Lowy Medical Research Institute, La Jolla, USA

Abstract

The retina is the light-sensitive area of neural tissue at the back of the eye that enables vision. Comprised of several layers of highly specialised cells, the retina receives, and transmits light signals to the brain via the optic nerve. Optical coherence tomography (OCT), a widely used method, is a non-invasive imaging technique, from which morphological measures such as retinal thickness, may be derived. Changes in retinal thickness are a key feature of several eye diseases, and more recently have been suggested as a biomarker for neurodegenerative disease. The regions in which retinal thinning, or thickening occurs can be highly disease specific.

Over 67,000 individuals from the UK Biobank underwent OCT imaging, as part of an ophthalmic assessment. Using these OCT data and AI-enabled image segmentation methods, we generated high resolution maps of retinal thickness, giving measurements at 29,041 locations (pixels) across the retina. We combined these retinal thickness measures, with genome-wide genotype data, to uncover the genetic drivers of retinal thickness using two approaches: 1) 2D-functional principal component (fPC) analysis was used to identify gross spatial patterns representing the major contributors to the variation of retinal thickness. We then undertook genome-wide association analyses (GWAS) using the top fPC scores as phenotypes. 2) by undertaking GWAS for all 29,041 pixels, allowing the examination of spatial patterns of associated genetic variants, at ultra-high resolution. We also explored associations between retinal thickness, metabolites and a range of diseases, and used Mendelian Randomisation methods to infer causal relationships.

112

What learning activities and resources do the students prefer for their learning in a large first-year enrolment statistics service unit?

Dr Karol Binkowski ORCID iD, Dr Nicholas Tse ORCID iD
Macquarie University, Sydney, Australia

Abstract

The first-year statistics service unit is offered to students studying towards a Bachelor of Science degree. Due to its large student cohort, introductory statistical unit assessments are designed as online mastery quizzes within a Learning Management System (a version of Moodle). Students are provided with pre-recorded lectures, Q&A lectures, and tutorials that could be taken online or on campus. Asynchronous discussion forums support opportunities for further questions and answers. Weekly practice quizzes are designed to help students to master weekly learning. They are created using R with a random seed so that students get a different version of the questions every time they attempt the quiz. The emphasis is on individualised learning and students' progress through the material at their own pace over five modules. In this paper, we explore students' engagement across discussion forums, online non-graded weekly practice quizzes, and the frequency of their access to the pre-recorded lectures. The relationship between students' engagement in learning resources and activities and their learning outcomes will be presented and measured by their final marks for the unit. The results provide information for better unit design based on students' engagement.

113

Embedded WIL experiences and career development learning in tertiary statistics subjects: A Singapore and Australian perspective

Dr Jing Feng Lau[1], A/P Michelle Eady ORCID iD[2], Dr Matthew Moores ORCID iD[2]
[1]Singapore University of Social Sciences, Singapore, Singapore. [2]University of Wollongong, Sydney, Australia

Abstract

This project is going to work with students in mathematical statistics courses in Singapore and Australia. Work Integrated Learning (WIL) are programs that link university students to a workplace related to their field of study. Career Development Learning (CDL) is learning which helps students to acquire knowledge, concepts, skills and attitudes that will equip them to manage their careers. We will implement an embedded work-integrated learning (WIL) and career development learning (CDL) activity in efforts to increase the students' confidence for future career choices and to develop improved employability readiness. This is a multicultural, global project that focuses on a group of cross-cultural students of similar study disciplines. The participants will be of varying ages, genders and experience who all have an interest in discussing their future careers in mathematical sciences post-graduation. Research will be disseminated through a report, journal articles and conference presentations.

114

A Bayesian approach to Mendelian randomization using summary statistics with correlated pleiotropy

Dr Andrew Grant[1], Dr Stephen Burgess[2]
[1]University of Sydney, Sydney, Australia. [2]University of Cambridge, Cambridge, United Kingdom

Abstract

Instrumental variables (IVs) can be used to make causal inference using observational data where there is unmeasured confounding. For a variable to be a valid instrument, it must be associated with the exposure and not associated with the outcome other than via the exposure. It can then be treated as an unconfounded proxy for the exposure, and an association with the outcome provides evidence of a causal effect. Mendelian randomization is a popular technique in health research which uses naturally occurring genetic variants as IVs. There is good justification for using genetic variants as instruments: they are independent of many social and environmental factors which typically confound the relationship between putative causal exposures and disease outcomes. However, their validity as instrumental variables can still be questioned because of pleiotropy, which is where a genetic variant associates with multiple traits. A number of methods for performing Mendelian randomization which are robust to instrument invalidity have been proposed in recent years. These methods typically swap one or more of the IV assumptions for other, often very strong and almost always untestable, assumptions, and therefore many commonly used approaches remain restrictive. In this talk, I will present a Bayesian framework for robust Mendelian randomization which provides valid causal inference under very general settings. The method may be performed without access to individual-level data, using only summary statistics of the type commonly published by genome-wide association studies.

115

MENTORING STUDENTS TO BECOME PROFESSIONAL STATISTICIANS THROUGH PROJECT WORK

Dr Tania Prvan ORCID iD, A/Prof Ayse Bilgin ORCID iD
Macquarie University, Sydney, Australia

Abstract

Project work units are an ideal way for students to demonstrate that they can analyse data that is messy, and then comprehensively communicate the results based on a research question and data at hand. Being able to prepare a data set for analysis, analysing data and communicating results are essential skills for finding employment as a statistician.

In statistics, the ideal way to help students transition from learning to job ready graduates is to design project work activities and assessments in such a way that students are provided with feedback at each step of their work to enable them to improve what they are doing and to become better communicators of their findings. In a sense, it is similar to an apprenticeship where novices begin their journey on becoming experts under the mentorship of their lecturers.

There are different ways of designing project units and different sources for real industry problems. However, the common theme for such project units is providing opportunities for students to demonstrate their prior learning in statistics. So, the emphasis is not on new statistical knowledge but it is on applying statistical knowledge to real problems. In this paper, we will demonstrate different ways of mentoring students in an undergraduate work integrated learning unit where the real industry problems are presented by industry partners and a postgraduate unit where students identify their own research questions (problems) from a set of given data sets. Where possible, we will share student work to support the value of mentorship.

117

Twinkling and Blinking: Resampling sparse radio lightcurves using Gaussian Processes

Mr Shih Ching Fu ORCID iD
Curtin University, Perth, Australia

Abstract

Time-series data are ubiquitous in astronomy. Typical examples are so-called lightcurves that describe the changing brightness of astronomical objects over time. Astronomers are particularly interested in linking the structure of lightcurves to the physics that underlie the object being observed. For example, oscillatory patterns give insight into the physical scale of a black hole.

However, lightcurves are often irregularly sampled and extremely sparse because of how astronomical observations are captured. For example, a field-of-interest might only be sporadically visible to an observatory during the course of many months. Consequently, dealing with sparsely sampled data is a major challenge in astronomy, and certainly relevant to upcoming endeavours such as the Square Kilometre Array (SKA) and Vera Rubin Observatory.

Techniques such as ARIMA and spectral analysis require evenly spaced data for the calculation of covariances and Fourier transforms, and are hence unsuitable for modelling lightcurves. A technique gaining popularity in astronomical time-series analysis is Gaussian Process (GP) regression, an approach that accommodates unevenly sampled time-series, and provides estimates of model uncertainty.

Gaussian processes are likened to 'infinitely dimensional' multivariate Gaussian random variables where a covariance kernel function describes the relationship between observations. Together with a mean function, the kernel function fully characterises a GP, where the choice of functional form and hyperparameters describes a plausible structure for the underlying data generating process.

Here we apply GP regression to modelling lightcurves from MeerKAT, an SKA precursor telescope situated in South Africa. As with other astronomical surveys, these data can contain observations of varied astrophysical phenomena, including orbiting exoplanets, neutron star stellar binaries, and black holes accreting matter from a companion star. We evaluate the efficacy of various kernel functions, provide guidance on the initial choice of hyperparameters, as well as comment on the interpretation of the power density spectra of the fitted models.

118

Developing the future workforce – supporting Industry 5.0 thinking and the cradle-to-career lifecycle

Professor Peter Howley[1,2,3], Emeritus Professor Tim Roberts[4]
[1]MCB Business Partners, Newcastle, Australia. [2]Hunter Innovation and Science Hub, Newcastle, Australia. [3]Hunter Medical Research Institute, Newcastle, Australia. [4]The University of Newcastle, Newcastle, Australia

Abstract

Industry 5.0 complements and extends upon the current 4th industrial revolution and STEM focus by specifically placing research and innovation at the service of the transition to a sustainable, human-centric and resilient Australian industry.

Research and innovation and the associated cross-functional skills required are fundamental to, and supported by, Statistics, Systems thinking, Sustainability and STEM (SSSS) skills... the focus of a Commonwealth Department of Education supported initiative in 2023 “Preparing for Industry 5.0 and beyond in light of COVID19: Facilitating the cradle-to-career life cycle” (www.ssss.org.au).

The presentation will describe this initiative which involves professional development for primary and high school educators, student workshops, and two national student competitions, aimed at positively impacting the attitudes, aspirations and abilities of school educators and students and preparing students for the future of work and Industry 5.0.

The emphasis will be on Statistics and Systems thinking, and how the initiative engages schools, supports the national curriculum learning outcomes, the education system, the future workforce and industry.

119

With non-linear dimension reduction, when can you believe what you see?

Jayani P.G. Lakshika ORCID iD[1], Dr Dianne Cook ORCID iD[1], Dr Paul Harrison ORCID iD[1], Dr Michael Lydeamore ORCID iD[1], Dr Thiyanga Talagala ORCID iD[2]
[1]Monash University, Melbourne, Australia. [2]University of Sri Jayewardenepura, Colombo, Sri Lanka

Abstract submission

Download file

120

Variable importance in likelihood-based regression models using the R package ‘vibe’

Stanislaus Stadlmann ORCID iD[1,2], Thomas Kneib ORCID iD[2]
[1]Sydney Informatics Hub, University of Sydney, Sydney, Australia. [2]Georg-August Universität Göttingen, Göttingen, Germany

Abstract

Classical regression methods are popular for quantifying relations between dependent and independent variables and making statements about their significance, but not made for ranking explanatory variables by their relative importance. In many fields, such as Psychology or Market Research, finding variables with the highest impact on the target variable is required. A viable relative importance metric should

a) take into account variable cross-correlation,

b) be independent of their order of inclusion,

c) measure error-reduction conditional on other covariates and

d) combine effects with multiple degrees of freedom.

Within the field of linear models, two metrics fulfill these requirements: “hierarchical partitioning” (often referred to as Shapley value), in which the average contribution of each variable in all possible model subsets is calculated and “relative weights”, a fast approximation of hierarchical partitioning based on orthogonalized predictors. Beyond linear models, however, a sizeable gap of variable importance measures remains. To fill this vacuum, we propose an extension of both previously mentioned metrics to likelihood-based models with linear predictors including generalized linear models (GLM) and generalized additive models for location, scale and shape (GAMLSS) with linear predictors. We present a package using R called vibe (Variable ImportanceBeyond Linear Regression). Our proposal’s effectiveness is proven with an extensive simulation and presented with two analyses: drivers of resident satisfaction in India described using a latent ordinal GLM and childhood malnutrition in India modeled with a multi-parametric GAMLSS.

Abstract submission

Download file

121

Assessment of the practices and processes of professional statisticians in a statistics capstone course.

Mrs Rachel Passmore
University of Auckland, Auckland, New Zealand

Abstract

In 2019 the University of Auckland (UoA) introduced compulsory capstone courses for all undergraduates in the Faculty of Science in recognition that students required more support to transition from being a student of a discipline to a practitioner. This research examined the practices and processes of professional statisticians/data scientists and categorised each practice and process within the framework of the UoA graduate profile. Course work submitted by students enrolled in a statistics capstone course was examined to determine whether and how such practices and processes could be demonstrated and/or developed in a statistics capstone course. Suggestions for capstone course models and assessment design in the light of this research will be presented.

122

Causal mediation and sensitivity of mediation effects using data from a randomised controlled trial of exercise, with and without diet, on function in knee osteoarthritis.

Dr Belinda Lawford ORCID iD[1], Prof Rana Hinman ORCID iD[1], Ms Fiona McManus ORCID iD[1], A/Prof Karen Lamb ORCID iD[1], Dr Thorlene Egerton ORCID iD[1], Dr Catherine Keating[2], Ms Courtney Brown[2], Ms Kathryn Oliver[1], Prof Kim Bennell ORCID iD[1]
[1]University of Melbourne, Parkville, Australia. [2]Medibank Private, Melbourne, Australia

Abstract

Introduction. Explore mediators of effects and sensitivity of causal mediation effects of two 6-month exercise programs, with and without diet, on function improvements in adults with knee osteoarthritis.

Methods. Secondary analysis of 345 participants from a three-arm randomised controlled trial of exercise (Exercise) and diet plus exercise (Diet+Exercise) versus information (Control). The outcome was 12-month change in function (WOMAC, 0-68). Potential mediators were change at 6 months in i) attitudes towards self-management, ii) fear of movement, iii) arthritis self-efficacy, iv) weight, and v) physical activity, and, at 6-months, vi) willingness for knee surgery. Causal mediation analyses were conducted where two regression models were simultaneously fitted for the outcome, considering each potential mediator for each treatment group comparison separately. Indirect (mediated) effects were estimated. Sensitivity analyses for indirect effects, with and without omitting an observed confounder, were conducted to investigate how robust the mediation analyses results were to violation of the sequential ignorability assumption.

Results. Mediators of Exercise vs Control may be reduced fear of movement (accounting for -1.11 [95% confidence interval: -2.15, -0.07] units improvement in function) and increased arthritis self-efficacy (-1.66 [-3.04, -0.28]). Mediators of Diet+Exercise vs Control may include reduced fear of movement (-1.13 [-2.17, -0.08]), increased arthritis self-efficacy (-5.15 [-7.34, -2.96]), and weight loss (-5.79 [-7.96, -3.63]). Where the mediation analysis results suggested mediation may be present, the sensitivity analyses suggest changes in the direction of the indirect effect occur at values of the sensitivity parameter of at least 0.2. Omitting an observed confounder reduced the sensitivity parameter by at most 0.13, suggesting 0.2 is a large critical value.

Discussion. Increased arthritis self-efficacy, reduced fear of movement, and weight loss may partially mediate effects of exercise programs, with and without diet, on function in knee osteoarthritis. Results appear insensitive to violation of the ignorability assumption.

124

Teach a student how to debug, and they can code for a lifetime: Open-source interactive self-paced R learning website

Danyang Dai[1,2], Dr. Emi Tanaka[3,2], Mitchell O'Hara-Wild[2]
[1]The University of Queensland, Brisbane, Australia. [2]Monash University, Melbourne, Australia. [3]The Australian National University, Canberra, Australia

Abstract

Learning and teaching statistics can be significantly enhanced by the use of statistical software, such as the R programming language. The current focus of teaching statistical software is to teach what is relevant to the subject and assume students know the basics of writing the code for statistical software. Students are thus often thrusted to learn programming while juggling the core statistical concepts. The challenge is that students come from all walks of life – some have no programming experience while others may have learnt different programming languages. In light of this challenge, we developed an online, self-paced interactive learning website called Learn R as an attempt to uniform students' knowledge. The Learn R website covers all the basics including R and RStudio installation, basic R syntax, data import, linear regression, data visualisation using ggplot2 and writing reproducible reports with R Markdown and Quarto.

One chapter of the Learn R website focuses on how students can debug and ask for help using a minimal reproducible example. We promote the idea of autonomous learning: "teach a student how to debug, and they can code for a lifetime" as one of the teaching goals of the Learn R website. By teaching students how to seek help, they would be better able to self-identify the issue themselves. In this presentation, I will share the design of the Learn R website and the key concepts of developing a self-guided interactive learning tool for statistical programming. This presentation aims to emphasise teaching students the best practices for debugging and seeking help, and provides potential for statistics educators to focus on core statistical materials by directing students to the self-guided Learn R website for self-study. For more information: https://learnr.numbat.space/

126

Inferring Changes to the Global Carbon Cycle with Wombat V2.0, a Hierarchical Flux-Inversion Framework

Dr Michael Bertolacci[1], A/Prof. Andrew Zammit-Mangion[1], Dr Andrew Schuh[2], Dr Beata Bukosa[3], A/Prof. Jenny Fisher[1], Dr Yi Cao[1], Dr Aleya Kaushik[4,5], Dist. Prof. Noel Cressie[1]
[1]University of Wollongong, Wollongong, Australia. [2]Colorado State University, Fort Collins, USA. [3]National Institute of Water and Atmospheric Research, Wellington, New Zealand. [4]National Oceanic and Atmospheric Administration, Boulder, USA. [5]University of Colorado, Boulder, USA

Abstract

The natural cycles of the surface-to-atmosphere fluxes of carbon dioxide (CO_2) and other important greenhouse gases are changing in response to human influences. These changes need to be quantified to understand climate change and its impacts, but this is difficult to do because natural fluxes occur over large spatial and temporal scales and cannot be directly observed. Flux inversion is a technique that estimates the spatio-temporal distribution of a gas' fluxes using observations of the gas' mole fraction and a chemical transport model. To infer trends in fluxes and identify phase shifts and amplitude changes in flux seasonal cycles, we construct a flux-inversion system that uses a novel spatially varying time-series decomposition of the fluxes. We incorporate this decomposition into the Wollongong Methodology for Bayesian Assimilation of Trace-gases (WOMBAT, Zammit-Mangion et al., Geosci. Model Dev., 15, 2022), a Bayesian hierarchical flux-inversion framework that yields posterior distributions for all unknowns in the underlying model. We also extend WOMBAT to accommodate physical constraints on the fluxes, and to take direct in situ and flask measurements of trace-gas mole fractions as observations. We apply the new method, which we call WOMBAT v2.0, to a mix of satellite observations of CO_2 mole fraction from the Orbiting Carbon Observatory-2 (OCO-2) satellite and direct measurements of CO_2 mole fraction from a variety of sources, in order to estimate the changes in the natural cycles of CO_2 fluxes that occurred from January 2015 to December 2020. We show that inferences from our method on how the global trends and seasonal cycles have changed during this period largely corroborate what is expected from a bottom-up understanding of the physical processes involved.

Abstract submission

Download file

127

Parasite clearance estimation for knowlesi malaria

Jeyamalar T Thurai Rathnam[1], Matthew Grigg[2,3], Timothy William[3], Siti Saimah Sakam[3], Daniel Cooper[4,3,2], Giri Rajahram[3,5], Bridget Barber[6,3,2], Nicholas Anstey[2,3], Saber Dini[1], Ali Haghiri[1], Megha Rajasekhar[1], Julie Simpson[1]

[1]University of Melbourne, Melbourne, Australia. [2]Menzies School of Health Research and Charles Darwin University, Darwin, Australia. [3]Infectious Diseases Society Sabah-Menzies School of Health Research Clinical Research Unit, Kota Kinabalu, Malaysia. [4]Department of Medicine, School of Clinical Medicine, University of Cambridge, Cambridge, United Kingdom. [5]Clinical Research Centre, Queen Elizabeth II Hospital, Ministry of Health, Kota Kinabalu, Malaysia. [6]QIMR Berghofer Medical Research Institute, Brisbane, Australia

Abstract

Background

Zoonotic *Plasmodium knowlesi* is the predominant cause of human malaria infections in Malaysia, driving the need for clinical studies to monitor the efficacy of antimalarial drugs. A key outcome measure for antimalarial drug efficacy is parasite clearance. This study aims to compare the standard two-stage approach, commonly used for *Plasmodium falciparum*, and an alternative Bayesian hierarchical modelling approach to estimate parasite clearance rates.

Methodology

Longitudinal parasite clearance data from 714 patients infected with *P. knowlesi* and enrolled in three clinical trials conducted in Sabah, Malaysia were analysed. The parasite clearance rates were estimated using the first stage of the standard two-stage approach which analyses each patient profile individually and the Bayesian hierarchical framework, which incorporates all profiles simultaneously. Both methods require the use a model that incorporates a lag, slope, and tail phase for the parasite clearance profile.

Results

The standard two stage approach estimated the clearance rates for 678 (95%) patients. The Bayesian method estimated a faster population mean parasite clearance estimate and visually, better model fits were observed. The artemisinin-based combination therapies were more effective in treating *P. knowlesi* compared to chloroquine, with estimated parasite clearance half-lives of 2.5 and 3.6 hours respectively using the standard two-stage method, and 1.8 and 2.9 hours respectively using the Bayesian method.

Conclusion

The standard two-stage approach is recommended for frequent parasite measurements as it is user-friendly and straightforward. Investigators should consider using the Bayesian hierarchical method if fewer parasite measurements are available to prevent potential selection bias.

128

Hybrid teaching- Effects and implementation of hybrid teaching in first-year tertiary statistical learning

Dr. Mitra Jazayeri ORCID iD, Dr. Andrew Buldt ORCID iD, Dr. Xia Li ORCID iD, Dr. Alysha De Livera ORCID iD, Associate Professor Andriy Olenko ORCID iD, Mr. Daniel Laurence ORCID iD
La Trobe University, Melbourne, Australia

Abstract

In response to the COVID-19 pandemic, educational institutions have implemented hybrid teaching and learning solutions that combine online and in-person students in concurrent synchronous learning sessions. However, research into the effectiveness of such solutions has produced mixed results. This study aims to address this gap by 1) identifying technological and pedagogical challenges encountered during the delivery of a statistical subject for first-year tertiary health sciences students and 2) examining whether students' intrinsic and extrinsic learning motivation affects their engagement and satisfaction with hybrid learning. The data analysis will be supported by utilizing well-designed instruments and data from both the student information system and learning management system (LMS). This presentation will provide novel theoretical and practical insights into the application of hybrid teaching in tertiary statistical education.

129

Understanding our students' viewing habits (of our recorded lectures)

Dr Emily Nordmann[1], Dr Craig Alexander[1], Dr James Bartlett[1], Dr Mitchum Bock[2], Dr Eilidh Jack[1], Dr Gaby Mahrholz[1]
[1]University of Glasgow, Glasgow, United Kingdom. [2]University of Glasgow, University of Glasgow, United Kingdom

Abstract

While lecture video recording is a widely adopted technology, its use has changed due to the pandemic and the shift to remote learning (and back again). As a result, the way students engage with lectures has changed with the normalization of online learning. Lecture recording analytics have the potential to provide valuable insights into student engagement, learning outcomes and teaching effectiveness. For example, such analytics provide a source of data for evidence-based reflections on patterns of student engagement with recorded lectures.

However, there are barriers to engaging with learning analytic data such as lecture recording analytics, including a lack of awareness, skills and time. To address the skills barrier, we present open-access tutorials that support educators to analyse the learning analytic data generated by the lecture recording platform Echo360 using R. Our tutorials have been designed to be accessible to users from all backgrounds with little or no expertise using R, enabling them to take advantage of the data available. Our intention is that the tutorials will help develop skills and confidence in educators to investigate their own teaching and learning practices using data for an evidence-based approach. Whilst the tutorials are focused on Echo360 data, the skills covered are applicable to other sources of data from a wider range of platforms which can also provide insights into, and critical reflection on, learning and teaching practices.

We describe three tutorials available as open-access materials:

1. Getting started with Echo360 data in R

2. Exploring Echo360 video/course level data in R

3. Analysing Echo360 data with other data sources in R

Applications of the tutorials will be presented using video recording analytics from several undergraduate courses taught at the University of Glasgow with critical reflections on blended learning and student engagement.

This project was funded by an Echo360 Impact Grant.

130

Multivariate meta-analysis methods for high-dimensional data

Dr Alysha De Livera ORCID iD[1,2], Ms Jayamini Liyanage[1], Prof Luke Prendergast ORCID iD[1]
[1]La Trobe University, Melbourne, Australia. [2]The University of Melbourne, Melbourne, Australia

Abstract

Meta-analysis is a statistical method that combines quantitative results from multiple independent studies on a particular research question or hypothesis, with the goal of making inference about the population effect size of interest. Traditional meta-analysis methods have focused on combining results from multiple independent studies, each of which has measured an effect size associated with a single outcome of interest. Modern studies in evidence synthesis, such as those in biological studies have focused on combining results from studies which have measured multiple effect sizes associated with multiple correlated outcomes. This presentation will introduce multivariate meta-analysis methods for obtaining summary estimates of the effect sizes of interest for high-dimensional data, and describe the statistical challenges in doing so, with applications to real and simulated high-dimensional data.

131

Performance of multivariate time series models over univariate approaches in modelling correlated count data

Mr Kalu Arachchilage Nishantha Kumara Karunarathna[1,2], Dr. Nicholas J. Clark[1]
[1]The University of Queensland, Gatton, Qld, 4343, Australia. [2]Eastern University, Sri Lanka, Vantharumoolai, Sri Lanka

Abstract

Time series modelling is used for analysis and forecasting in every field. Most of traditional time series models are not able to well model series of discrete data. There is also a lack of tools to deal with collections of count-valued time series. Analysis and forecasting of count responses from correlated entities become tedious task because of interdependent structure among entities.

There is a need for modelling tools that can handle interdependent count time series with options for estimating contemporaneously correlated errors, moving average processes and lag interdependences. We showcase how multivariate count time series can be analysed in a Bayesian framework. Our models can quantify nonlinear covariate relationships while capturing unobserved (and possibly multivariate) temporal dynamics in a joint probabilistic framework with hierarchical impacts using Hamiltonian Monte Carlo (HMC) simulation with Stan language in R environment.

It is essential to understand how key user decisions impact model estimation, inferences on parameters, interpretation process and forecasting performance. We explore consequences of model misspecifications (models, order, structure), distributional assumptions (including common transformation routines that assume Gaussian observations), modelling assumptions on parameters (time invariant mean and variance-covariance structures) and prior specifications.

Using both simulations and empirical data, we show that a wide variety of latent multivariate trend models can be fit with HMC. Compared to univariate modelling approaches, Bayesian based multivariate time series models tend to perform better in forecasting tasks and can uncover important insights about temporal dependence. However, we find that overall performances of modelling depend on goals of analysis and decisions on models' specifications. Among these specifications, trend dimension (univariate-multivariate), number of series to be modelled jointly, compositional structures of models, distributional assumptions and prior choices are vital to consider.

132

Visually exploring local explanations to understand complex machine-learning models

Professor Dianne Cook ORCID iD[1], Dr Nicholas Spyrison ORCID iD[1], Dr Przemyslaw Biecek ORCID iD[2]

[1]Monash University, Melbourne, Australia. [2]Warsaw University of Technology and University, Warsaw, Poland

Abstract

Explainable Artificial Intelligence (XAI) is an emerging field of methods being developed to help interpretability for complex nonlinear models. At the core of XAI are local explanations, which provide insight into what factors most contribute to an individual prediction. Local interpretations examine the local neighbourhood of a predicted value. They can be converted into a linear combination of variables, which can then be interactively examined, against the predicted and observed values, using manual tour methods. This talk will describe how local explanations work, and also how to investigate the validity of any interpretations using visualisation.

133

Confidence intervals for proportions estimated by group testing based on Firth's bias correction

A/Prof Graham Hepworth[1], Dr Brad Biggerstaff[2]
[1]Statistical Consulting Centre, The University of Melbourne, Melbourne, Australia. [2]Centers for Disease Control and Prevention, Fort Collins, USA

Abstract

Group testing (or pooled testing) arises when units are pooled together and tested as a group for the presence of an attribute, such as a disease. It originated in blood testing, but has been applied in many fields, including plant disease assessment and prevalence estimation of mosquito-borne viruses – the two fields in which we have encountered the technique.

Confidence intervals for proportions estimated from group testing have been studied by a variety of researchers, who have examined both exact and asymptotic methods. Recent work in point estimation (Hepworth and Biggerstaff 2017) has seen the use of Firth's correction (Firth 1993) to maximum likelihood estimation to reduce bias effectively. Firth's correction is to shift the score by a parameter-dependent function incorporating both the bias of the MLE and the Fisher information, so that the resulting maximiser of the corrected score is a bias-corrected estimator.

Considering the Firth-corrected score as the first derivative function of a penalised likelihood, we develop confidence intervals wholly within this framework. We evaluate the performance of the intervals, and compare them to the existing, recommended asymptotic method.

References:
Hepworth G. and Biggerstaff, B.J. (2017) Bias correction in estimating proportions by pooled testing. JABES, 22, 602–614.
Firth, D. (1993) Bias reduction of maximum likelihood estimates. Biometrika, 80, 27–38.

134

Quantifying streamflow trends in mining impacted catchments

Rajitha Athukorala ORCID iD[1,2,3], Maria Dubikova[3], Quinn Ollivier ORCID iD[3], R. Willem Vervoort ORCID iD[1,2]

[1]ARC Industrial Transformation Training Centre in Data Analytics for Resources and Environments, Sydney, Australia. [2]School of Life and Environmental Sciences, The University of Sydney, Sydney, Australia. [3]WaterNSW, Parramatta, Australia

Abstract

Identification and quantification of trends in hydrological data can be a challenging task particularly in catchments that are stressed with anthropogenic influences such as mining. The complexity of hydrological and climate systems, and their many interactions, requires methodologies that can identify the individual contributions from multiple drivers, such as rainfall, mining, and climate change.

We propose a Generalized Additive Mixed Model (GAMM) approach to analyse streamflow, which combines individual effects from a range of predictor variables.

As a first step, rainfall, seasonality, and time were used as the predictor variables to model the log transformed monthly averaged streamflow. An auto regressive first order process was assumed in the model to account for the auto correlation in the time-series data.

Preliminary results indicate how each effect on streamflow is quantified in a hydrologically consistent pattern in relation to rainfall, seasonality, and trends in time. The modelled trends in time will be further investigated to deconstruct the combined effects of mining, climate change and possible other impacts. Current research involves the use of mining progress as a predictor variable along with climate indices like SOI to better understand the system's dynamics and responses. The limitation of the study is in the limited amount of data available for pre-mining periods to draw comparisons with baseline catchment conditions.

Abstract submission

Download file

135

A Plot is Worth a Thousand Tests: Assessing Residual Diagnostics with the Lineup Protocol

Mr Weihao Li ORCID iD[1], Professor Dianne Cook ORCID iD[1], Dr Emi Tanaka ORCID iD[2,1], Assistant Professor Susan VanderPlas ORCID iD[3]
[1]Monash University, Melbourne, Australia. [2]Australian National University, Canberra, Australia. [3]University of Nebraska, Lincoln, USA

Abstract

Regression experts consistently recommend plotting residuals for model diagnosis, despite the existence of hypothesis procedures. Assessing model fit with residual plots alone, however, poses an issue with the lack of calibration: humans tend to over-interpret the patterns in a plot. The issue can be mitigated by the use of the lineup protocol: by embedding the residual plot among other plots generated from the null hypothesis. This process is part of the framework of "visual inference". In this talk, we provide evidence for why plotting residuals for model diagnostics is good advice, using data from a visual inference experiment. We show that conventional tests have significant drawbacks, and how the lineup protocol can be used instead to yield reliable and consistent evaluation of residual plots for better model diagnosis.

136

Interpretable Models - When accounting for confounders is just dodgy accounting

Mr christopher howden ORCID iD
Sydney University, Sydney, Australia

Abstract

Statistics often finds itself in competition with other quantitative fields not just for funding, but also relevance. This talk begins by looking at that often-heard expression “I accounted for variable X by including it in the model” and asking the questions “What do researchers mean by this?”, “Are they doing it correctly, or is it just dodgy bookkeeping?”, and “Why is it needed for interpretable models, and why are they so important?”.

We finish by using the prior exposition on “How to account for variable X” to explore how statisticians can become researchers preferred quantitative partners. Discussing our strengths and where we need to do better to ensure they see us as trusted advisors on the complex issues they face when building interpretable models. These models being the engines driving knowledge creation not just in science, but in an increasingly diverse range of domains in the digital age. Doing so should lead to a greater share of those scarce resource dollars and also boost our profession’s profile and involvement in ground-breaking impactful research. Ensuring our relevance in a fast-moving quantitative research world.

137

SimpleLMM: An easy-to-use RShiny interface for testing longitudinal data

Rodrigo Canovas ORCID iD[1], James Doecke[2]
[1]Australian E-Health Research Centre, CSIRO, Melbourne, Australia. [2]12Australian E-Health Research Centre, CSIRO, Brisbane, Australia

Abstract

Background

Longitudinal analysis is essential in clinical research where participants are studied over time, recording all the parameters of interest at each visit (for example, their risk factors, biomarkers, health outcomes, etc). While there are several available libraries to perform this kind analysis, we are still missing a user-friendly interface that allows statisticians and non-statisticians the ability to perform simple longitudinal analysis that can graphically display results in a easy way to understand.

Methods & Results

In this work we present the RShiny “SimpleLMM” application, which allows the user to compute a linear mixed model (LMM) over a dataset. SimpleLMM presents a user interface (UI) that allows the user to select and create formulae to compute the LMM over the data. After this step, the tool enables the user to explore the created model by showcasing the results from statistical modelling using the parameters selected and offering different plot options. Moreover, by modifying various experimental parameters, such as variable constraints or stratification parameters, researchers are able to see how these measures change in different scenarios. Within the application, users can statically and dynamically select relevant parameters to perform statistical analysis for calculating cohort requirements during the planning phase of an experimental set-up. That is, the application tool provides statistical information that helps to design and test possible scenarios, providing useful information to guide the design of future cohort requirements.

Summary

This app presents a longitudinal analysis statistical summary tool, allowing researchers to access a range of statistical analyses through numerical and graphical representation. This provides a fast way to assess LMM results over a given data set in a clear and simple way, enabling statisticians and non-statisticians to explore a simple assessment and design considerations needed in setting up new longitudinal research projects based on their data.

138

Approximate Bayesian computation for long memory processes.

Dr Clara Grazian[1,2], Mr James Gabor[3]
[1]University of Sydney, Sydney, Australia. [2]DARE, Sydney, Australia. [3]UNSW, Sydney, Australia

Abstract

This work investigates a Bayesian approach to estimating the parameters of long memory models, in particular ARFIMA models. Long memory, i.e. the phenomena of hyperbolic autocorrelation decay in series, has attracted much attention, since in many situations the assumption of short memory, for example the Markovianity assumption, can be considered too strong. Applications ca be easily found in astronomy, finance, and environmental sciences; however, current parametric and semiparametric approaches to long-memory modelling present difficulties, especially in the estimation procedure. We present a novel approach to approximating the joint posterior distributions of ARFIMA model parameters using approximate Bayesian computation (ABC), which allows to approximate the posterior distributions of the parameters given the observed finite series, without making use of asymptotic arguments. Acceptance of simulated long-memory parameters is based on the periodogram: an estimate of the spectral density which captures the dominance of long term non-negligible correlations, characteristic of long-memory ARFIMA processes. A simulation study and an example on daily log returns for Standard and Poor's 500 index will show the advantages of the proposed approach.

139

Constructing Large Nonstationary Spatio-Temporal Covariance Models via Compositional Warpings

Quan Vu[1], Andrew Zammit Mangion[1], Stephen Chuter[2]
[1]University of Wollongong, Wollongong, Australia. [2]University of Bristol, Bristol, United Kingdom

Abstract

Understanding and predicting environmental phenomena often requires the construction of spatio-temporal statistical models, which are typically Gaussian processes. A common assumption made on Gausian processes is that of covariance stationarity, which is unrealistic in many geophysical applications. In this talk, we introduce a new approach to construct descriptive nonstationary spatio-temporal models by modeling stationary processes on warped spatio-temporal domains. The warping functions we use are constructed using several simple injective warping units which, when combined through composition, can induce complex warpings. A stationary spatio-temporal covariance function on the warped domain induces covariance nonstationarity on the original domain. Sparse linear algebraic methods are used to reduce the computational complexity when fitting the model in a big data setting. We show that our proposed nonstationary spatio-temporal model can capture covariance nonstationarity in both space and time, and provide better probabilistic predictions than conventional stationary models in both simulation studies and on a real-world data set.

140

A quick and flexible visualisation system for the designs of experiments

Dr Emi Tanaka ORCID iD
Australian National University, Canberra, Australia. Monash University, Melbourne, Australia

Abstract

Experimental data are hallmarks of scientific evidence to prove or disprove theories or hypotheses. Multiple people with different expertise are typically involved in planning and executing experiments but rarely is the communication easy or seamless, especially across people from different domains, yet we predicate on the assumption that misapprehensions will be somehow sorted out. This assumption leaves the success of an experiment at the mercy of the interpersonal communication skills of people involved. Rather than leaving the success of an experiment to serendipity, I propose a novel framework to robustify the workflow of the construction of experimental designs that encourages users to deliberate on understanding the experimental structure. This framework, called "the grammar of experimental designs", considers an object oriented system to encapsulate the experimental structure in a cognitive programming approach. In this talk, I extend the "the grammar of experimental designs" system in the edibble R-package to automate construction of the visualisation object; this system is implemented as the deggust R-package.

141

Mediation analysis with multiple mediators: a target trial approach

Associate Professor Margarita Moreno-Betancur ORCID iD
Murdoch Children's Research Institute, Melbourne, Australia. University of Melbourne, Melbourne, Australia

Abstract

Many research questions concern the causal pathways that are hypothesised to mediate an association. Invariably, the translational intent of such research is to inform potential intervention targets, but until recently mediation effect definitions did not acknowledge this interventional intent. In a recent body of work, we proposed a novel framework that conceptualises mediation effects by mapping to a hypothetical “target” randomised trial evaluating mediator interventions. This approach is particularly relevant for mediators that do not correspond to well-defined interventions, which arise frequently, and perhaps can only be addressed by considering hypothetical interventions that would shift the mediators’ distributions. The approach proposes specifying a target trial to capture the research question in the form of a measure of the impact of shifting joint mediator distributions to user-specified distributions that represent the effect of the hypothetical intervention. These estimand assumptions are distinguished from identifiability assumptions, which are needed to emulate and thus estimate the effects with the observed data. By its nature, the approach is context-specific. Drawing on learnings from applications to several longitudinal cohort studies, five of which are already published, this talk will provide an overview of this framework and discuss alternative approaches to effect definition according to the question in diverse contexts. A workflow that will assist researchers in applying the method in practice is described.

142

Combining multiple imputation and inverse probability weighting to address covariate missing data and loss to follow-up in longitudinal studies

Ms. Melissa Middleton, Dr. Cattram Nguyen, Prof. John Carlin, A/Prof. Margarita Moreno-Betancur, Prof. Katherine Lee
Murdoch Children's Research Institute, Melbourne, Australia

Abstract

Missing data are ubiquitous in longitudinal studies. When estimating the effect of an exposure on a later outcome, loss to follow-up can result in a large proportion of missing outcome data, with covariate missingness often present as well. Omitting incomplete records from analyses may result in biased parameter estimates and reduced precision. Two approaches commonly used to handle missing data are multiple imputation (MI) and inverse probability weighting (IPW). Both MI and IPW can reduce bias in the effect estimate over a complete-case analysis (CCA) through the inclusion of auxiliary information, with MI additionally improving precision. However, when imputing missing values in a large proportion of records, MI may result in biased inferences if the imputation model is misspecified. Conversely, IPW may not be efficient when there are multiple incomplete variables with different missingness patterns. An alternative that might capitalise on the relative strengths of each of these two approaches is to use MI to handle incomplete covariates and IPW to handle the incomplete outcome.

We conducted a simulation study to assess the performance of a combined MI/IPW approach, compared with MI-only, IPW-only, and CCA, in a longitudinal study estimating the effect of an exposure on an outcome at a later timepoint. The study considered a range of scenarios, varying the amount of missing data, outcome type, missing data mechanism, and sample size. We also illustrated the approaches in a case study.

MI/IPW showed a slight bias in the effect estimate, particularly when the outcome was continuous, or a large proportion had missing data. MI-only resulted in unbiased estimation of the effect estimate and its standard error across all scenarios considered. IPW-only produced biased effect estimates and standard errors in small samples. Overall, MI-only may be the preferred approach when handling covariate and outcome missingness in longitudinal studies.

143

Uncovering biological signatures of Alzheimer's Disease through integrative omics

Dr Michael Vacher ORCID iD[1,2], Dr Rodrigo Canovas[3], Dr James Doecke[4]
[1]CSIRO, Kensington, Australia. [2]ECU, Joondalup, Australia. [3]CSIRO, Parkwood, Australia. [4]CSIRO, Herston, Australia

Abstract

Abstract submission

Download file

144

METHODOLOGY FOR SYSTEMATIC IDENTIFICATION AND ANALYSIS OF MULTIPLE BIASES IN CAUSAL INFERENCE: A TARGET TRIAL APPROACH AND A SIMULATION STUDY

Dr Rushani Wijesuriya ORCID iD[1,2], Prof John B Carlin ORCID iD[1,2], A /Prof Rachel L Peters ORCID iD[1,2], A/Prof Jennifer L Koplin ORCID iD[1,3], A/Prof Margarita Moreno-Betancur ORCID iD[1,2]

[1]Murdoch Children's Research Institute, Melbourne, Australia. [2]University of Melbourne, Melbourne, Australia. [3]Child Health Research Centre, University of Queensland, Queensland, Australia

Abstract

See attached

Abstract submission

Download file

145

Are house prices reflecting the risks posed by climate change?

Dr Patricia Menendez[1], A/Prof Maria Jesus Barcena[2], Dr Cristina Gonzalez[2], Prof Fernando Tusell[2]

[1]Monash University, Melbourne, Australia. [2]University of the Basque Country, Bilbao, Spain

Abstract

Owning a home is often considered one of the most significant and valuable assets for households, comprising a substantial portion of their net worth. However, the latest report by the Intergovernmental Panel on Climate Change warns of the accelerating rise of global mean sea level, which could potentially affect homes situated in coastal areas and near fluvial regions in the Basque Country. The purpose of this study is to investigate whether house prices have taken into account the risk of flooding associated with climate change. Our dataset comprises more than 300,000 observations including dwelling transaction prices, rental values, and property characteristics. We use geographically weighted regression with tailored neighbourhoods to analyse the differences in house prices between high-risk and low-risk areas. We use visual representations of data and risk areas to facilitate our analysis and enhance our interpretation of results.

146

Sensitivity analysis for outcome missingness assumptions in causal inference: accommodating the substantive analysis

Jiaxin Zhang[1,2], S.Ghazaleh Dashti[1,2], John B. Carlin[1,2], Katherine J. Lee[1,2], Margarita Moreno-Betancur[1,2]

[1]the University of Melbourne, Melbourne, Australia. [2]Murdoch Children's Research Institute, Melbourne, Australia

Abstract

When conducting multiple imputation (MI), maintaining compatibility between the imputation model and substantive analysis is important for avoiding bias. Recently, two compatible MI approaches have been developed: the “Substantive-Model-Compatible-Fully-Conditional-Specification (SMCFCS)” which accommodates the substantive model in the FCS procedure to ensure compatibility; and a stacked-imputation-based approach (SMC-stack), which multiply imputes non-outcome variables ignoring the outcome and stacks them into a single dataset that is analysed using weights proportional to the substantive model density. Both methods are guaranteed to be unbiased under the “missing-at-random” assumption. In practice, however, it is common for an incomplete outcome to be a cause of its own missingness or to be associated with its missingness, which implies a violation of this assumption. Although methods such as “not-at-random (NAR) FCS” provide an appealing approach for sensitivity analysis (SA) with multivariable missingness, compatible approaches are lacking, which is a key gap as using incompatible imputation in SA may induce bias. This is particularly pertinent when estimating the average causal effect (ACE) using g-computation, which uses an outcome model including exposure-confounder interactions. To address this gap, we propose two approaches for compatible SA for the missingness assumptions when an incomplete outcome causes. The proposed approaches, NAR-SMCFCS and NAR-SMC-stack, extend SMCFCS and SMC-stack, respectively, by incorporating the outcome missingness indicator with an associated sensitivity parameter when imputing the outcome. We evaluated the performance of our proposed methods through a simulation study motivated by a real case study, which we also analysed. We considered a range of outcome models and multivariable missingness mechanisms, where the substantive analysis aimed to estimate the ACE using correctly specified g-computation. The simulation results showed that both approaches reduced the bias in ACE estimation compared to default NARFCS approach. We conclude that NAR-SMCFCS and NAR-SMC-stack are preferred to conduct SA for missingness assumptions in causal inference.

147

Network Analysis of Relationships and Change Patterns in Depression, Multiple Chronic Diseases and Other Factors

Dr Xia Li
La Trobe University, Melbourne, Australia

Abstract

Objectives: The aim of this study is to explore the relationships of depression, chronic diseases, inflammation, and other factors which can be conceptualized as a network system in a 7-year Health and Retirement Longitudinal Study dataset. Method: We estimated Mixed graphical model (MGM) networks for the selected items and then explored the network interconnections, stability, network temporal difference, network community and bridge nodes. Results: Our network analyses reveal among a variety of factors in a large cohort sample of middle-aged participants, the most central items, and strongest associations. These results were robust to stability tests. No temporal difference was found between year 2011 and 2018. Five communities were found, and depression was found the important bridge symptom in the Neurological Disorders community which connecting to the other communities. Conclusions: Network analysis is a useful tool to explore the interconnections in a multivariate and multi causal framework. Future studies should be focus on verifying our findings in other samples and provide novel insights. Finally, we discuss limitations and challenges for future research.

148

Spatial prediction of non-negative spatial processes using asymmetric loss functions with environmental applications

Distinguished Professor Noel Cressie ORCID iD, Mr Alan Pearse ORCID iD, Dr David Gunawan ORCID iD
University of Wollongong, Wollongong, Australia

Abstract

A major component of inference in spatial statistics is that of spatial prediction of an unknown value of a latent spatial process, based on noisy measurements of the process taken at various locations in a spatial domain. The most commonly used predictor is the conditional expectation of the unknown value given the data, and its evaluation is obtained from hierarchical-statistical-modelling assumptions about the probability distributions of the latent process and the measurements of that process. By considering the spatial-prediction problem from a decision-theoretic viewpoint, one can recognise the conditional expectation as optimal for squared-error loss (SEL). In this talk, we consider spatial prediction of processes that take non-negative values, for which SEL and the conditional expectation are poorly adapted. However, the family of phi-divergence loss functions are well defined for predictor and predictand when the spatial process is non-negative, and losses are asymmetric (unlike SEL). Taking a hierarchical spatial-statistical-modelling approach, it can be seen that this new class of asymmetric loss functions generate new optimal spatial predictors, for which new characteristics of the predictive distribution are needed. We feature the important sub-class of power-divergence loss functions, which is indexed by the choice of a power parameter. An application is given to spatial prediction of zinc concentrations in soil on a floodplain of the Meuse River in the Netherlands.

149

How to train animal disease detectives in the Asia-Pacific – A practical guide for developing e-learning modules on statistics

Dr Alexandra Green[1,2], Dr Harish Tiwari[2,3], Amanda Evans[2], Professor Annette Burgess[4], Dr Imas Yuyun[5], Dr Martha Simanjuntak[5], Dr Widagdo Sri Nugroho[6], Dr Khrisdiana Putri[6], Dr Caitlin Pfeiffer[7], Professor Mark Stevenson[7], Associate Professor Navneet Dhand[2]
[1]The Sydney Informatics Hub, Core Research Facilities, The University of Sydney, Sydney, Australia. [2]Sydney School of Veterinary Science, The University of Sydney, Sydney, Australia. [3]Jyoti and Bhupat Mehta School of Health Science and Technology, Indian Institute of Technology Guwahati, Guwahati, India. [4]Sydney Medical School, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia. [5]Directorate of Animal Health, Directorate General of Livestock and Animal Health Services, Ministry of Agriculture, Jakarta, Indonesia. [6]Faculty of Veterinary Medicine, Universitas Gadjah Mada, Yogyakarta, Indonesia. [7]Melbourne Veterinary School, The University of Melbourne, Parkville, Australia

Abstract

There is a significant risk of an emerging infectious disease (EID) outbreak in Australia from neighbouring countries, and EIDs can severely impact farmer livelihoods, as well as human and animal health. To combat this, the Asia-Pacific Consortium of Veterinary Epidemiologists (APCOVE), formed in 2020, aims to strengthen the field veterinary epidemiology capacity in the Asia-Pacific by training animal disease detectives in outbreak investigation and disease surveillance. To this end, APCOVE has developed and delivered 36 e-learning modules on various epidemiological topics, including three modules on the fundamentals of data analysis. In 2022, 93/139 (67%) trainees completed all modules delivered asynchronously over six months. Geographical barriers to participation were mitigated by the online delivery to provide a cost-effective means to widen participation. Notably, the APCOVE model has provided a valuable tool to educate participants in statistical concepts.

The modules were tailored to veterinarians in Vietnam, Laos, Myanmar, Cambodia, the Philippines, Indonesia, Timor-Leste, and Papua New Guinea. They considered the skill-levels of all trainees and contained context-specific scenarios and datasets. The modules were developed in English and are now being translated into six other languages to increase accessibility. To enhance trainee engagement, the modules were built using the highly interactive, easy-to-navigate e-learning platform Rise360. The choice of statistical software was based on its ease of use, and to demonstrate statistical techniques using this software, 21 videos were included in the modules. But to allow transmission in areas of target countries with low bandwidth, the videos were kept to a short duration. Feedback received from the trainees in focus group discussions and questionnaires has also been incorporated into the modules to improve their clarity. The modules are now undergoing a rigorous referencing and copyright check and will be made available on the APCOVE website free of charge for anyone to utilise.

150

Juror understanding of statistical evidence: a systematic review.

Dr Alanah Cronin ORCID iD[1], Professor Ben Mathews ORCID iD[2], Associate Professor Janet Chaseling[3], Associate Professor Dimitrios Vagenas[4]

[1]School of Biomedical Sciences, Faculty of Health, Queensland University of Technology, Kelvin Grove, Australia. [2]Faculty of Law, Queensland University of Technology, Brisbane, Australia. [3]School of Environment and Science, Griffith University, Nathan, Australia. [4]Research Methods Group, School of Public Health and Social Work, Faculty of Health, Queensland University of Technology, Kelvin Grove, Australia

Abstract

As legal evidence in criminal cases is usually presented with some degree of uncertainty, the correct understanding and interpretation of statistical evidence by jurors, as well as the presentation of such evidence by expert witnesses and legal professionals should be central to the correct function of the criminal justice system. There has been an increase in queries regarding applied methods used to report statistical evidence and the way it is presented to a lay audience. Numerous attempts have been made to understand how jurors interpret statistical evidence that offer different opinions (see Martire, 2018). However, there is neither consensus on the most appropriate approach of presenting evidence in practice, nor an in-depth understanding of how laypeople understand probability and statistics in varying contexts.

The systematic literature review presented attempted to systematically and critically evaluate current attempts to understand how jurors interpret statistical evidence and provide an quantification of the overall quality of existing literature. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were followed (Page et al., 2021), and identified 23 empirical studies from 14 multidisciplinary databases up to the 30th April 2020.

The main findings were: (i) a lack of consistency in how we measure understanding and comprehension, (ii) variation in the recommended presentation of statistical evidence, (iii) broad failures in the reporting of quality metric items, and (iv) four themes that were identified to have a potential impact on juror comprehension of statistical evidence. The themes identified by the empirical studies included in this review were (a) the individual's perception of evidence and the justice system, (b) prior beliefs regarding the severity of the crime, (c) the individual's learning style (i.e., visual, auditory, or written), and (d) demographic factors.

151

Spatial Models for Alcohol Effects

Dr John Henstridge, Dr Fiona Evans, Mrs Anna Hayes, Ms Elise Corless, Mr Lachlan Robinson
Data Analysis Australia, Perth, Australia

Abstract

A key assumption in many efforts to limit the harms caused by alcohol is that higher physical availability in the form of retail outlets increases many forms of harm. Hence it is not surprising that many statistical studies have explored the relationship between outlet density or outlet proximity and levels of harm in the community. However this is challenging from both data availability and quality as well as statistical modelling issues where the spatial and temporal dependences must be managed, making it relatively easy to highlight shortcomings of many studies. The current project used Australian data to develop a best practice methodology to overcome these shortcomings whilst still being practical. A key component has been the detailed predictive structure using point (such as an outlet) to area (over which harm is measured) relationships using road distances as a proximity measure within a constrained temporal-spatial generalised linear model. The results suggest that at least some of the findings in the literature are the result of inadequate models. This is likely to be in part due to the inadequate incorporation of socio-economic and demographic differences, as well as poorly structured models that do not properly handle scale effects.

152

Facilitating juror comprehension amongst statistical evidence challenges: a survey of lay comprehension in Australia.

Dr Alanah Cronin ORCID iD[1], Associate Professor Janet Chaseling[2], Professor Ben Mathews ORCID iD[3], Associate Professor Dimitrios Vagenas[4]
[1]School of Biomedical Sciences, Faculty of Health, Queensland University of Technology, Kelvin Grove, Australia. [2]School of Environment and Science, Griffith University, Nathan, Australia. [3]Faculty of Law, Queensland University of Technology, Brisbane, Australia. [4]Research Methods Group, School of Public Health and Social Work, Faculty of Health, Queensland University of Technology, Kelvin Grove, Australia

Abstract

Presenting evidence with absolute certainty is rare in criminal justice. Usually there is a degree of uncertainty, which must be stated, quantified, and understood for justice to be delivered. This study focuses on the understanding of laypeople in their capacity as jurors and attempts to contribute knowledge on the understanding of statistical evidence by laypeople and how their statistical capability differs between two scenarios: (*i*) in everyday probability (dice and lottery) and a (*ii*) in forensic context.

A mixed-recruitment survey of 204 lay Australians was conducted that measured participants' self-nominated understanding of: (*i*) the Likelihood Ratio (LR) and (*ii*) the Random Match Probability (RMP), (*iii*) a dice statement, and (*iv*) a lottery statement on a 5-point Likert scale. Self-nominated understanding was followed by a knowledge manipulation check that asked participants to correctly interpret the same statement. Polychoric correlation showed a strong positive correlation between the self-perceived understanding of the RMP and LR (ρ = 0.79, $\rho\rho$ = 0.04, p < 0.001) and between the dice and lottery statements (ρ = 0.82, $\rho\rho$ = 0.05, p < 0.001). Results of a multivariate ordinal logistic regression suggested that participants with higher educational qualifications appeared to nominate higher self-perceived understanding of forensic statements and that education may indicate an individual's confidence in the interpretation of statistical evidence, but not general probability.

When presented with a knowledge manipulation check, the LR and RMP were correctly interpreted by 66% of participants, the dice statement correctly interpreted by 68%, and the lottery statement correctly interpreted by 74% of participants. The results for suggest that perhaps the method of presenting statistical statements is not as influential to juror understanding as previously thought (e.g., see Koehler, 1997; Nance and Morris, 2005; Thompson and Newman, 2015) and that jurors may misinterpret the evidence regardless of the statement type.

153

Implementation of g-computation in practice: a new diagnostic tool to guide outcome model specification

Dr Daisy Shepherd ORCID iD[1,2], Stijn Vansteelandt ORCID iD[3], A/Prof Margarita Moreno-Betancur ORCID iD[2,1]
[1]The University of Melbourne, Melbourne, Australia. [2]The Murdoch Children's Research Institute, Melbourne, Australia. [3]Ghent University, Ghent, Belgium

Abstract

Causal inference is a central goal of clinical and public health studies, investigating the effect of an exposure on an outcome of interest. For many studies, reliance on observational data is common, requiring confounding-adjustment methods to estimate causal effects. G-computation is one such method, which in the point-exposure setting extends outcome regression by allowing exposure-confounder interactions in the outcome model and predicts counterfactual outcomes across the sample under each exposure. Consistent estimation with g-computation relies on correct specification of the outcome model, which cannot be empirically verified. It is recommended that variables included in the model should be driven by expert-knowledge. However, there is no formal guidance or diagnostic tool available to aid the parametric specification of the outcome model, for example which interaction or non-linear terms to include, presenting a challenge when applying g-computation in practice.

In this work we aimed to address this gap, proposing a new diagnostic tool to guide the outcome model specification in g-computation. Our proposed tool distinguishes between candidate outcome model specifications based on the expected bias in the resulting g-computation estimator. This bias is derived from the efficient influence curve and depends on the outcome model specification and propensity scores (PS). Our method estimates the PS flexibly and equally across candidate outcome model specifications using SuperLearner. Thus, the differences in the resulting statistic depend solely on the outcome model specification.

We investigated the performance of the diagnostic tool in a simulation study based on the Longitudinal Study of Australian Children (LSAC), considering a range of true outcome generation models and sample sizes. Results indicated the statistic was optimised for the correctly specified model in most settings, and appropriately discriminated the model that minimized the bias in effect estimates. Implementation of the tool is available as an R function.

154

The noise and the signal: conveying the importance of uncertainty in data visualisation and communication

Ms Harriet Mason
Monash University, Melbourne, Australia

Abstract

Data visualisation is critical in understanding and communicating complex information so it is no surprise that graphs are ubiquitous in news, articles and many other avenues. Unfortunately, the graphs presented in mass media, as well as some journal articles, often hide the underlying uncertainty inherent in the statistics used to construct these graphics. Authors often opt to show the signal over the noise seeing one as a beacon of truth and the other as a nuisance. However, is uncertainty the pointless trouble it is made out to be? Should uncertainty itself not be seen as something worthy of being estimated and understood? Do we even understand uncertainty well enough to give a definitive answer?

Uncertainty is central to conveying the importance of a signal and incorporating risk into decision making, two important aspects of communicating statistics. When we quantify and visualise our uncertainty, it is difficult to do so without a clear definition of uncertainty and a clear idea of exactly what “uncertainty” we are trying to capture. Additionally, the visualisation should also change if we are treating uncertainty as something to be discovered in the data exploration stages, or something to be calculated and communicated in the final stages of a project. Due to these vague definitions and graphics designed without purpose, the current discussion of uncertainty methods is completely isolated from the ways in which uncertainty is visualised in our day to day lives.

In this work, we review the current state of uncertainty visualisation and clarify the language to describe different forms of uncertainty as well as discuss the reasons we visualise uncertainty at all. We illustrate this with examples

155

Tractable expectation propagation in non-standard settings

Mr Jackson Zhou, Associate Professor John Ormerod, Dr Clara Grazian
University of Sydney, Sydney, Australia

Abstract

Expectation propagation (EP) is an approximate Bayesian inference (ABI) approach which has seen increasing use across machine learning and statistics, owing to its high accuracy and speed. The main way in which the EP algorithm is made tractable is through the use of the affine transformation trick for multivariate Gaussians. Current applications focus on transformations to a univariate derived variable, which reduces the EP moment calculations to one-dimensional integrals. However, this limits the scope of feasible models. We show that more general versions of this trick can still make the algorithm tractable for some non-standard EP settings. The examples we cover include the Bayesian variants of heteroscedastic, lasso, and quantile regression models. EP is compared to other ABI methods across simulations and benchmark datasets, and is shown to offer a good balance between accuracy and speed.

156

Group Variable Selection via Unconstrained Continuous Optimization

Dr Sarat Moka ORCID iD[1,2], Mr Anant Mathur[1], Dr Zdravko Botev[1], Prof. Benoit Liquet-Weiland[2]

[1]The University of New South Wales, Sydney, Australia. [2]Macquarie University, Sydney, Australia

Abstract

In this research, we address the problem of selecting grouped variables in linear regression, with the goal of selecting a subset of groups that best fit the response variable. This type of group variable selection is particularly useful in applications such as genomics, image processing, and natural language processing, where the predictors naturally fall into groups. To solve this discrete constrained subset selection problem, we propose a new method that uses an unconstrained continuous optimization approach. We compare the effectiveness of our method with existing methods, including the popular Group Lasso, which is an extension of the Lasso regularization method. Our numerical results demonstrate the efficacy of our proposed method and its potential applications in various fields.

157

Robust Lasso Regression Using a Redescending Scale Estimate

Mr Alistair Martin ORCID iD, Dr Brenton Clarke ORCID iD
Murdoch University, Murdoch, Australia

Abstract

Regression model accuracy can be negatively impacted in high-dimensional settings, when many predictors are present. The adaptive Lasso is a widely adopted shrinkage model used to identify independent variables that are truly associated with the response, however it is highly sensitive to outliers and contaminated data. To address this issue, various robust Lasso formulations have been proposed in the literature, including those with redescending location M-Estimators. Typically, a robust scale estimate such as MADN is used, however none have utilised a redescending M-Estimator for scale. If the error distribution is indeed normal, then one ought assign zero weight beyond a reasonable threshold, so improvements in robustness may be realised when all model parameters are fit using redescending estimates. This study evaluates the robustness, model fit, and accuracy of the generalised robust adaptive lasso with redescenders. Both asymptotic and empirical evaluation is provided, under varying levels of contamination.

158

moostr: Enhancing Exam Creation in Moodle with Customisable Rewards, Shuffling, and Tolerance Levels

Dr Iris Jiang ORCID iD, Dr Thomas Fung ORCID iD
Macquarie University, Macquarie University, Australia

Abstract

Assessment is a crucial component of education as it provides valuable information on student progress, helps educators adjust their teaching methods to better meet student needs, and motivates students to learn and engage with the subject matter. In recent years, the use of automatically marked E-learning quizzes has gained popularity in education, and they offer several advantages over traditional assessments, such as immediate feedback to learners, consistency in marking, and more importantly they can be easily scaled to cater to a large number of students.

The exams package in R provides a framework for automatic generation of randomised exams/tests/quizzes, which is especially useful for large-scale assessments. The current implementation of the exams package only allows users to assign rewards to a single correct option/combination of options for single/multiple choice questions. For numerical entries, the package can either accept an exact match to a single correct answer or allow for a tolerance level as the distance from the single correct answer.

Despite the algorithm in exams package provides some level of flexibility when assigning rewards to correct answers, it does not support multiple correct answers or partially correct answers to be rewarded. moostr is a R package that expands the capabilities of the exams package and allows users to conveniently allocate rewards for multiple correct and/or partially correct answers. It allows users extra flexibility to shuffle options in selected single/multiple choice questions. In addition, for numerical questions, it also allows users to specify different tolerance levels for each correct/partially correct answers.

160

Estimating and visualising statistics of social information flows

Lewis Mitchell ORCID iD
The University of Adelaide, Adelaide, Australia

Abstract

Measuring and modelling online mis/disinformation spread is a pressing public concern for societies worldwide. Fundamentally, such phenomena are underpinned by information flows, which need to be estimated from the content of online social postings. While proxies for information flow inferred from metadata (shares, likes, timings) are simple to use and model, they do not make use of the full extent of content available: the text comprising postings themselves. This necessitates the use of information-theoretic estimators such as entropy and cross entropy, for which non-parametric estimators exist, but are poorly understand within an appropriate statistical framework.

This talk develops a statistical framework for estimating social information flows using non-parametric information-theoretic estimators, leveraging the text + timings of online social postings. We describe efficient estimation algorithms, as well as novel significance tests, which can be used to extract information flow networks from social media data. We discuss some of the challenges in visualisation of these large-scale derived social network datasets and showcase some examples where high-dimensional data can be visualised meaningfully.

We illustrate the estimation and visualisation techniques using examples from contemporary mis/disinformation campaigns on social media, including from the 2022-23 Russian war in Ukraine.

161

Estimation of the total loss in class actions

Professor Ian Gordon ORCID iD
Statistical Consulting Centre, University of Melbourne, Melbourne, VIC, Australia

Abstract

Representative legal proceedings, commonly known as “class actions”, are a prominent feature in Australian law. The Federal Court Act (1976) requires the Court “not to make an award of damages ... unless a reasonably accurate assessment can be made of the total amount to which group members will be entitled under the judgment”.

Statistical approaches can make an important contribution to addressing this requirement.

In this talk I will discuss issues in designing and analysing samples for the purpose of estimating the total loss, in several Australian class actions over the last 15 years or so. Statistical issues, including stratified sampling, proxy variables and missing data, will be discussed.

Effective communication of the process, for lawyers and judges, is of critical importance. The challenges and opportunities involved in presenting methods and results to a legal context will be described.

162

Estimation of Population Size with Heterogeneous Catchability and Behavioural Dependence: Applications to Air and Water Borne Disease Surveillance

Dr Prajamitra Bhuyan[1], Dr Kiranmoy Chatterjee[2]
[1]Indian Institute of Management, Calcutta, Kolkata, India. [2]Bidhannagar College, Kolkata, Kolkata, India

Abstract

Population size estimation based on the capture-recapture experiment is an interesting problem in various fields including epidemiology, criminology, demography, etc. In many real-life scenarios, there exists inherent heterogeneity among the individuals and dependency between capture and recapture attempts. A novel trivariate Bernoulli model is considered to incorporate these features, and the Bayesian estimation of the model parameters is suggested using data augmentation. Simulation results show robustness under model misspecification and the superiority of the performance of the proposed method over existing competitors. The method is applied to analyse real case studies on epidemiological surveillance. The results provide interesting insight on the heterogeneity and dependence involved in the capture-recapture mechanism. The methodology proposed can assist in effective decision-making and policy formulation.

163

Specifications Grading for Success in Statistics Education

Dr Wesley Burr ORCID iD
Trent University, Peterborough, Canada

Abstract

Specifications (or mastery-based) grading methods have become increasingly popular as assessment tools in education since the publication of Nilson's *Specifications Grading: Restoring Rigor, Motivating Students, and Saving Faculty Time* in 2015. This has extended to the mathematical, statistical and computational sciences, with numerous authors reporting success with assessments using this approach for laboratory sessions, calculus courses, and programming courses. Since 2019, I have implemented a series of variations on this theme across the statistics curriculum at Trent University (Ontario, Canada), which has resulted in significant perceived benefits on the part of students, and reported success in student learning outcomes and future knowledge retention.

In this talk, I will discuss two variations implemented in two separate undergraduate statistics courses: Mathematical Statistics, and Linear Models. The first, Mathematical Statistics, has evolved three times since first implementation, and in its current iteration consists of a "pure" mastery system, with a large (45) number of pass/fail deliverables, which break down the course into the smallest atomic pieces feasible; each of these deliverables can be resubmitted once. The second, Linear Models, is a simplified system more appropriate for senior students, with dynamically chosen problem sets which can be resubmitted as many times as desired, but with each individual problem within the set being set to pass/fail, and a "best of" grade summary used. In both systems, the marking quantity is significantly increased, but the complexity of each unit of marking is dramatically decreased, resulting in approximately equivalent workloads as compared to more classic systems of assessment. I will discuss the outcomes from these experiments, from both the student and faculty point of view, with some thoughts on the difficulty of determining success in a change in assessment such as this.

164

Planned but never published: analysing trends in clinical prediction models registered on clinicaltrials.gov

Dr Nicole White ORCID iD, Mr Rex Parsons, Dr David Borg, Prof Adrian Barnett
Queensland University of Technology, Brisbane, Australia

Abstract

Clinical prediction models combine patient-level data with statistical algorithms to estimate the risk of being diagnosed with or experiencing outcomes from a disease or health condition. The potential for clinical prediction models to support medical decision-making has made them an increasingly popular approach in health and medical research. Unfortunately, despite many new models being published, most are of insufficient quality to be used in everyday practice due to sources of bias introduced at the planning stage. Furthermore, the tendency to publish 'positive' findings means that of the models published, many more will have been planned but never completed. The analysis of study registries offers a unique opportunity to evaluate research before it is published, to determine current areas of research focus and recommend improvements in study design as early as possible to improve model quality.

In this talk, I will present findings from an analysis of clinical prediction models registered on clinicaltrials.gov since 2000. Our analysis had three objectives: to summarise current trends in planned diagnostic and prognostic modelling, to evaluate planned versus actual sample sizes as a common source of prediction modelling bias, and to estimate times to study completion and publication. Our search strategy found 928 studies aiming to predict diagnostic and/or prognostic outcomes. Sixty percent of studies proposed the development of new models without mention of internal or external validation. Less than half of the studies were matched to a subsequent publication within five years of registration.

Based on these findings, I will discuss potential solutions to improve the visibility of clinical prediction models to foster improvements in research quality.

165

Question Error, Students Correct: A solution to Error Carried Forward for online assessment

Dr Connor Smith
Macquarie University, Sydney, Australia

Abstract

Error carried forward has been a staple part of assessment grading in mathematics and statistics for a long time. Without these ECF marks, it is easy for students to apply the correct logic and not receive any marks. Through the deployment of major online assessments, we have observed that students will either accept their marks without question or question everything (and everything in-between). To help resolve this issue and provide correct marking for units with larger cohorts, the R package afteRexam is developed.

Initially designed for Moodle (with extensions for other LMS systems), the R package afteRexam is designed to be a bridge between no ECF marks and a fully integrated systems with a simple user interface. Although many methods exist for assessment design, the majority of question designers have a strong background in paper based exams and always employing perfect assessment design is a challenge. We do not address the different methods of online assessment design, but rather provide a solution to ensure that the students receive positive affirmation (marks) for correctly identifying the required question logic.

This solution was developed for large first year statistics cohorts to assist in the development of a strong statistical understanding from every student.

167

Accident Hotspot Road Segments in the Melbourne LGA vs Rest of Melbourne

Mr Mahamendige Asel Anthony Mendis ORCID iD
Swinburne University, Melbourne, Australia

Abstract

Accident hotspots have been traditionally modeled using Poisson regression and spatial clustering for accident count data. However, these models have failed to consider a broader range of variables, study areas encompassing metropolitan and regional characteristics, and use GIS for feature engineering. Accident Hotspot Road Segments (AHRS) are identified in this study using the Jenks natural breaks clustering method. These clusters are used to generate 2 groups (hotspots and non-hotspots) that will be considered as the dependent variable in binary logistic regression. This method is applied separately for Melbourne's Greater Capital City Statistical Area (GCCSA) (excluding Melbourne LGA) and the Melbourne LGA. Two separate models are built for these 2 regions as they exhibit very different land use and vehicular movement characteristics. These two regions require different explanatory variables for the modeling of AHRS. However, one common variable is the number of liquor license venues, increasing the odds of an AHRS (GCCSA Melbourne: 1.8 times and Melbourne LGA:8.6 times). Tram lines are another common variable, however, in the GCCSA Melbourne model, a tram line reduces the odds of an AHRS by 26% whereas, in the Melbourne LGA, it increases the odds of an AHRS by 200%. The GCCSA model has a Recall rate of 72% and an AUCROC of 77% [0.7526,0.7816] with 20 statistically significant independent variables. The Melbourne LGA model has a recall rate of 66% with an AUCROC of 79% [0.7214,0.8601] with 5 statistically significant independent variables.

168

Development of Australian Social Vulnerability Index

Tiernan Byrne, Kay Cao, Matt Beaty
Australian Bureau of Statistics, Canberra, Australia

Abstract

This presentation describes the development of an Australian Social Vulnerability Index (ASoVI). Vulnerability is one of the components of disaster risk and is defined in this context as 'the conditions determined by physical, social, economic and environmental factors or processes which increase the susceptibility of an individual, a community, assets or systems to the impacts of hazards'. For this research, the focus was on the social factors which affect the susceptibility of communities to the impacts of natural hazards.

There is a recognised need to develop a robust and up-to-date index to assess the social and economic vulnerability, resilience and capacity to recover of a community. The variables included in the ASoVI are based off those used in the U.S. SoVI, which is a component of the U.S. National Risk Index.

We used the Principal Components Analysis (PCA) method to develop the ASoVI. The index was calculated at the Statistical Area 2 level using the 2021 Census and other sources. We conducted a sensitivity analysis by comparing results from a single component to a multiple-component PCA. The results were also compared to the 2016 SEIFA Index of Relative Socio-Economic Advantage and Disadvantage (IRSAD). The index was also mapped to identify areas of high and low vulnerability for more detailed assessment by decision makers in crisis planning and management.

The results show a large amount of overlap between ASoVI and IRSAD reflecting their similarity in measuring social disadvantage, but ASoVI may be more suitable for emergency management applications. The results also highlight the contextual differences between Australia and the U.S. that need to be taken into account when developing indicators based on international best practice.

170

An Approximated Collapsed Variational Bayes Approach to Variable Selection in Linear Regression

Dr Chong You[1], A/Prof John Ormerod[2], Dr Xiangyang Li[1], Prof Cheng Heng Pang[3], Xiao-Hua Zhoua[1]
[1]Peking University, Beijing, China. [2]Sydney University, Sydney, Australia. [3]University of Nottingham, Ningbo, China

Abstract

In this work, we propose a novel approximated collapsed variational Bayes approach to model selection in linear regression. The approximated collapsed variational Bayes algorithm offers improvements over mean field variational Bayes by marginalizing over a subset of parameters and using mean field variational Bayes over the remaining parameters in an analogous fashion to collapsed Gibbs sampling. We have shown that the proposed algorithm, under typical regularity assumptions, (1) includes variables in the true underlying model at an exponential rate in the sample size, or (2) excludes the variables at least at the first order rate in the sample size if the variables are not in the true model. Simulation studies show that the performance of the proposed method is close to that of a particular Markov Chain Monte Carlo sampler and a path search based variational Bayes algorithm, but requires an order of
magnitude less time. The proposed method is also highly competitive with penalised methods, expectation propagation, stepwise AIC/BIC, BMS, and EMVS under various settings.

172

Regression analysis: “modelling the data” or answering a question?

John Carlin ORCID iD[1,2], Margarita Moreno-Betancur[1,2]
[1]Murdoch Children's Research Institute, Melbourne, Australia. [2]University of Melbourne, Melbourne, Australia

Abstract

Statisticians and users of statistics all too frequently fit regression models without a sufficiently clear purpose. The teaching of statistics encourages this practice, by focussing on the technicalities of models and estimation methods, leading to widespread belief in the “true model myth”, the idea that the essential purpose of regression analysis is to build the best possible model for the data, under the implicit assumption that a range of substantive conclusions may then be drawn from the fitted model. Unfortunately, it is clear on brief reflection that a regression model identified and fitted from finite data will very rarely if ever provide a full representation of the data generating process, particularly in complex areas such as health and social sciences. In fact, how regression models might be useful depends strongly on the purpose of the proposed analysis, so statistical analysts should clearly define the research question or purpose before starting to develop models. Hernan and others have emphasised a natural classification of the “tasks of data science” as addressing three types of research question: description, prediction, and causal inference. Regression methods can play a useful role in answering questions of all three types, but different considerations apply in each case for model development and interpretation. Unfortunately, the applied literature still abounds with papers that fit multivariable regression models for ill-defined purposes such as “identifying risk factors” or “exploring the effects of [a list of variables]”. We provide brief illustrations of the (mis)use of regression models and outline a new approach to the teaching and application of regression that emphasises the primacy of the three types of question. Students of statistics should learn that the aim is not to identify the best/correct model, but that once a clear purpose is defined, regression models may be useful, with important limitations.

173

Time Series Signal Detection using Polynomial Phase Demodulation

Benjamin Ott[1], Kian Blanchette[2], Dr Glen Takahara[1], Dr Wesley Burr ORCID iD[3]
[1]Queen's University, Kingston, Canada. [2]Statistics Canada, Ottawa, Canada. [3]Trent University, Peterborough, Canada

Abstract

Detection of periodic signals in the presence of noise within time series data is an important problem in many scientific fields such as geophysics, helioseismology and oceanography. Within the realm of harmonic analysis, Thomson's harmonic F-statistic is well understood to be a robust, effective test for this problem. However, this statistic loses effectiveness in the case of signals subject to frequency modulation, as well as when the signal-to-noise ratio is low. In his 2009 paper *Polynomial Phase Demodulation in Multitaper Analysis*, Thomson proposed a method for dealing with data with these problems.

In this talk, we will discuss two recent advancements on this problem. The first proposed test, which we will refer to as $\tilde{F_3}$, uses Slepian sequences as projection filters, complex demodulates the spectrum around a candidate frequency, and forms an instantaneous frequency series under the assumption of low-degree polynomial form. The resulting test statistic formed from this series is then asymptotically distributed $F_{1, K-P}$, where K is a function of the user-chosen time-bandwidth parameter, and P is the maximum degree of the polynomial modulation. The second, which we will refer to as F_4, addresses problems observed with $\tilde{F}_3$, and demonstrates that the choice of K has implications for the test statistic due to the even/odd nature of the Slepian sequences. A correction for this is also proposed. These methods are applied to data from the GOLF satellite, which presents time series containing signals which are known to experience modulation.

174

Emulation of greenhouse-gas sensitivities using variational autoencoders

Laura Cartwright ORCID iD, Dr Andrew Zammit-Mangion, Dr Nicholas M. Deutscher
University of Wollongong, Wollongong, Australia

Abstract

Flux inversion is the process by which sources and sinks of greenhouse gases are identified from observations of gas mole fraction. At the core of the flux inversion process is an atmospheric transport model, often a Lagrangian particle dispersion model (LPDM), that simulates particle trajectories in the atmosphere. These models are highly sophisticated, but are also highly computationally intensive. In related studies, researchers have proposed emulating the output of LPDMs at a small set of locations using a small subset of the LPDM inputs. However, existing approaches are limited to situations involving a single, known source of a specific atmospheric pollutant that could only have been emitted from that source. With greenhouse gases, the spatial sources and sinks constitute entire spatial fields. We thus need to emulate plumes over an entire spatial domain.

In this talk I will present a method to emulate the sensitivity plumes of methane mole fractions to a methane flux field. We use a convolutional variational autoencoder (CVAE) to reduce the dimension of the sensitivities (via the encoder segment), and fit a spatio-temporal Gaussian Process emulator on this reduced-dimensional space. Using the decoder segment of the CVAE, we obtain corresponding plumes from emulated variables. We then use our model to emulate a collection of sensitivities obtained using the Numerical Atmospheric Modelling Environment (NAME). For comparison, we repeat the study using empirical orthogonal functions (EOFs) for dimensionality reduction. Based on performance metrics computed on the emulated sensitivities, the CVAE performs marginally better than the EOFs. We conclude that our emulation-based approach can be used to considerably reduce the amount of computing time needed to generate LPDM outputs for use in high-resolution flux inversions.

175

Introducing predictive modelling in a large introductory statistics course

Dr Anna Fergusson ORCID iD, Mrs Emma Lehrke
University of Auckland, Auckland, New Zealand

Abstract

Teaching recommendations for implementing data science at the introductory tertiary level include placing greater emphasis on predictive modelling and immersing students in data-rich contexts by sourcing dynamic (“live”) data from the internet. Education research suggests that a different perspective is needed to teach predictive modelling than how simple linear regression has been traditionally introduced. More emphasis is needed on training and testing models with different data sets, validation through residual analysis and predictive accuracy, and concepts such as overfitting, underfitting, and generalisability. Hence, we designed and implemented an informal approach for introducing predictive modelling to students in our very large introductory statistics course (n ≈ 2200). In this paper, we will describe the learning progression, activities and data technologies used, and discuss our experiences with teaching and assessing this new approach with students.

176

Professional statistical training in agricultural science: promoting statistical confidence through communities of practice.

Sharon Nielsen[1], Sam Rogers[1], Annie Conway[1], Matthew Prescott[1], Isabel Munoz Santa[1,2]
[1]The University of Adelaide, Adelaide, Australia. [2]University of Valencia, Valencia, Spain

Abstract

Statistical rigour is key for the efficiency of agronomic research, where poor statistical design and analysis can lead to considerable wastage and financial loss. At the Biometry Hub, University of Adelaide, we have developed a statistical education workshop series aimed at researchers in agricultural and plant science, that promotes a full pipeline of statistical good practice; beginning from design of experiments, followed by appropriate statistical modelling and analysis, and finally to streamlined reporting and reproducibility. The workshops also introduce participants to the statistical software package R.

Following on from the training workshops is a broader focus on promoting ongoing statistical practice through the "Community of Practice." All workshop participants are invited to participate in the monthly community of practice meetings, where participants can bring their own questions to be worked through by the group. This gives workshop participants an incentive to continue to use R and reinforces and extends the workshop material.

As part of our teaching program, we have also developed the R package "biometryassist", which provides user-friendly access to statistical tools commonly required in agronomy and plant science. The package is used during the workshops and provides easy access to statistical functions for users who are not proficient in programming. Ongoing access to the package on CRAN allows participants to continue utilising the material after they have completed the workshops.

Through participant surveys we demonstrate how the workshop series has been beneficial in promoting confidence among participants, and how our ongoing support for participants raises the standard of statistics in the industry.

177

Renewal Epidemic Type Aftershock Sequence (ETAS) model for modelling the spatiotemporal distribution of earthquakes

Dr Tom Stindl ORCID iD
UNSW, Sydney, Australia

Abstract

The Epidemic Type Aftershock Sequence (ETAS) model is one of the most successful stochastic point process models for describing the distribution and clustering of earthquakes in both time and space. The ETAS model employs parametric forms based on empirical laws, such as the Omori-Utsu law and the Gutenberg Richter law, to describe the temporal and magnitude distribution of aftershocks. Because of this, the ETAS model has been very successful in modelling aftershock activity. However, an important aspect that has potential to improve the model, is in the modelling of the background seismicity (main shocks). ETAS models often assume that the background seismicity is constant in time and spatially varying. I will discuss a generalization of the ETAS model, which employs a renewal process to describe the main-shock arrival process, which serves as an alternative to the standard homogeneous Poisson process assumption. I will discuss methods for fitting the model, including direct likelihood maximization and an EM algorithm, and a declustering method to separate the main shocks from the aftershocks, which assists in estimating the spatial variation in background seismicity. An application of the proposed model and methods to a New Zealand earthquake catalogue is discussed, demonstrating its potential to improve the modelling of the spatiotemporal distribution of earthquakes and inferences drawn from the model.

178

Teaching Interdisciplinary Collaboration Across Cultures

Dr. Eric Vance ORCID iD[1], Professor Khairil Anwar Notodiputro[2]
[1]University of Colorado Boulder, Boulder, USA. [2]IPB University, Bogor, Indonesia

Abstract

Interdisciplinary collaboration is an integral part of the work of many statisticians and data scientists. Improving the effectiveness of collaborations can increase the impact of one's work. Are we, as statistics and data science practitioners and educators keeping up? Are we teaching collaboration? How does one teach collaboration skills? What are the essential collaboration skills we should teach to prepare our students for success in careers in statistics and data science? How do these skills vary based on culture?

This talk will describe the ASCCR framework for collaboration to help statisticians and data scientists learn and teach interdisciplinary collaboration skills. ASCCR stands for Attitude, Structure, Content, Communication, and Relationship. This framework was developed in the United States and is now being translated and adapted to teach Indonesian students how collaborate more effectively. We will discuss which components of collaboration are universal and which depend on the local culture.

179

A Gaussian Process Approach to Fitting Dynamic Network Models to Cross-Sectional and Sampled Data

Mr Christiopher Gordon[1], Dr Pavel Krivitsky ORCID iD[1], Dr Daniel Gladish ORCID iD[2]
[1]UNSW, Sydney, Australia. [2]CSIRO, Dutton Park, Australia

Abstract

Modelling and simulation of social networks, particularly networks that evolve over time, has manifold applications. A number of statistical models for network evolution has been developed, but the limiting factor is often availability of data. For example, for romantic networks, which are critical to understanding and predicting spread of sexually transmitted infections (STIs), it is important to accurately model not just the frequency of relationships but their timing. However, logistical and privacy considerations limit the kind of data that can be collected about them to cross-sectional "snapshots" of network features of interest (not networks themselves).

The Generalised Method of Moments Estimator (GMME) provides a way to estimate the parameters of the network evolution process based on these limited data, by tuning the model parameters until the simulated networks produced by this process in the long run have (on average) features similar to those observed. This is usually represented by specifying an objective function calculating a weighted squared difference between the observed and the simulated, which is then minimised. However, this simulation is extremely costly, the estimates of the objective functions are noisy, and the relationship between the parameters of the network evolution process and the features that can be observed can be complicated.

We propose to use Gaussian processes to emulate the GMME objective function and its components; Bayesian optimisation techniques are then used to quantify uncertainty in our estimates, and suggest subsequent guesses in order to minimise it and find the GMME.

This allows us to efficiently fit complex models for network evolution from limited data, even for large, difficult to observe networks. We demonstrate our approach on synthetic and real-world network data.

180

Compositional data and rapid implementation of score matching estimators

Dr Kassel Hingee ORCID iD
Australian National University, Canberra, Australia

Abstract

Score matching is an estimation technique that avoids normalising constants in model densities and can thus be used in many cases that maximum likelihood estimation cannot. Score matching estimation is sensitive to transformations of the data, which brings benefits and difficulties. For compositional data, work by Janice Scealy, Andrew Wood and John Kent has shown that some transforms allow estimation when zero-valued components occur frequently and other transformations appear to produce better estimates of correlation-like terms. Rapid implementation of score matching estimators enables faster empirical investigations of estimator behaviour. I have developed an R package for exactly this purpose using the automatic differentiation library CppAD. My package already includes numerous estimators that can be used on compositional data (and directional data too). Also available in my package is a general form of Windham robustification for exponential families.

181

Performance of interim analyses in a two-by-two factorial design with a time-to-event outcome: a simulation study of the VAPOR-C trial

Dr Anurika De Silva, Ms Meg Tully, Ms Yue Yang, Ms Sabine Braat
The University of Melbourne, Melbourne, Australia

Abstract

We can monitor accumulating data during a trial for stopping for early superiority or futility. These interim looks at the data are guided by statistical stopping rules that maintain overall Type I and Type II errors. While this topic has been well-documented for parallel group trials, there is limited guidance for factorial trial designs. The Volatile Anaesthesia and Perioperative Outcomes Related to Cancer (VAPOR-C) trial is an international multicentre randomised trial with a two-by-two factorial design and a time-to-event outcome. We plan to conduct a single interim analysis of the accumulated data using pre-defined stopping rules for superiority. We conducted a simulation study mimicking the VAPOR-C trial to examine the performance of conventional stopping rules under no, synergistic, or antagonistic interaction. If we observe a statistically significant interaction at interim, data will be analysed as multi-arm. If any treatment effects are statistically significant, the control arm will be dropped, followed by a final analysis as multi-arm, otherwise all arms will proceed at interim and an interaction test will be performed at final. If a statistically significant interaction is not observed at interim, data will be analysed as factorial. We drop the control arms based on the statistical significance of the marginal treatment effects, followed by a final two-arm analysis, otherwise all four arms will proceed at interim and an interaction test will be performed at final. We investigate the probability of a statistically significant interaction when testing for interaction between the main treatment effects at interim and final. Following the interaction test, we investigate the probability that the trial is analysed as factorial only, multi-arm only, or a combination of factorial and multi-arm across interim and final analyses, and the Type I error rates and Powers when testing for treatment effects using the stopping rules. Simulation results will be presented.

182

The Extraordinary Potential of Statistical Collaboration Laboratories

Dr. Eric Vance ORCID iD[1], Professor Khairil Anwar Notodiputro[2]
[1]University of Colorado Boulder, Boulder, USA. [2]IPB University, Bogor, Indonesia

Abstract

This talk will explore the powerful concept of the statistical collaboration laboratory (or “stat lab”), whose main idea is to simultaneously educate and train students in statistical consulting and collaboration by offering them the opportunity to collaborate with researchers and decision-makers to apply statistics and data science for the benefit of society. I will discuss the components of a stat lab, the models I have used to create new stat labs, and the model of the new lab currently being developed at IPB University in Indonesia.

I will also discuss the LISA 2020 Global Network, which is a network of 35 recently established stat labs in developing countries. These stat labs train local statisticians and data scientists to collaborate with local researchers and policy makers to transform evidence into action for sustainable development. Since 2019, our labs have trained more than 2968 data scientists to collaborate with more than 2672 domain experts to discover local solutions for local challenges and have taught 499 short courses and workshops to improve the statistical skills of 17670 attendees.

183

Multidimensional versus univariate sequential analysis

Dr Alexander Ek ORCID iD[1,2], Floyd Everest ORCID iD[3], Dr Michelle Blom ORCID iD[3,2], Prof. Ronald L. Rivest ORCID iD[4], Prof. Philip B. Stark ORCID iD[5], Prof. Peter J. Stuckey ORCID iD[1,2], A/Prof. Vanessa J. Teague ORCID iD[6,7], A/Prof. Damjan Vukcevic ORCID iD[1,2]
[1]Monash University, Melbourne, Australia. [2]ARC Training Centre in Optimisation Technologies, Integrated Methodologies, and Applications (OPTIMA), Melbourne, Australia. [3]University of Melbourne, Parkville, Australia. [4]Massachusetts Institute of Technology, Boston, USA. [5]University of California, Berkeley, USA. [6]Thinking Cybersecurity Pty. Ltd., Melbourne, Australia. [7]Australian National University, Canberra, Australia

Abstract

Sequential analysis methods carry out inference using a sample size that is not fixed in advance. Typically, the goal is to minimise the number of observations while ensuring sufficient evidence for a conclusive result, for example sufficient power and control of the false positive rate in the context of a hypothesis test. The methods protect against “stopping when the data look good”, providing rigorous control of error rates despite allowing the practitioner the freedom to choose when to stop collecting data. Our motivating application is statistical auditing of preferential elections. This requires inference in a very high-dimensional parameter space using a sample of ballot papers.

Many recent developments in sequential testing have focused on univariate settings, leveraging martingale theory to derive efficient and widely applicable sequential tests. To make use of these for election auditing, a successful strategy for many elections has been to take projections of the parameter space into several, single-dimensional spaces, in which the aforementioned tests can be applied. The test results are then combined together to form an overall result.

A natural question is whether anything is lost (or gained) by taking such projections, as compared to inference approaches that operate directly in the whole parameter space. Is there any loss or gain of statistical efficiency? How does computational efficiency compare?

We explore these questions both analytically and empirically, comparing different modes of inference (both Bayesian and frequentist) in the context of simple and complex elections.

184

Optimal staircase cluster randomised trial designs and when to use them

Dr Kelsey Grantham[1], Prof Andrew Forbes[1], Prof Richard Hooper[2], A/Prof Jessica Kasza[1]
[1]Monash University, Melbourne, Australia. [2]Queen Mary University of London, London, United Kingdom

Abstract

The stepped wedge design is a useful tool for testing interventions applied at the cluster level that cannot be removed once implemented. However, these designs require clusters to contribute measurements for the entire trial duration. Staircase designs are emerging as a less burdensome alternative to the stepped wedge. Visually, the trial design resembles a staircase: clusters are randomly assigned to sequences made up of a limited number of measurement periods (control periods followed by intervention), where sequences start measurement at different times. Recent work has found the basic staircase design, which has just one control period followed by one intervention period in each sequence, to be a particularly lean design with power that can rival that of the stepped wedge in certain situations. However, to meet or exceed the efficiency of a stepped wedge, a basic staircase design typically requires clusters to measure more participants in each period, or include additional clusters. In this talk we aim to find optimal staircase designs that could more closely rival the stepped wedge.

We will identify optimal staircase designs by considering the variance of the treatment effect estimator, a key component of power calculations, using a linear mixed model under different trial settings. We will examine whether there is a benefit to having different numbers of control and intervention periods in a sequence, moving beyond the basic staircase design. We will also determine the optimal allocation of clusters to sequences for different trial configurations. Surprisingly, our results show that imbalanced designs, where clusters do not contribute equal numbers of measurements in control and intervention conditions, can be optimal, for certain common trial configurations and modelling assumptions. Furthermore, the optimal allocation of clusters to sequences in a staircase design does not assign the same number of clusters to each sequence.

185

Sequential testing using adaptively weighted martingales, for auditing preferential elections

Dr. Alexander Ek ORCID iD[1,2], Dr. Michelle Blom ORCID iD[3,2], Prof. Ronald L. Rivest[4], Prof. Philip B. Stark ORCID iD[5], Prof. Peter J. Stuckey ORCID iD[6], A/Prof. Vanessa J. Teague ORCID iD[7,8], A/Prof. Damjan Vukcevic ORCID iD[1,2]

[1]Department of Econometrics and Business Statistics, Monash University, Melbourne, Australia. [2]ARC Training Centre in Optimisation Technologies, Integrated Methodologies, and Applications (OPTIMA), Melbourne, Australia. [3]School of Computing and Information Systems, University of Melbourne, Melbourne, Australia. [4]Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Boston, USA. [5]Department of Statistics, University of California, Berkeley, Berkeley, USA. [6]Department of Data Science and AI, Monash University, Melbourne, Australia. [7]Thinking Cybersecurity Pty. Ltd., Melbourne, Australia. [8]Australian National University, Canberra, Australia

Abstract

Counting the votes in an election can be complex and prone to errors and malicious interference, especially when technology and automation are thrown into the mix. Statistical audits that manually interpret the votes on paper ballots have been introduced as a way to provide confidence in the reported election results; some jurisdictions around the world require such audits by law. These are usually designed as sequential procedures: using relatively few ballots if the sample clearly supports the reported election outcome, and otherwise requiring that more ballots be sampled.

For simple voting systems, such as ‘first-past-the-post’ and scoring rules, methods to carry out such audits have been well developed. In contrast, some preferential voting systems, including those used in Australian elections, involve complex counting algorithms that make them challenging to audit efficiently. The best current methods for those systems either do not provide a convenient way to control the chance that an incorrect outcome will fail to be corrected, or require a digitised version of all voter preferences (‘cast vote records’; CVRs). We develop a method that avoids both issues.

In common with some existing methods, our approach projects the full parameter space into many single-dimensional spaces, then conducts statistical tests in these spaces, which are combined into an overall verdict about the correctness of the outcome. For preferential voting, many such projections are needed. Existing methods use CVRs to select an economical subset of projections to test, on the assumption that the CVRs have relatively few errors. We develop a test using adaptively weighted nonnegative martingales that finds an economical set of projections directly from the sampled ballots, while rigorously controlling the chance of failing to correct wrong electoral outcomes. This allows our method to be more widely applicable, and robust to scenarios where CVRs are contaminated with substantial errors.

186

Strategic model reduction by analysing model sloppiness: a case study in coral calcification

Ms Sarah Vollert ORCID iD[1], Dr Christopher Drovandi ORCID iD[1], Dr Gloria Monsalve-Bravo ORCID iD[2], Dr Matthew Adams ORCID iD[1]

[1]Queensland University of Technology, Brisbane, Australia. [2]University of Queensland, Brisbane, Australia

Abstract

It can be difficult to identify ways to reduce the complexity of large models whilst maintaining predictive power, particularly where there are hidden parameter interdependencies. Here, we demonstrate that the analysis of model sloppiness can be a new invaluable tool for strategically simplifying complex models. Such an analysis identifies parameter combinations which strongly and/or weakly inform model behaviours, yet the approach has not previously been used to inform model reduction. Using a case study on a coral calcification model calibrated to experimental data, we show how the analysis of model sloppiness can strategically inform model simplifications which maintain predictive power. Additionally, when comparing various approaches to analysing sloppiness, we find that Bayesian methods can be advantageous when unambiguous identification of the best–fit model parameters is a challenge for standard optimisation procedures.

187

Between-centre differences in trial treatment effects in multicentre anaesthesia trials

Vanessa Pac Soo[1], Sabine Braat[1], Anurika De Silva[1], Philip Peyton[2]
[1]University of Melbourne, Melbourne, Australia. [2]Austin Health, Melbourne, Australia

Abstract submission

Download file

188

Creating flexible e-learning quizzes with literate programming

Mr Mitchell O'Hara-Wild ORCID iD[1], Dr Emi Tanaka ORCID iD[2,1], Dr Thomas Fung ORCID iD[3], Dr Iris Jiang ORCID iD[3]

[1]Monash University, Melbourne, Australia. [2]Australian National University, Canberra, Australia. [3]Macquarie University, Sydney, Australia

Abstract

E-learning quizzes with automated marking systems allow educators to assess students' abilities, motivate learning, give automatic and instant feedback, and are scalable to a large number of students. Many quiz questions prevalent in teaching statistics use numerical and single/multiple choice responses. These types of questions can be algorithmically generated to create multiple versions of the question with different instances of randomised tasks, datasets and/or inputs. In practice, quiz creators must conform to a constrained and/or rigid structure due to the limitations of the deployment platform, which may not align with the natural order of thinking.

In this talk, I provide a higher-order overview of creating algorithmic e-learning quizzes and propose a set of literate programming design principles for a robust, yet flexible system to create these quizzes. The design of this system is implemented in the moodlequiz R package for the free and open-source learning management platform Moodle. The core set of literate programming design principles can be extended to other learning management platforms.

189

Model selection for high dimensional genomic data using hypercubes

Aline Kunnel[1,2], Dr Simon Tuke[1], A/Prof Gary Glonek[1]
[1]The University of Adelaide, South Australia, Australia. [2]South Australian Health and Medical Research Institute, South Australia, Australia

Abstract

Gene expression data consists of a large number of genes and a small number of observations. Analysis of this type of data is quickly met with issues due to high dimensionality. This project uses a recently introduced method by Cox & Battey motivated by a hypercube geometry. This comprehensive method allows to select and assess sets of combinations of predictors and dramatically reduces the time needed. We extended the hypercube approach further to allow for strong correlation structures present in gene expression data. Results find it allows the identification of potentially relevant genes to a greater extent than the LASSO. We compare the predictive power of the extended hypercube approach and regularization methods to real-world data.

190

Deep Learning for the Repair of Administrative Data

Mr Sean Buttsworth
Australian Bureau of Statistics, Canberra, Australia

Abstract

Administrative data can provide fine level statistics not possible with survey data. However these statistics may contain biases because of coverage or conceptual errors. If a representative survey exists, collecting the same data items as found on the administrative dataset, we can adjust the administrative data to better align with the population. A method has been developed for doing this, in collaboration with QUT, based on transformer neural networks and transfer learning. This approach does not require the survey and admin datasets to be linked. We evaluate the method on real-world economic data in the ABS Datalab.

191

Generalized Partial Least Square in Deep Neural Network

Miss Qian Jin[1,2], Dr Clara Grazian[3,2], A.Prof Pierre Lafaye de Micheaux[1,4]
[1]UNSW, Sydney, Australia. [2]DARE, Sydney, Australia. [3]University of Sydney, Sydney, Australia. [4]Universite Paul Valery Montpellier 3, Montpellier, France

Abstract

While deep learning has shown exceptional performance in many applications, the model's mathematical understanding, model designing, and model interpretation are still new areas. Combining the two cultures of deep and statistical learning can provide insight into model interpretation and high-dimensional data applications. This work focuses on combing deep learning with generalized partial least square estimation. In particular, Bilodeau et al. (2015) proposed a generalized regression with orthogonal components (GROC) model, which is more flexible than the standard partial least square (PLS), because it may involve more complex structure of dependence and the use of generalized additive model (GAM) instead of linear regression. We propose a deep-GROC (DGROC) model, which allows for different measures of dependence (through their copula representation) to be used and shows a high prediction accuracy. Hyperparameter selection and transfer learning in the training loop are included to boost model performance. The superiority of the proposed method is demonstrated on simulations and real datasets, which show that our method achieves competitive performance compared to GROC, PLS and traditional Neural Networks.

192

Comparison of estimation methods for two-stage sampling with differential non-response

Ms Victoria Leaver
University of Wollongong, Wollongong, Australia

Abstract

Various weighting methods can be used to address differential non-response in sample surveys. However, the estimation process may become inefficient if the weights are highly variable, which can occur in practice when response rates within some subcategories are very low. To explore the performance of selected estimation methods, including weight smoothing and some likelihood methods, a synthetic population has been created from public use sample data. A simulation study using two-stage samples with varying rates of non-ignorable non-response will be used to compare the efficiency of the estimation methods under different non-response scenarios. The comparison will use the results from weighted linear regression where the outcome variable is related to the pattern of non-response.

193

Directed weighted network reconstruction using random graph ensembles

Mr Paul Fortuin
Australian Bureau of Statistics, Sydney, Australia

Abstract

Network reconstruction methods are being used increasingly in the analysis of financial and trade networks. The advantages of using probabilistic methods for weighted network reconstruction is explored in cases where there is limited information available. Focus is placed on the methods of Parisi, Squartini and Garlaschelli, who generate a graph ensemble of Bernoulli distributed edges and exponentially distributed weights to model networks. The quantification of model uncertainty and calculation of confidence intervals is also explored. Weighted Network Reconstruction has applications in modelling internet traffic, transport networks, banking, debt, trade networks and food distribution. This talk will focus on directed weighted network reconstruction with application to business supply chain networks.

194

Data literacy skills for transition into tertiary education: An Indonesian context

Ms Charanjit Kaur ORCID iD, Dr Joan Tan, Dr Ririn Yuniasih
Monash University, Melbourne, Australia

Abstract

Given the prevalence of data, it is crucial for Secondary students to develop data literacy skills to prepare them for the transition to tertiary education and employment. In the current context of Indonesian Secondary education, there needs to be more emphasis on practical applications and contextual understanding. This has created a gap in the learning skills required when students progress to higher education. To address this gap, we designed online boot camps for Secondary students from Indonesia with the aim of enhancing their statistical thinking and conceptual understanding. This pilot intervention is aligned with the Indonesian Education initiative, aimed at the early integration of data literacy skills in the high school curriculum. In this paper, we analyse students’ perceptions about the skills learnt through student surveys. Our findings show that students' understanding of statistical concepts and their application increased post-boot camp, with no gender differences. However, there was a significant difference between students from the Science and Math and Social Sciences disciplines. These findings provide policy recommendations that can be used as guidelines for integrating data literacy into the secondary education curriculum.

195

Ordered logit, multinomial logit or partial proportional odds model? Demographic risk factors of patients with low Vitamin C levels between 2017 – 2021 in NSW, Australia

Dr Kathrin Schemann ORCID iD[1], Dr Puja Bhattacharyya[2,3], Dr San San Min[4], Associate Professor David Sullivan[5,6], Associate Professor Stephen Fuller ORCID iD[7]
[1]Sydney Informatics Hub, Core Research Facilities, The University of Sydney, Sydney, Australia. [2]WSLHD Research and Education Network- Blacktown Hospital Sydney, Sydney, Australia. [3]The University of Sydney, Sydney, Australia. [4]NSW Health Pathology, Sydney, Australia. [5]Department of Clinical Biochemistry, Royal Prince Alfred Hospital, Sydney, Australia. [6]The University of Sydney Central Clinical School, Faculty of Medicine and Health, Sydney, Australia. [7]The University of Sydney Nepean Clinical School, Faculty of Medicine and Health, Sydney, Australia

Abstract

Low levels of vitamin C may cause general unwellness, nausea, reduced appetite, and impaired wound healing. Extremely low levels of vitamin C can cause scurvy, a disease that causes anaemia, gum disease and skin haemorrhages. Scurvy was once associated with long-distance sea travel; however recent case reports suggest it is now resurfacing in Australia.

This retrospective study is the largest Australian analysis of risk factors for low vitamin C using blood samples processed over a 5-year period at a central laboratory in Sydney. The main outcome in univariable and multivariable models was vitamin C status measured by integer concentration and categorised into normal (40-250 µmol/L), hypovitaminosis C (12-39 µmol/L) and significant deficiency (≤11 µmol/L).

Of 12,934 unique patients tested, more than half had abnormal levels of Vitamin C: 24.5% had significant deficiency and 29.9% had hypovitaminosis C. With the proportional odds assumption violated by some predictor variables, a partial proportional odds model was considered; yet a significant interaction term necessitated a final multinomial approach for an identifiable model. Males, older patients, and patients from remote or socio-economically disadvantaged areas had higher odds of hypovitaminosis C or a significant deficiency. In the final multivariable model study year and age were interacting and remoteness was excluded.

In addition to known risk factors, this study identifies remoteness as a risk, however this may be influenced by vitamin C degradation due to inadequate, pre-analytical sample transport. The findings show that vitamin C deficiency cases can be hidden in communities that are considered advantaged within pockets of socio-economic disadvantage. The presence of vitamin C deficiency in the elderly, males and people from disadvantaged and remote areas indicate a significant public health issue that requires a population-wide assessment of vitamin C levels.

196

Mental health care needs and access among Australian men: Data from the Australian Longitudinal Study on Male Health

Dr Clement Wong, Karlee O'Donnell, Dr Jennifer Prattley, Dr Rebecca Jenkinson, Dr Katrina Scurrah, Dr Sean Martin
Australian Institute of Family Studies, Melbourne, Australia

Abstract

Background

Men’s mental health remains a significant public health concern and a priority issue for Australia. Australian men tend to have lower usage of health care in general compared to women, despite comparable health needs. This study examines Australian men’s reported depressive symptoms against mental health treatment usage (and non-usage) between 2013 and 2021.

Method

Data was drawn from the first three waves of Ten to Men, a national longitudinal survey of Australian males (n=8,887; 2013/14-2020/21). Depressive symptoms were assessed using a 9-item validated measure (Patient Health Questionnaire; PHQ-9). Medicare MBS and PBS linkages were obtained from participants inclusive from 2012-2021 to assess usage of mental health services (by defined MBS item numbers) and prescriptions (by defined ATC classification codes). A control set of selected variables on socio-economic, demographic, masculinity norms, financial and disability status were used in multi-adjusted linear regression models of use of mental health care by PHQ-9 category.

Results

Overall, approximately 13% of men reported moderate to severe depressive symptoms at each survey wave across 2013—2020. Over 30% of Australian men accessed a mental health service or prescription medication between 2012—2021. Mental health care usage increased during this period, primarily in GP visits for mental health reasons and prescriptions of antidepressants. Among men with greater depressive symptoms, older age, identifying as an Aboriginal & Torres Strait Islander, unemployment, and not being in the labour force were associated with higher usage of mental health care, while conformity with masculine norms was associated with lower usage.

Conclusion

Mental health treatment is becoming more widely accessed among Australian men, while self-reported depressive symptoms remain steady. Despite this, many men with greater depressive symptoms did not subsequently access mental health treatment, suggesting that unmet need and barriers to health care remain.

197

Introducing novices to posing and answering investigative questions about categorical data.

Mrs Malia Puloka, Honorary Associate Professor Maxine Pfannkuch, Associate Professor Stephanie Budgett
University of Auckland, Auckland, New Zealand

Abstract

Posing and answering investigative questions is an important aspect of statistical enquiry. Little research has been done on posing and answering questions about categorical data involving two variables. In an exploratory study with 15 students in a mixed-ability class in a low socio-economic school, a teaching intervention was conducted. In this paper the first teaching episode is described. The posing and answering of student-posed questions revealed that while students were able to answer simple proportion questions, they had difficulty determining denominators when answering questions involving conditional and joint proportions. Delving into novices' thinking processes when answering questions could help teachers to design learning strategies to assist students deconstruct questions about two categorical variables.

198

Modelling COVID and crime in the US as hierarchical time series

Dr Thomas Fung ORCID iD[1], Dr Joanna Wang[2]
[1]Macquarie University, Sydney, Australia. [2]University of Technology, Sydney, Sydney, Australia

Abstract

Crime time series data can often be naturally disaggregated by various attributes of interest, either by their crime type or geographical location. When modelling this type of data, the current recommended practice in crime science is to model each series at the most disaggregated level as it helps to identify more subtle changes. However, authorities and stakeholders are often only interested in the big picture, which requires researchers to either simply summing the fitted value series up or model the aggregated series independently. This leads to poor forecasting performance at the higher levels of aggregation in practice, as the most disaggregated series often have a high degree of volatility, while the most aggregated time series is usually smooth and less noisy. Intuition also requires the forecasts to add up the same way as the data, but one can’t guarantee that would be true when series are modelled independently. In this talk, we will explain why the hierarchical and grouped time series method of Wickramasuriga et al. (2019) should be considered as the default technique for modelling this kind of data. US COVID and crime data of Abrams (2020) will be used as an example.

200

STEPHEN: A Hidden semi-Markov Model for Estimating Sedentary Behaviour Using Wrist-worn Accelerometers

Dr Agus Salim ORCID iD
University of Melbourne, Melbourne, Australia. Baker Heart and Diabetes Institute, Melbourne, Australia

Abstract

Sedentary behaviour (SB) has been linked to increased risk of cardiovascular diseases. Until recently, thigh-worn accelerometers have been used to measure SB. But due to the burden they placed on participants, there has been increasing interest in using wrist-worn accelerometers for long-term monitoring of SB. Challenges for detecting SB using wrist-worn accelerometers includes the detection of body posture, since the accelerometer is placed far away from the hip/waist area, differentiating standing from sitting can be problematic. Wrist-worn devices are known to produce anomalously large step counts for activities that do not involve stepping. To address these challenges, we developed a Hidden semi-Markov model (HSMM) called STEPHEN (STEP and HEart Rate ENcoder) that used routine step counts and heart rate data from a consumer-grade wrist-worn accelerometer to classify SB. To differentiate standing from sitting postures, the model takes advantage of the fact that heart rate increases immediately when standing from sitting/lying position. To minimize the effect of anomalous large steps, the model assumes that physical activities will lead to increases in heart rate, hence any increases in step counts without significant increase in heart rate is more likely due to SB. STEPHEN assumes there are four hidden states and under each state, we assume that the step count and heart rate data are distributed as Negative Binomial (NB) random variables with state-specific parameters. An individual is allowed to move between any two states, e.g., from sedentary behavior state to light physical activity with the amount of time (minutes) spent consecutively under a particular state (sojourn) following a Gamma distribution. We compared STEPHEN's performance to proprietary software and validated its applicability using Fitbit data from two different cohorts with different demographic characteristics and living conditions.

201

Highway to the “Danger Zone”: A practical framework for evaluating early-career collaborative opportunities

Dr Taya Collyer ORCID iD
Monash University, Melbourne, Australia. National Centre for Healthy Ageing, Frankston, Australia

Abstract

The literature and the internet abound with ‘advice’ for early-career statisticians wishing to engaging in collaborative projects. However, these lists of do’s and don’ts mainly state the obvious and appear systematically deaf toward the complex interpersonal dynamics of such projects. In this talk I share a framework for the evaluation of collaborative opportunities, which I have developed and refined via a large number of mistakes over my short career. I begin by presenting four questions which junior analysts should ask prior to joining a collaborative project: i) What are you doing (is this project business as usual, a favour to someone)? ii) Why are you doing it? iii) What is the strategic importance of the outcome (to you, to your employer, and to your collaborators)? and iv) Is it possible for you to withdraw your involvement, should you want to? Next I present a 2x2 matrix which I have found extremely helpful for understanding the strategic importance of opportunities which come my way, with special focus on what I term the ‘Danger Zone’ of collaborative opportunity. In Part 2 I discuss challenges which should be anticipated, including technical challenges and interpersonal challenges such as scope-creep. I conclude with 4 practical tips for overcoming these challenges, and presentation of my Diplomatic Communication Toolkit.

202

Changes in depression symptoms in a nationwide, representative prospective cohort of Australian men before, during and after COVID-19

Dr Katrina Scurrah ORCID iD, Dr Clement Wong, Dr Jennifer Prattley, Melissa Suares, Leanne Howell, Jennifer Renda, Kipling Walker, Karen Biddiscombe, Sarah Carr, Dr Sean Martin
Australian Institute of Family Studies, Melbourne, Australia

Abstract

Background: Mental ill health including depression is common in Australia. Previous studies have reported that severity of depression increased during the COVID-19 pandemic, especially for men. However, many of these estimates were derived from smaller, cross-sectional samples. This has limited estimates on whether depression symptoms were altered overall for men and for which specific population sub-groups as a result of COVID-19.

Methods: Data from *Ten to Men*: The Australian Longitudinal Study on Male Health, was used to investigate changes in depression symptoms in Australian men over time. Four waves of data are available, obtained in 2013-2014, 2015-2016, 2020-2021, and 2022. Depressive symptoms were assessed using a validated measure (Patient Health Questionnaire; PHQ-9). Effects were estimated using linear mixed effects models with adjustment for baseline age. The main covariate was time (modelled as a categorical exposure), and random intercepts were included at the participant level.

Results: PHQ9 scores were available for all 4 waves for over 5100 adult men. Preliminary results suggest that compared with 2015-2016, average PHQ9 increased slightly during COVID-19 (β=0.26, 95% CI 0.15 to 0.37) and remained higher post- COVID-19 (β=0.23,95% CI 0.12 to 0.34). Larger increases were apparent for younger participants ($\leq$ 25 years at Wave 1) during and post- COVID-19 (β=0.91, 95% CI 0.51 to 1.30 and β=0.80, 95% CI 0.41 to 1.19 respectively). Observed individual changes in PHQ-9 varied from between -23 and 24, with 10% of participants changing by $\geq$ 10.

Conclusion: Overall, a small but significant increase in depression symptoms was observed in Australian men around the COVID-19 period, especially among younger men. Large proportions of men were observed to have minimum clinically important differences in depression symptoms. Future work will characterise these participants in order to inform targeted policies for ongoing support to improve mental health.

203

Stochastic variational inference for heteroskedastic time series models

Mr Hanwen Xuan[1], Dr Feng Chen[1], Dr Clara Grazian[2], Dr Luca Maestrini[3]
[1]University of New South Wales, Sydney, Australia. [2]University of Sydney, Sydney, Australia. [3]Australian National University, Sydney, Australia

Abstract

Stochastic variational inference algorithms are derived for fitting various heteroskedastic time series models using Gaussian approximating densities. Gaussian, t and skew-t response GARCH models are examined. We implement an efficient stochastic gradient ascent approach based upon the use of control variates or the reparameterization trick and show that the proposed approach offers a fast and accurate alternative to Markov chain Monte Carlo sampling. We also present a sequential updating implementation of our variational algorithms, which is suitable for the construction of an efficient portfolio optimization strategy.

204

A Latent Variable Approach to Bayesian Inference for Self-Inhibited Hawkes Processes

Mr Arya Karami ORCID iD[1,2], Dr Pavel Krivitsky ORCID iD[1], Dr Xuhui Fan ORCID iD[3], Prof Scott Sisson ORCID iD[1]

[1]UNSW, Sydney, Australia. [2]Sharif University, Tehran, Iran, Islamic Republic of. [3]University of Newcastle, Newcastle, Australia

Abstract

Hawkes models for temporal point processes represents influence of events on the rate of future events by specifying a *kernel* function that controls how an event either increases (“excites”) the rate of the events that follow or decreases (“inhibits”) it. They have been applied to neurons firing in neuroscience, market events data in finance, incidence of infections in epidemiology, and messages sent in social networks.

In Bayesian inference for self-exciting Hawkes processes, it is common to use a latent variable for each observed event, indicating which prior event, if any, had “excited” it. This greatly simplifies the likelihood and, in particular, its compensator (the probability of no events occurring between the observed events), often enabling a conjugate prior for the background rate parameters.

Self-inhibited Hawkes processes present their own computational challenges, which have received less attention. For one, effects of inhibition can make the compensator intractable. For another, it is in principle possible for a series of events to inhibit the rate of the process to 0, complicating the form of the likelihood. This is commonly addressed by transforming the predicted rate of the process away from 0, but this can affect the model's interpretation.

We introduce a latent variable approach to self-inhibited Hawkes processes that performs a similar function to that in the self-exciting process: it separates the compensator from the inhibition effects for the purposes MCMC sampling. In addition, for some forms of inhibition kernels, we provide a likelihood calculation algorithm linear in the number of observations. This facilitates scalable Bayesian inference for Hawkes processes, making transformations unnecessary.

We illustrate our approach on univariate and multivariate synthetic and real-world event data.

205

A Linear Errors-in-Variables Model with Unknown Heteroscedastic Measurement Errors

Dr Linh Nghiem[1], Dr Cornelis Potgieter[2]
[1]University of Sydney, Syndey, Australia. [2]Texas Christian University, Fort Worth, USA

Abstract

In the classic errors-in-variables model, covariates are contaminated by independent additive measurement errors. We consider parameter estimation in a linear errors-in-variables model where the unknown measurement error distributions are symmetric and heteroscedastic across observations. We propose a new nonparametric estimation method that corrects measurement error by combining two different techniques: a moment correction approach and a phase function-based approach. The former method requires underlying distributions to have four finite moments, while the latter relies on covariates having asymmetric distributions. These methods are combined using a generalized method-of-moments approach. The new estimator is shown to be consistent and asymptotically normal under appropriate regularity conditions. Furthermore, the new estimator has a strong finite sample performance in simulation studies, especially when the measurement errors follow non-Gaussian distributions.

206

Combining online survey data and area-level external data to create local estimates of social cohesion in Australia

Mr Sam Slamowicz[1], Mr Andrew Ward[1], Dr Farzana Jahan ORCID iD[2]
[1]Social Research Centre, Melbourne, Australia. [2]Murdoch University, Perth, Australia

Abstract

Social cohesion is the sense of connectedness and unity within communities and is seen as a key influence on individual well-being and on social, economic and political stability. Surveys that monitor social cohesion have become common tools for researchers and policy-makers, and there is increasing need for measures at finer levels of detail than what surveys alone can provide. In this work responses from a national probability online survey of Australian adults were combined with census and other auxiliary data to generate estimates of attitudes to social issues and immigration at the local government area (LGA) level. Two methods were employed - the first was a basic unit-level model using the online survey responses and the second was a Bayesian spatial meta-analysis built on top of estimates from the first model and incorporating the spatial association of measures between neighbouring areas. The study compared the estimates of scores with and without spatial smoothing in terms of validity, meaningfulness, and precision, with a view to recommending methods for future waves of the survey

207

A novel decision-making framework for Bayesian adaptive trial designs

Mr Michael Dymock ORCID iD[1], Professor Tom Snelling ORCID iD[2], Dr Charlie McLeod ORCID iD[1,3,4], Professor Peter Richmond ORCID iD[1,3,5,6], Dr Julie Marsh ORCID iD[1]

[1]Wesfarmers Centre of Vaccines and Infectious Diseases, Telethon Kids Institute, Perth, Australia. [2]School of Public Health, University of Sydney, Sydney, Australia. [3]Centre for Child Health Research, The University of Western Australia, Perth, Australia. [4]Infectious Diseases Department, Perth Children's Hospital, Perth, Australia. [5]School of Medicine, University of Western Australia, Perth, Australia. [6]General Paediatrics and Immunology Departments, Perth Children’s Hospital, Perth, Australia

Abstract

COVID-19 has sparked an increase in the implementation of adaptive trial designs to evaluate the efficacy of treatments and vaccines. In comparison to fixed designs, adaptive designs aim to maintain power and type one error with regular scheduled analyses, potentially requiring fewer participants and shorter timelines. Pre-specified decision rules allow the trial design to evolve as data accumulates. The decision rules are traditionally based on direct comparisons between statistical quantities (e.g., means, log odds) representing intervention effects. For example, one might stop the trial early and conclude treatment superiority if the probability that the treatment arm mean is greater than the control arm mean exceeds some pre-specified threshold.

Our research challenges the notion that clinical trials must involve direct intervention comparisons. For example, all COVID-19 vaccines reduce the risk of hospitalisation due to COVID-19 infection, but less is known about the variation in immune markers of protection following immunisation. We propose a novel decision-making framework where stopping rules are based on the precision of a statistical quantity. One can reasonably expect that as data accumulates, the precision in a statistical quantity increases and this translates to greater certainty regarding the treatment response. However, there should exist a limit whereby the desire to achieve additional precision in the statistical quantity is outweighed by the incremental cost in trial resources, hence the motivation to define decision rules based on precision.

We will demonstrate the utility of implementing precision-based decision rules using the PICOBOO trial, which evaluates different COVID-19 booster vaccines in immunocompetent children and adults. The trial is designed to generate high-quality data to inform national COVID-19 priming and boosting vaccination practice and policy. Results from the first scheduled analysis will be presented with a walkthrough on the interpretation of a precision based decision-making trial framework.

208

The impact of incorrectly parameterising time in segmented regression models applied to interrupted time series

Dr Louise Marquart-Wilson[1,2], Mrs Elizabeth Underwood[1], Dr Louise Wilson[1], Prof Jenny Doust[1]

[1]School of Public Health, University of Queensland, Brisbane, Australia. [2]QIMR Berghofer Medical Research Institute, Brisbane, Australia

Abstract

Interrupted time series (ITS) studies are quasi-experimental methods used to assess the impact of interventions, such as public health measures or policy changes, on system performance or outcomes over time. Segmented regression (SR) models estimate the intervention effect by comparing the level and slope of the outcome before and after the intervention. In a SR model the parameterization of time in the interaction term can impact the estimation and inference of the level change parameter. Correct parameterisation requires the interaction term to use time since intervention (Xiao et al 2021 Int. J. Epidemiol 50:1011-15).

Preliminary investigation of 75 ITS studies from 2022 that implemented SR showed that 27% had incorrectly parameterised time, 29% had correctly parameterised time; and the remaining studies provided insufficient detail. The impact of parameterising time in SR models with one or multiple interventions was assessed by re-analysing these data and performance assessed using simulation studies. Simulations were motivated by an ITS study of introducing high-sensitive troponin assays on myocardial infarction rates, where parameterisation of time resulted in different statistical inference for level changes. Performance of SR models was evaluated when the interaction term(s) was incorrectly parameterised as time since the start of pre-intervention phase or correctly parameterised as time since intervention(s).

For ITS with a single series and single interruption, incorrect parameterisation of time resulted in biased estimation of the level change parameter exceeding 67% and higher type I error rates. As the number of time-points increased from quarterly to monthly data, the magnitude of percentage bias also increased. Incorrect parameterisation had accurate model-based standard errors for all parameters. Similarly for ITS with two interventions, both level change parameters were substantially biased (percentage bias>10%). Correct parameterisation of time in interaction terms of SR models is necessary to avoid biased estimates and differing statistical inference.

209

Subbagging Variable Selection for Massive Data

Mr Xian Li, Dr Tao Zou, Dr Xuan Liang
The Australian National University, Canberra, Australia

Abstract

Massive datasets usually possess the features of large N (number of observations) and large p (number of variables). In this work, we propose a subbagging variable selection approach to select relevant variables from massive datasets. Subbagging (subsample aggregating) is an aggregation approach originally from the machine learning literature, which is well suited to the recent trends of massive data analysis and parallel computing. Specifically, we propose a subbagging loss function based on a collection of subsample estimators, which uses a quadratic form to approximate the full sample loss function. The shrinkage estimation and variable selection can be further conducted based on this subbagging loss function. We then theoretically establish the consistency and selection consistency for this approach. It is also proved that the resulting estimator possesses the oracle property. However, variance inflation is found in its asymptotic variance compared to the full sample estimator. A modified BIC-type criterion is further developed specifically to tune the hyperparameter in this method. An extensive numerical study is presented to illustrate the finite sample performance and computational efficiency.

210

Can swapping be differentially private? A refreshment stirred, not shaken

James Bailie[1], Professor Ruobin Gong[2], Professor Xiao-Li Meng[1]
[1]Harvard University, Cambridge, MA, USA. [2]Rutgers University, New Brunswick, NJ, USA

Abstract

This paper presents a formal privacy analysis of data swapping, a family of statistical disclosure control (SDC) methods which were used in the 1990, 2000 and 2010 US Decennial Census disclosure avoidance systems (DAS). Like all swapping algorithms, the method we examine has invariants -- statistics calculated from the confidential database which remain unchanged. We prove that our swapping method satisfies the classic notion of pure differential privacy (ε-DP) when conditioning on these invariants. To support this privacy analysis, we provide a framework which unifies many different types of DP while simultaneously explicating the nuances that differentiate these types. This framework additionally supplies a DP definition for the TopDown algorithm (TDA) which also has invariants and was used as the SDC method for the 2020 Census Redistricting Data (P.L. 94-171) Summary and the Demographic and Housing Characteristics Files. To form a comparison with the privacy of the TDA, we compute the budget (along with the other DP components) in the counterfactual scenario that our swapping method was used for the 2020 Decennial Census. By examining swapping in the light of formal privacy, this paper aims to reap the benefits of DP - formal privacy guarantees and algorithmic transparency - without sacrificing the advantages of traditional SDC. This examination also reveals an array of subtleties and traps in using DP for theoretically benchmarking privacy protection methods in general. Using swapping as a demonstration, our optimistic hope is to inspire formal and rigorous framing and analysis of other SDC techniques in the future, as well as to promote nuanced assessments of DP implementations which go beyond discussion of the privacy loss budget ε.

Keywords: Differential Privacy, Statistical Disclosure Control, Data Swapping, US Census, Confidentiality.

212

Statistical preprocessing for privacy-induced spatial mismatch: A multiple imputation approach to poverty prediction using deep learning

Dr. Mohammad Kakooei[1], James Bailie[2], Professor Xiao-Li Meng[2], Professor Adel Daoud[3,1]
[1]Chalmers University of Technology, Gothenburg, Sweden. [2]Harvard University, Cambridge, MA, USA. [3]Linköping University, Norrköping, Sweden

Abstract

With near global coverage from the 1970s onward, satellite data has the potential to improve sustainable development and aid effectiveness by addressing the lack of historical and geographical poverty data. Recent literature has combined satellite images with ground-level survey data via deep learning to assess living conditions over time and across Africa. However, noise is added to the survey data geolocations to protect respondents' privacy. To address the induced spatial mismatch between the satellite and the ground-level data, we impute the survey data locations and fit deep learning models with these imputations. We assess the performance of this approach by predicting night-time light measurements from day-time satellite images, using a human settlement map as an informative prior for the imputation model. The results are a 'triply synthetic' dataset, illustrating aspects of multi-source, multi-phase and multi-resolution inference, as is common across data science.

213

Australian Cancer Atlas 2.0: Advanced models and visualisations

Dr Susanna Cramb ORCID iD[1], Dr Jessica Cameron ORCID iD[2], Mr James Hogg ORCID iD[1], Distinguished Prof. Kerrie Mengersen ORCID iD[1], Prof. Peter Baade ORCID iD[2]
[1]Queensland University of Technology, Brisbane, Australia. [2]Cancer Council Queensland, Brisbane, Australia

Abstract

The original Australian Cancer Atlas launched in 2018, pioneering novel data visualisations and reliable small-area cancer incidence and survival ratios. Enthusiastically embraced, it has received multiple international awards and both the modelling approach and visualisations have been replicated in the Netherlands.

In providing the most detailed information to date, 'the Atlas' also raised additional questions, such as: How are these patterns changing over time? What are the underlying risk factor patterns? And how does cancer screening, stage or treatment vary geographically?

The Australian Cancer Atlas 2.0, launched in the second half of 2023, was an ambitious project designed to answer exactly these questions. Covering over 2200 areas and up to 20 annual time points, a range of new methods were introduced, including Bayesian spatio-temporal and small area estimation models. De-identified unit-level data was obtained from the Australian Cancer Database, National Births, Deaths and Marriages Database, National Health Survey, Hospital separations and other research projects.

The multitude of results presented a new challenge: we needed to develop a website that was intuitive to navigate and showcased the new information in an uncluttered and visually appealing way. But with the help of QUT's Visualisation and Interactive Solutions for Engagement and Research (VISER) group, and expert input, a range of cutting-edge maps (including some bivariate options) and graphs were developed and piloted in workshops.

This revolutionary new Australian Cancer Atlas shows how the combination of high-quality data, appropriate statistical models and innovative visualisations can illuminate information, and can be accessed at https://atlas.cancer.org.au (from late 2023). The project was funded by Cancer Council Queensland and the Australian Research Council (LP200100468).

214

Ten to Men @ Ten: The Australian Longitudinal Study on Male Health reaches its 10-year anniversary.

Dr Sean Martin, Dr. Clement Wong, Dr. Melissa Suares, Ms. Leanne Howell, Ms. Karen Biddiscombe, Ms. Sarah Carr, Dr. Katrina Scurrah
Australian Institute of Family Studies (AIFS), Melbourne, Australia

Abstract

On average, Australian men have a shorter life expectancy than women, and die more often than women from preventable causes. Differences in a range of health-related indicators are known to contribute to the generally poorer health outcomes of Australian boys and men.

Ten to Men: The Australian Longitudinal Study on Male Health was established as part of the National Male Health Policy (2010), one of only four national policies targeting men's health specifically worldwide. Over 16,000 boys (10-13y), young adults (14-17y), and adult (18-55y) males were recruited into the study in 2012-13. Ten to Men completed its fourth wave of data collection in 2023 (n=7,050) with data due for public release in September 2023. .

The overarching aims for *Ten to Men* include: i) examine key determinants of male health and wellbeing; ii) address key research gaps throughout the lifecourse; iii) identify policy opportunities for improving men's health. Currently, the study serves the National Men's Health Strategy 2020-2030 to address priority health issues (mental health, chronic conditions, sexual and reproductive health, injuries and risk taking, and healthy ageing), with a particular focus on priority population groups .

Ten to Men has a range of stakeholders across government, academic, interest groups, peak and representative bodies. With over 1100 variables collected to date (including linkages to Medicare and Pharmaceutical Benefits Scheme) over four waves of data collection, there is scope to continue to contribute to the methodological and technical literature using existing date or collaborate trialling new techniques and methods for future waves).

Ten to Men is seeking your assistance to expand its range of collaborators. Data are freely available for approved researchers. Further details are available at tentomen.org.au.

216

Assessing the influence of gender and pregnancy on disability of Multiple Sclerosis patients

Dr Farzana Jahan ORCID iD[1], Dr Pamela McCombe[2], DProf Kerrie Mengersen[3]
[1]Murdoch University, Western Australia, Australia. [2]University of Queensland, Queensland, Australia. [3]Queensland University of Technology, Queensland, Australia

Abstract

This paper investigates the influence of gender and pregnancy on the level of disability of Multiple Sclerosis (MS) patients. MS patient data from multiple visits were obtained from the MSBase registry. Disability was measured using their Expanded Disability Status Scale (EDSS) scores. Generalised additive mixed models were fitted to the longitudinal data for patients from all their visits to model the EDSS scores with gender and pregnancy, controlling for other confounding variables. The models were fitted separately for 16 different countries and a random-effects meta-analysis was then conducted to estimate the combined effect of gender and pregnancy on MS disability, measured by EDSS.

219

Fast Collapsed Gibbs Sampling Algorithms for Bayesian Penalized Regression in High-Dimensional Data

Dr Mohammad Javad Davoudabadi, A/Prof John Ormerod
University of Sydney, Sydney, Australia

Abstract

Bayesian penalised regression models are widely used for simultaneous parameter estimation and variable selection. They can be fit using computationally expensive Markov chain Monte Carlo sampling especially in high-dimensional settings where the number of predictors is much larger than the number of observations. Block Gibbs sampler is an approach to speed up computation but can be slow when the number of predictors is large (even just a few thousand). In this study, we introduce several fast collapsed Gibbs sampling approaches to address these issues. We also introduce Rao-Blackwellization schemes to reduce the number of samples required. We will demonstrate our collapsed Gibbs samplers are highly efficient on benchmark datasets.

Abstract submission

Download file

220

LACSH: a spatial and causal framework for modeling socio-economic health

Dr Fui Swen Kuh ORCID iD[1,2], Prof Grace Chiu ORCID iD[3,4,5,6,7], Dr. Anton Westveld ORCID iD[6]

[1]Monash University, Melbourne, Australia. [2]University of Adelaide, Adelaide, Australia. [3]Virginia Institute of Marine Science, Williamsburg, USA. [4]Virginia Commonwealth University, Richmond, USA. [5]University of Waterloo, Waterloo, Canada. [6]Australian National University, Canberra, Australia. [7]University of Washington, Seattle, USA

Abstract

We develop a model-based Latent Causal Socioeconomic Health (LACSH) index, with uncertainty bounds, at the national level. Extending the latent health factor index (LHFI) modeling approach to assess ecosystem health, LACSH integratively models the hierarchical relationship among the nation's societal health or well-being (latent / intangible), socio-economic metrics (e.g., GDP), the covariates that drive the notion of well-being (e.g., natural resources), and a continuous variable that reflects policy (e.g., government mandated maternity leave days). In addition to making statistical inference for socio-economic health, LACSH facilitates the evaluation of any causal impact of the policy on health. A formal spatial component in the LACSH framework allows us to compare the socio-economic health of countries around the world based on various metrics, covariates, and two different policy variables that pertain to socio-economic well-being. This is joint work with GS Chiu and AH Westveld (https://arxiv.org/abs/2009.12217).

221

Statistical Consulting Network Panel: This was not in the textbook!

Dr Peter Humburg[1], Dr Alun Pope[2], Professor Virginia Wheway[3]
[1]University of New South Wales, Sydney, Australia. [2]Analytical Insight, Melbourne, Australia. [3]University of Wollongong, Wollongong, Australia

Abstract

Many of the challenges in statistical consulting are well-met by skills in the technical aspects of statistics, combined with skills in communicating with clients. But there are times in consulting when we need to deal with issues or make decisions for which there are no textbook answers. These can sometimes pose an ethical dilemma for consultants. For example: “What do I say to a client who is only interested in the P-values in an analysis?” “If the ‘right’ analysis is too complicated for a client to understand, what can I do – present something that’s not quite right?”

Four statistical consultants will share their experience and expertise on how they deal with these types of issues in a practical and ethical way: Peter Humburg (University of New South Wales), Alun Pope (Analytical Insight), Annie Solterbeck (Statistical Revelations), and Virginia Wheway (Optiver & University of Wollongong).

If you would like to submit a question to the panel, email the session chair: sfinch@unimelb.edu.au

223

Statistical data analysis on Riemannian manifolds

Dr Zengyan Fan
Singapore University of Social Sciences, Singapore, Singapore

Abstract

This paper is concerned by generalizing statistical methods to analyze complex data sets in non-linear spaces. We propose to extend principal components to non-linear Riemannian manifolds and use them to analyze data variation under a fixed direction. In particular, we aim to find a flow with fixed starting and ending points for noisy multivariate data sets lying near an embedded non-linear Riemannian manifold. We construct a high-level algorithm to compute the random fixed boundary flow and the convergence of the algorithm is provided. We further prove that the random fixed boundary flow converges largely to the population fixed boundary flow with high probability. We illustrate how the proposed method can be used and interpreted and showcase its application in analyzing seismology events.

224

When are we going to die? A Bayesian latent variable approach to modelling Australian mortality data from January 2015

Dr John Holmes[1], Prof Lyle Gurrin[2]
[1]University of Canterbury, Christchurch, New Zealand. [2]University of Melbourne, Melbourne, Australia

Abstract

The Covid-19 pandemic has resulted in increased interest in excess mortality modelling. However, the statistical methods used tend to (1) be over-parametrised, (2) fit models that ignore variance heterogeneity, (3) avoid checking if postulated models are consistent over the entire year, (4) stop short of asking what factors drive short-term variation in death rates.

Since 2020, the Australian Bureau of Statistics has been releasing mortality statistics monthly, including age-standardised death rates per week split into a large number of underlying cause categories. Modelling the resulting multivariate mortality time series using a Bayesian approach to multivariate regression plus factor analysis allows us to show

1. While mortality rates in Australia contains a true seasonal component, most seasonality is driven by winter epidemics of Pneumonia causing pathogens. However, over 80 % of these epidemic associated deaths do not list Pneumonia as the underlying cause, and over 60 % do not list any Respiratory condition as the underlying cause.

2. Deviations from the expected winter respiratory illness epidemic mortality in any week split into the different underlying cause categories in the same proportions as expected epidemic mortality.

3. Since January 2020, mortality trends are unchanged except no winter respiratory illness epidemic is visible in 2020 mortality data and Covid-19 circulation increases all-cause mortality rates.

4. In the Omicron period, only 55 \% of estimated Covid-19 associated deaths list Covid-19 as the underlying cause. The reported underlying causes of the remaining Covid-19 associated deaths split into categories in similar proportions to the non-Respiratory underlying causes associated with the winter respiratory illness epidemic.

While estimated Covid-19-associated mortality at wave peaks is comparable to estimated peak winter epidemic mortality for 2015-2022, the greater frequency of waves means the Covid-19 associated mortality burden in 2022 was 3 to 4 times greater that of all winter epidemic respiratory infections combined.

225

WoylieR: An R package for generating a tour of multivariate data using frame to frame interpolation

Zoljargal Batsaikhan ORCID iD[1], Dianne Cook ORCID iD[1], Ursula Laa ORCID iD[2]
[1]Monash University, Clayton, Australia. [2]BOKU University, Vienna, Austria

Abstract

The challenge of visualizing high-dimensional data on 2D screens is addressed by grand tour, which uses linear projections to view data in more than three dimensions. The guided tour combines projection pursuit with the grand tour and allows users to locate the projection of high-to-low dimensional space that exposes the most interesting feature of the data. Projection pursuit optimizes a projection pursuit index function to determine the interestingness of the projected data, and the guided tour uses this to choose new projections to show. Plane-to-plane (geodesic) interpolation between planes makes the motion smooth. However, some projection pursuit indexes are NOT rotation invariant, and the value can be higher or lower depending on the rotation within a plane. This work focuses on the Givens interpolation method to provide frame-to-frame interpolation so that the guided tour will arrive at the optimal 2D projection, and orientation. It is implemented in the R package "woylier."

227

Identification of disease sub-groups from longitudinal patient trajectories

Dr Marcela Cespedes[1], Dr Timothy Cox[2], Dr Hamid Sohrabi[3], Dr Chrostopher Fowler[4], Prof Paul Maruff[5], Assoc. Prof Stephanie Rainey-Smith[6], Prof Colin Masters[4], Dr Jurgen Mejan-Fripp[1], Dr James Doecke[1]
[1]CSIRO, Herston, Australia. [2]CSIRO, Melbourne, Australia. [3]Murdoch University, Murdoch, Australia. [4]Florey Institute of Neuroscience and Mental Health, Melbourne, Australia. [5]Cogstate Limited, Melbourne, Australia. [6]Centre of Excellence for Alzheimer's Disease Research and Care, Joondalup, Australia

Abstract

Alzheimer's disease (AD) can take 30 years to progress from cognitively unimpaired (CU) to clinical dementia. Sigmoid curves allow for three disease progression stages: stable/healthy, progression and clinical disease. For AD, early identification of participants at risk of transitioning between pre-clinical and progressive stages for different cognitive tests is desperately needed for intervention trials. Using longitudinal data from the Australian Imaging, Biomarkers and Lifestyle (AIBL) study of ageing we derived a method that identified participants with progressive cognitive decline prior to clinical identification.

First, we used a data driven approach to estimate sigmoid inflection points via the function's derivatives (Figure 1). Second, participants annual rate of decline and overall mean were mapped to the sigmoid first derivative, allowing for participants age to be offset from their chronological age, creating a disease adjusted age. Inflection points are then calculated via the third derivative. Participants whose longitudinal trajectory cross the first inflection point are classified as progressors, those who do not cross remain CU. Variability is estimated using 1000 bootstrap iterations. We assessed the methods ability to identify progressive cognitive decline using the pre-clinical Alzheimer's cognitive composite (PACC) score; a well-known measure of early cognitive change in pre-clinical AD.

After successful validation of this method on simulated data, it was applied on the PACC score data from 1049 AIBL participants who were at various stages along the AD continuum. From the complete cohort, 619 (59%) were classified as progressors, 287 (27%) as CU and stable, and 143 (14%) in the clinical disease group (Figure 2). The inflection point threshold was found to be at the disease age of 72.4 years (standard deviation of 0.13 years).

Applied to real-world data, this fast and simple data-driven approach disaggregates longitudinal trajectories into three classifications using only participant age and cognitive data.

Abstract submission

Download file

228

Embedding statistical thinking in student learning: what's new and what's not?

Professor Helen MacGillivray
QUT, Brisbane, Australia

Abstract

Avoiding semantic pitfalls, the "big tent" of statistical thinking has been described over time in general terms referring to data needs and production, and awareness, modelling and interpretation of variability. Statistical thinking includes, but is not limited to: the full process of data investigations, framing of problems from contexts, scaffolding on which models are built and used, disposition to query data and data-based reporting, and statistical communication. The learning of statistical thinking is necessarily gradual, scaling from universal essentials for citizenship, to foundational across disciplines for work and future learning, to more discipline-specific higher levels. Leading statisticians have frequently observed that statistical education in other disciplines - of future users, collaborators and information producers – is as important for the statistical and data sciences profession as educating future statistical professionals, with comments also that the latter should have the same type of foundational learning as students in other disciplines.

The environments in which statistical thinking is needed are continuously developing and expanding, sometimes with such acceleration as to (falsely) appear to be dramatically large discontinuities. The long-established and often repeated advocacy that the teaching of statistics should reflect statistical practice, means that such changing environments must be taken into account. However, many of the core principles and good practices built on work over decades still apply, and can be enriched or adapted taking account of teaching and institutional contexts and restrictions. Unfortunately, many previous problems and poor practices still persist; tackling these are ongoing and substantial challenges. This presentation discusses achievements and failures, commenting on what is new and what is not, and busting some myths. It is notable that achievements at all educational levels have involved authentic, committed and egalitarian participation from statisticians. Embedding statistical thinking in statistical and data science education is vital for the statistical and data sciences profession.

229

The use of nudging to enhance learning outcomes

Ms Charanjit Kaur ORCID iD, Dr Joan Tan, Dr Ririn Yuniasih
Monash University, Melbourne, Australia

Abstract

The problem of student attrition is encountered by universities around the globe. Various strategies have been developed to enhance student completion rates. However, many of these strategies mainly rely on performance data that is collected at the level of individual courses or programs. We argue that in order to combat attrition, an effective strategy that focuses on improving completion rate within individual units is required. In light of this, we promote the use of customised feedback as a means of nudging students at risk of academic failure. This approach is based on evidence from 'nudge analytics', which proves the effectiveness of subtle nudges in motivating students to improve their learning outcomes. In line with this approach, we developed an early intervention framework which uses performance data from assessments in an Introductory Statistics unit in order to identify those at-risk of failure. Performance data was used to segment students and provide personalised feedback based on students' achievement. We evaluated the effectiveness of the nudge as an early intervention mechanism, by analysing their post-nudge performance and engagement in learning activities. Results of the analysis show differences in the completion rates among students after the intervention. Forty percent of students who were nudged successfully completed the unit. The findings also indicate differences in students behaviour in learning as well as assessment completion. Those nudged garnered considerable levels of productivity increases in gaining marks within quizzes compared to before the nudge. Implications of this study reveal the positive impact of nudges on student performance. It recommends early intervention measures in order to improve the learning outcomes of those at risk of academic failure.

232

Statistics x Sociology: Relaxing outdated assumptions to estimate literacy proficiency for students in the margins.

Dr Daniela Vasco, A/Professor Samantha Low-Choy, Professor Parlo Singh
Griffith University, Mount Gravatt, Australia

Abstract

Large standardised educational tests are commonly used worldwide, to evaluate and compare literacy of students, teachers, schools and educational systems. They include Australia's National Assessment Program – Literacy and Numeracy (NAPLAN). Like international tests, NAPLAN's value is recognised by some, but highly criticised by others. This critique mainly concerns test design and fairness, including how results are used. We consider a fairness issue: How are students with socio-educational disadvantage handled by statistical analysis underlying NAPLAN?

Literacy scores from NAPLAN are estimated via Item Response Theory (IRT). Australia adopts the simplest, Rasch model: the chance that a student correctly answers a test item relates to item difficulty (fixed effects) and student proficiency (random effects). Proficiency across students is presumed to follow a bell-curve. Our concern was that students "in the margins" with low NAPLAN scores were hidden during reporting, instead absorbed into a coarse "band" of low proficiency. We upgraded to the 3-parameter logistic form, since guessing is prevalent with low literacy. We extended this IRT model, to describe proficiency by a flexible mixture of Normal distributions, rather than a single Normal.

We encountered multifarious challenges. Modelling challenges included: model complexity, label mixing, posterior predictive checks, and prior distributional choices. Computational challenges included limited data, MCMC design and implementation. Several challenges were more socio-political, arising from: controversies surrounding NAPLAN, its high societal impacts, and its uncomfortable position straddling quantitative/qualitative research cultures. Starting with the sociological research question, statistical goals evolved through intensive reading, discussions and debate, whilst navigating political issues. Forensic investigations enhanced transparency of modelling practices, which then enabled testing of model assumptions. Interactions with sociologists in education highlighted a major barrier: the negative connotations of well-entrenched statistical terminology, due to outdated links with controversial theories. Overall, open dialogue with sociologists stimulated statistical thinking, enriching the applied statistics.

233

A Study on Teachers' Awareness of Ethics on Infographics

Salma Banu Nazeer Khan ORCID iD, Associate Professor Ayse Aysin Bombaci Bilgin ORCID iD, Professor Deborah Richards ORCID iD, Professor Paul Formosa ORCID iD
Macquarie University, Sydney, NSW, Australia

Abstract

With the evolution of technology, teaching and learning are revolutionising and new advancements such as infographics are widely embraced in the education sector. There is significant growth in the usage of infographics among students and researchers to communicate results, findings, and reports. Infographics are high-level means of communication encompassed with data visualisation and meaningful narratives to communicate complex statistical information in a story-telling manner. In the context of universities, teachers teach infographic design guidelines to students studying statistics and data science, however, there is still a need for providing ethical guidance to students to identify underlying ethical errors in infographics. These ethical errors can mislead the users of infographics to misinterpret the information and affect their decisions and judgements. Hence, importance needs to be given to ethical guidance in teaching and designing infographics. Our study aims to "train the trainer" to create awareness of ethics on infographics for which we used five ethical principles from the AI4 Peoples framework: non-maleficence, beneficence, justice, autonomy, and explicability in the data science ethics context. We recruited teachers for our study through conferences and mailing (N=15). Our findings show that 60% (9) taught infographics and an overwhelming 44% (4 out of 9) did not teach ethics in infographics. There was a 30% increase in the number of teachers who identified ethical issues after ethical awareness training. Hence, training teachers can significantly influence the education of students, as they are the first point of contact, who can give insight from their experiences. Because teachers are domain experts, it is important to understand ethical perspectives and interpretations of infographics and record knowledge gained through the training. This acquired knowledge can be reused and shared as a teaching module to benefit both other educators and students.

234

Detecting viral antibodies in wildlife (and potentially humans) via Gompertz-mixture distributions: statistical efforts from forensic code analysis and elicitation to computation and decision analysis.

A/Prof Samantha Low-Choy ORCID iD[1,2], Dr Trevelyan McKinley ORCID iD[3], Dr Laura Pulscher ORCID iD[1]

[1]Centre for Planetary Health & Food Security, Griffith University, Nathan, Australia. [2]Griffith Institute for Educational Research, Griffith University, Mount Gravatt, Australia. [3]Exeter University Medical School, Exeter, United Kingdom

Abstract

Detecting viruses in wildlife is important, not only for biodiversity conservation, but more broadly for planetary and human health, as underlined by global pandemics. Many viruses that affect human populations may have a zoonotic origin (from animals). Hence, learning about these viruses in wildlife populations may inform us about similar viruses in humans: improving understanding, technology, and statistical methodologies.

Virus detection uses serological assays to detect antibodies, developed through past exposure to viruses circulating within a population. Unexposed individuals conform to baseline distribution of assay readings, lower than readings of infected individuals. Thus populations with a mixed history of exposure may yield a mixture of baseline and higher readings. Of interest is a threshold that differentiates infected from healthy individuals. A rule-of-thumb sets this threshold at three times the average baseline level, determined from a control sample known to be uninfected. However, this rule has drawbacks. It ignores uncertainty in the threshold, and may misclassify diseased individuals if the assay is non-definitive, i.e. serology distributions overlap for healthy/infected individuals. Thus, ideally, thresholds could be estimated with uncertainty, whilst also managing misclassification.

Previously, serology for flying fox viruses was modelled using a Bayesian mixture of Normal distributions. As serology exhibited a skewed distribution, we adopted a mixture of Gompertz distributions, using new code from a recent study. Forensic analysis of this code, was necessary for clarifying the model and computation. Expert elicitation uncovered knowledge about expected values, so computation was refined. A decision-analytic layer provided a flexible and coherent way to estimate the threshold.

Working with ecologists, statistical efforts were distributed evenly across detective work, technical work, as well as problem-solving and communication. The latter reshaped decision-making by reshaping statistical analysis. Overall, we unearthed a microcosm of statistical practices that affect this method’s usability, reproducibility and relevance.

235

A comparison of methods for external validation of the Kidney Donor Risk Index in the UK transplant population in the presence of semi-competing events

Stephanie Riley[1], Kimberly Tam[1], Wai-Yee Tse[2], Andrew Connor[2], Yinghui Wei[1]
[1]University of Plymouth, Plymouth, United Kingdom. [2]University Hospitals Plymouth NHS Trust, Plymouth, United Kingdom

Abstract

Background

Transplantation represents the optimal treatment for many patients with end-stage kidney disease. The Kidney Donor Risk Index (KDRI) was developed to predict graft failure following kidney transplantation. The survival process following transplantation consists of semi-competing events, where recipient death precludes graft failure but not vice-versa. We sought to externally validate the KDRI in the UK kidney transplant population, and assess whether validation under a competing risks framework had an impact on predictive performance. Additionally, we updated the KDRI using data from the United Kingdom to explore whether this improved the predictive performance.

Methods

We analysed data from recipients of deceased donor single kidney-only transplants held by NHS Blood and Transplant. Our outcomes of interest were one- and five-year graft failure. Considering the semi-competing events, we modelled the outcome in two ways: censoring the recipient at the time of death, and modelling death as a competing event. Cox proportional hazard models were used to validate the KDRI when censoring for death, and cause-specific Cox models were used to account for death as a competing event. KDRI performance was assessed by discrimination, calibration, and overall accuracy of predictions.

Results

The KDRI consistently underestimated event probabilities for those at higher risk of graft failure. When predicting five-year graft failure, discrimination was poorer in the semi-competing risks model, but predictions were more accurate. Calibration plots were similar regardless of whether death was modelled as a competing event or not. The updated KDRI showed worse calibration, but marginally improved discrimination.

Conclusions

Predictive performance for one-year graft failure was similar between death-censored and competing event graft failure, but there were some differences when predicting five-year graft failure. The updated index, developed using UK kidney transplant data, did not show superior performance so we conclude that updating the KDRI is not required.

236

Star Trek in statistical teaching: Drama-based pedagogy for teaching statistical thinking to researchers.

A/Prof Samantha Low-Choy ORCID iD[1], Dr Clair Alston-Knox ORCID iD[2], Adjunct A/Prof Madonna Stinson[1]

[1]Griffith Institute for Educational Research, Griffith University, Mount Gravatt, Australia.

[2]Predictive Analytics Group, Cotswald Hills, Australia

Abstract

Drama-based pedagogy is a well-established collection of methods, ideas and techniques for teaching through role-play, acting-out and other theatrical techniques. As one of the early pioneers of these methods, Brecht aimed to engage people in thinking through the messages of Hitler. Decades later, we consider how these pedagogies can engage researchers in thinking through new statistical approaches, as alternatives to null hypothesis (and significance) tests (NHST). Since the banning of NHST by some journals, and the subsequent social media storm, researchers new to quantitative analysis are now faced with more choices of statistical “paradigm”. Our aim was to create experiences, as a focal point for helping researchers grapple with the ideas behind classical (Frequentist) and Bayesian alternatives to NHST. Researchers from fields—such as social, behavioural and environmental sciences—are often new to statistical analysis and statistical thinking, with low mathematical confidence and/or skills. Our aim was to move beyond a traditional focus on mathematical skills, to encourage dialogue, and hence stimulate statistical thinking which also involves logical skills and an appreciation of scientific method. Here we describe the theatrical techniques that we employed in writing a script inspired by the Star Trek television series from the 1960s. Actors drawn from the class become the main characters (Captain Kirk, Spock, Scottie, and Bones). Class members are given a mission, adopting roles as crew members on the Starship Enterprise. In small groups, the crews design, analyse and interpret the results of an experiment, sending probes down to the planet’s surface. Throughout these statistical activities they are encouraged to discuss and test interpretations. Here we outline some pedagogical techniques we introduced, that combine active learning with theatre. We share evidence on how the science fiction context helped participants to engage more deeply in statistical thinking, through imagination, dialogue and questioning.

237

Statistical thinking to design interviews about risks/benefits of big data: Injecting techniques for structured elicitation of expert knowledge.

A/Prof Samantha Low-Choy ORCID iD[1,2], Dr Judy Rose ORCID iD[3], Adjunct Professor Ross Homel ORCID iD[4], Professor Ilan Katz ORCID iD[5]
[1]Griffith University, Mount Gravatt, Australia. [2]Centre for Planetary Health & Food Security, Griffith University, Nathan, Australia. [3]Griffith Institute of Educational Research, Griffith University, Mount Gravatt, Australia. [4]Griffith Criminology Institute, Griffith University, Mount Gravatt, Australia. [5]Social Policy Research Centre, University of New South Wales, Kensington, Australia

Abstract

Often, interviewing seeks information about a few cases, in depth, to capture complexity or individualised contexts. In contrast quant(itative) methods (such as surveys) may aim to characterise a larger population, with a “broad brush”. Yet this dichotomy does not always apply. A mixed methods approach may deliberately and independently adopt both qual and quant approaches, to address one research question from both perspectives. Our project adopted mixed methods for designing interviews. We adapted techniques from structured elicitation of expert knowledge (SEEK), which evolved for risk analysis and subjective Bayesian analysis.

In business and government, data linkage is a rapidly evolving practice for combining big data from multiple sources, to address new questions or address existing questions more comprehensively. Our project sought insights on the risks and benefits involved in data linkage and subsequent analytics. We chose a practical, strategic and holistic approach that interviewed with this experience, where big data supported social policy towards wellbeing in vulnerable children.

SEEK techniques made subtle though numerous contributions that complemented the qualitative orientation of design and conduct of interviews. SEEK contributed to design of the interview, particularly preparation of experts, and the logical sequencing of questions. In addition, SEEK tuned the focus of questions, since the research about risks and benefits is inherently causal: certain practices [could] have certain consequences (positive or negative) in certain scenarios.

This work sits in an emerging arena at the nexus of statistical and qualitative methods, for research across disciplines, of potential value to all sectors. Harnessing SEEK techniques to embed statistical thinking into design of data collection (here interviews) has strategic benefits. Importantly SEEK could help offset a temptation to analyse “without thinking”, an increasing danger as technologies become more sophisticated for recording and automating analysis of interviews.

238

Investigating the Impact of Lockdowns on Aviation Market Competition: A Case Study of Western Australia

Mr Hongyang Cui[1], Dr Cecilia Xia[2,1], Professor Felix Chan[1], Dr Brett Hughes[1]
[1]Curtin University, Perth, Australia. [2]Data Analysis Australia, Perth, Australia

Abstract

The COVID-19 outbreak has hit the aviation industry hard. With a sharp fall in travel demand and government restrictions, such as lockdowns, The air travel passenger flow decreased significantly. Many studies investigated the impact of COVID on passenger movement. However, limited studies have been done systematically to understand how airline competition reacts to the Covid lockdown measure. This research aims to examine how traffic volume and market share vary by market structure (monopoly, symmetric duopoly, asymmetric duopoly, and asymmetric oligopoly) and travel purpose (natural resource, community, and combination) before, during and after the lockdown in Western Australian (WA) aviation market. The results show that the aviation market in Western Australia was highly concentrated. The lockdown in February has reduced competition and increased market concentration slightly. The February lockdown had a very limited impact on symmetric duopoly market competition, while for the asymmetric duopoly market, the higher concentration of the market occurred during the lockdown. The impact of the lockdowns in February was much higher than the ones in April and June, which demonstrated the resilience of the WA aviation industry.

241

The difficulty of assessing model goodness for aggregate estimates

Dr Lauren Kennedy
University of Adelaide, Adelaide, Australia

Abstract

Multilevel regression and poststratification is a model-based method used for making population and sub-population estimates from surveys. In this method, the principal concern is estimating the population mean of a particular outcome. However, the goodness of this estimate is model dependent (unlike other estimators for this task, which can be doubly robust) – the estimate is only as good as the model is. This means that is very important to ensure the model is good.

In this talk I present recent findings discussing the challenge of assessing goodness when no external benchmarks exist. In recent work, Kuh, Kennedy, Chen & Gelman (2022) demonstrate that average individual goodness (as measured through leave one out cross-validation and expected log predictive density) is not an adequate measure of the goodness of a population estimator. One potential reason for this is that individual level errors can cancel out when creating a population level average, but scoring using the expected log predictive density does not allow for this.

In this talk I consider some alternative schemes that might solve this problem. One solution is to consider individual level scoring schemes other than expected log predictive density. Another is to consider alternative cross-validation schemes that do not rely on individual level goodness.

242

Development of Robust Metabolic Risk Score using a Correlation Structure-based Approach

Dr Dulari Hakamuwa Lekamlage[1], Dr Corey Giles[2], Prof Peter Meikle[2], A/Prof Agus Salim[1]
[1]University of Melbourne, Melbourne, Australia. [2]Baker Heart and Diabetes Institute, Melbourne, Australia

Abstract

Abstract submission

Download file

243

How is the job of a statistician different in academia, government and industry?

Dr Mark Griffin[1], Ms River Paul[2]
[1]Insight Research Services Associated, Brisbane, Australia. [2]Australian Small Business and Family Enterprise Ombudsman, Canberra, Australia

Abstract

With such a lofty title naturally the speakers for this session can only speak about their own experiences working in different sectors, and we also look forward to hearing from people in the audience about their own experiences.

In this session you will hear from:

- River Paul – a long-time employee in government
- Dr Mark Griffin – after spending the first portion of his career in academia, Mark started his own consulting company in 2011 working with clients from academia, government and industry

Some of the topics that we will include are:

- What are our experiences working in different sectors?
- What are our experiences conducting job interviews with candidates coming from different sectors?
- Terms like “statistician” and “data scientist” are bandied about within our community (where there is usually a love/hate relationship between these communities). What is the role of these two professions within the different sectors?
- Any project (statistical or otherwise) is a balance between time, cost and money. How does this balance change between projects and between sectors?
- What software do people use in different sectors?

This session would particularly suit audience members who are:

- Thinking about making a career move between different sectors
- Recruiting candidates from one sector to work in another
- Consultants working with clients in different sectors

244

An Interpretable Regression Framework for Bimodal Count Data

Dr Alan Herschtal, Prof Stephane Heritier
Monash University, Melbourne, Australia

Abstract

In various domains, count data have been empirically observed to be bimodal, indicative of an underlying data generating mechanism that may come from a Poisson or Negative Binomial (NB) two-component mixture. Such domains include seizure counts in epilepsy patients, hospital length of stay, and single cell RNA sequencing data. Historically, standard approaches to modelling such count data include the Negative Binomial (NB) model, the zero-inflated NB model, and estimation of the mixing component frequencies and parameters using the EM algorithm. However, the NB and zero-inflated NB models cannot flexibly accommodate the bimodal nature of count data, and EM based mixture modelling approaches often suffer from non-identifiability of the mixing components and computational tractability issues.

This work explores the use of a three-parameter Poisson-Beta (PB) model for bimodal count data. The PB model is an Poisson infinite mixture with a scaled beta density as the underlying mixing distribution. It contains the 2-parameter zero-inflated Poisson and NB distributions as special cases, but is more generic, allowing for modelling of bimodality in the data. We show that this model has advantages in terms of goodness of fit over commonly considered modelling alternatives such as the NB and zero-inflated NB for a range of synthetic and real-world datasets. Also, unlike mixture models, the non-hierarchical nature of the PB model facilitates inclusion of covariates in a straightforward manner. Further, the PB model allows for enhanced interpretability of findings which would improve communication with non-expert consumers of statistics and thus encourage uptake of the approach.

246

Fuzzy clustering of circular time series based on a new distance

Ángel López-Oriona
King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Abstract

Clustering of time series is a central problem in data mining with applications in many fields. Frequently, the clustering goal is to identify groups of time series generated by the same stochastic process. Whereas several clustering methods have been proposed in the literature for real-valued or even categorical time series, clustering of circular time series has received much less attention. In this work, we propose a new dissimilarity for circular time series, which is based on an association measure specifically designed to deal with the special features of this kind of series. The metric is used as starting point to develop a partitioning around medoids algorithm. An extensive simulation study shows that the proposed clustering procedure outperforms a wide range of alternative methods and exhibits robustness to the noise distribution besides being computationally efficient. Moreover, the method can be successfully applied to several real time series datasets involving wind directions or animal movement, among many others.

248

Identifying Indigenous People in Australian National Data: Challenges and Priorities

Associate Professor Kalinda Griffiths
Poche SA+NT, Flinders University, Adelaide, Australia. VCCC Alliance, Melbourne, Australia

Abstract

The inclusion of Aboriginal and Torres Strait islander people in official statistics in Australia, a realisation that occurred in 1967, brings to light a range of historical and contemporary issues that demand attention. This talk explores the definition of Indigenous peoples, the entities responsible for defining them, and the practical implementation of identification processes in official data collections. It also emphasises the significance of ensuring the completeness and accuracy of data on Aboriginal and Torres Strait islander people, as it directly impacts the measurement of health and wellbeing within the nation.

Official national reporting of the health and wellbeing of Aboriginal and Torres Strait Islander people typically relies on data derived from censuses, vital statistics, administrative data collections, and surveys. In accordance with human rights standards, individuals in Australia have the option to self-identify as ‘Indigenous.’ Historical context and limitations in decision making has made Aboriginal and Torres Strait Islander data a fraught discussion. Even as we further develop our national capabilities in data sharing, challenges persist in obtaining high-quality data, which can lead to biases in estimating the characteristics and progress of Aboriginal and Torres Strait Islander communities. The resulting measurement issues arising from incomplete and inaccurate data also require serious consideration, particularly when assessing our developments in ‘Closing the Gap’ within Australian society.

By recognising the historical and contemporary factors that shape data collection practices, policymakers and researchers should be aiming to support developments in Indigenous Data Sovereignty and self-determining governance. These will work towards enhancing the accuracy and completeness of Aboriginal and Torres Strait Islander data, contributing to more informed decision-making processes and the advancement of equitable outcomes for Aboriginal and Torres Strait Islander peoples in Australia.

249

Multi-resolution Meandering: Personalized Treatments, Individual Privacy, Machine Unlearning, and a World without Randomness

Professor Xiao-Li Meng
Harvard University, Cambridge, USA

Abstract

Data science revolutionizes the granularity of human inquiries and even offers the promise of personalized assessments. However, how can we assess individual treatment effect before treating the individual? Transitional Inference addresses this dilemma through the concept of "transfer to the similar," a notion that has been pondered by philosophers since Galen of the Roman Empire. This talk presents a Multi-Resolution Framework (Li and Meng, 2021, JASA) for transitional inference, where similarity is prescribed probabilistically by concomitantly specifying the sameness — the shared distributional form — and the differences — the individual realizations. This framework avoids the concept of randomness and defines "individual probability" as a deterministic limit with infinite resolution. These conceptualizations help us operationalize the meaning of personalized treatments, clarify what individual privacy is protected by differential privacy, and anticipate the challenges of preserving an individual's right to be forgotten through machine unlearning. Furthermore, it reveals a world that is resistant to overfitting when the resolutions of our data and (deep) learning far exceed the resolution necessary for pattern recognition.

250

Bridging the gap: integrating statistical modelling and mathematical biology

Professor Julie Simpson
University of Melbourne, Melbourne, Australia

Abstract

The success of malaria control relies on the availability of highly effective antimalarial drugs that can significantly improve individual treatment outcomes. The widespread emergence of drug-resistant parasites now threatens the efficacy of first-line treatments, necessitating the urgent development of novel regimens and combinations of existing and new therapeutic agents to ensure adequate cure of malaria.

Addressing this challenge, biostatisticians and mathematical biologists often approach the determination of optimal treatment regimens from a different starting point. Biostatisticians primarily analyse clinical data to estimate the effects of different treatment regimens on patient outcomes. While this approach provides valuable insights into the investigated dosing regimens, it doesn’t provide an appropriate model for predicting patient outcomes under different mechanisms of drug resistance or explore alternative dosing schemes — a necessity for improving the control of infectious diseases. Mathematical biologists begin by developing a model for prediction that captures the biological mechanisms of the infection, such as the life cycle of the malaria parasite within the red blood cell. However, when these “mechanistic” mathematical models are expanded to incorporate treatment actions and patient immunity, they often become highly complex, impeding their validation against clinical data within a proper statistical framework.

This presentation will outline an interdisciplinary approach that brings together mathematical biology and Bayesian statistical methods, demonstrating how it can be used to determine optimal treatment regimens and how this work has informed WHO treatment guidelines for malaria.

251

Trustworthy communication of data-derived evidence: what it is, and how we can get more of it

Professor Sir David Spiegelhalter
University of Cambridge, Cambridge, United Kingdom

Abstract

The recent pandemic has emphasised the key role played by evidence based on data. But how do we decide whether to trust all the claims that are made? Are the numbers being used to manipulate us? Using a wide range of examples, I will look at the way that statistics can be used to try and persuade audiences to think or act in a certain way, and contrast this with efforts to make communication 'trustworthy', by presenting balanced information that seeks to inform rather than persuade. Trustworthy communication should also acknowledge uncertainty and limitations in the quality of the underlying evidence (something sadly missing from ChatGPT).

But if authorities admit their uncertainties and present both the potential benefits and harms of innovations, won't people distrust them? The opposite appears to be the case, and I will discuss results from randomised trials by my colleagues that suggest that current one-sided information actively decreases trust in those sceptical of the innovation who, ironically, are the very people whose trust is being sought.

I will finish by showing some examples of trying to do things properly, in particular relating to the benefits and harms of the AstraZeneca COVID vaccine, and list the questions that anyone should ask whenever you are presented with claims that are based on numbers.

252

The new era of the Royal Statistical Society Centre for Statistics and Data Science Education

Dr Elinor Jones
University College London, London, United Kingdom

Abstract

This talk will outline and discuss international collaborations in relation to the newly established Centre for Statistics and Data Science Education (CSDSE) at the University of Surrey, UK. This exciting new centre will build on the experience of the former Royal Statistical Society Centre for Statistics Education (RSSCSE) in promoting the improvement in statistical and data science education. It will promote data literacy for all, help to develop subject-specific data skills in partnership with other bodies, and argue for specialist provision of a data science curriculum targeted at the 16-19 age range.

253

Multi-resolution Meandering: Personalized Treatments, Individual Privacy, Machine Unlearning, and a World without Randomness

Professor Xiao-Li Meng
Harvard University, Cambridge, MA, USA

Abstract

Data science revolutionizes the granularity of human inquiries and even offers the promise of personalized assessments. However, how can we assess individual treatment effect before treating the individual? Transitional Inference addresses this dilemma through the concept of "transfer to the similar," a notion that has been pondered by philosophers since Galen of the Roman Empire. This talk presents a Multi-Resolution Framework (Li and Meng, 2021, JASA) for transitional inference, where similarity is prescribed probabilistically by concomitantly specifying the sameness — the shared distributional form — and the differences — the individual realizations. This framework avoids the concept of randomness and defines "individual probability" as a deterministic limit with infinite resolution. These conceptualizations help us operationalize the meaning of personalized treatments, clarify what individual privacy is protected by differential privacy, and anticipate the challenges of preserving an individual's right to be forgotten through machine unlearning. Furthermore, it reveals a world that is resistant to overfitting when the resolutions of our data and (deep) learning far exceed the resolution necessary for pattern recognition.

254

Fast statistical inference with neural networks and amortisation: Golden ticket or red herring?

Associate Professor Andrew Zammit-Mangion
University of Wollongong, Wollongong, Australia

Abstract

Neural networks can provide solutions to tasks that were inconceivable just a few years ago and have benefitted society in numerous ways. These benefits primarily stem from a property often referred to as "amortisation": Training a neural network usually requires significant effort and resources but, once trained, the network can solve similar problems repeatedly and rapidly with virtually no additional computational cost. Hence, the substantial initial training cost of training neural networks is "amortised" over time. Amortisation can also be used to enable fast inference with parametric statistical models: Once a network is trained using observational data as input and inferential targets (e.g., model parameters) as output, the network can make inference with future data in a fraction of the computing time needed by conventional likelihood or Monte Carlo methods. These amortised inferential tools have several compelling advantages over classical methods: They do not require knowledge of the likelihood function, are relatively easy to implement, and facilitate inference at a substantially reduced computational cost. In this lecture I will first give a brief review of recent work that has leveraged the property of "amortisation" in statistical inference. I will then evaluate the merits and drawbacks of amortised inference from a statistician's perspective and conclude by outlining the challenges that need to be overcome for these inferential tools to gain widespread acceptance.

255

Bridging the gap: integrating statistical modelling and mathematical biology

Professor Julie Simpson
University of Melbourne, Melbourne, Australia

Abstract

The success of malaria control relies on the availability of highly effective antimalarial drugs that can significantly improve individual treatment outcomes. The widespread emergence of drug-resistant parasites now threatens the efficacy of first-line treatments, necessitating the urgent development of novel regimens and combinations of existing and new therapeutic agents to ensure adequate cure of malaria.

Addressing this challenge, biostatisticians and mathematical biologists often approach the determination of optimal treatment regimens from a different starting point. Biostatisticians primarily analyse clinical data to estimate the effects of different treatment regimens on patient outcomes. While this approach provides valuable insights into the investigated dosing regimens, it doesn’t provide an appropriate model for predicting patient outcomes under different mechanisms of drug resistance or explore alternative dosing schemes — a necessity for improving the control of infectious diseases. Mathematical biologists begin by developing a model for prediction that captures the biological mechanisms of the infection, such as the life cycle of the malaria parasite within the red blood cell. However, when these “mechanistic” mathematical models are expanded to incorporate treatment actions and patient immunity, they often become highly complex, impeding their validation against clinical data within a proper statistical framework.

This presentation will outline an interdisciplinary approach that brings together mathematical biology and Bayesian statistical methods, demonstrating how it can be used to determine optimal treatment regimens and how this work has informed WHO treatment guidelines for malaria.

256

The tale of a career with large variance and a central tendency of promoting good statistical practice

Dr Marijke Welvaert
Australian Red Cross Lifeblood, Sydney, Australia

Abstract

Starting out studying a degree in psychology, my career has taken many unexpected turns. Whilst it would be easy to state that this was always part of the plan, the truth is more like a random walk being the driver. I have counted moths and mice. I have been hooked up to machines. I have received electroshocks. And that all for the greater good of statistics.

In this lecture I will walk you through the varied path of my career, lessons I learned along the way and demonstrating that the career of an applied statistician can have large variance, but at the centre of it is the quest for improving statistical practice in all application areas.

257

Identifying Indigenous People in Australian National Data: Challenges and Priorities

Associate Professor Kalinda Griffiths
Flinders University, Adelaide, Australia

Abstract

The inclusion of Aboriginal and Torres Strait islander people in official statistics in Australia, a realisation that occurred in 1967, brings to light a range of historical and contemporary issues that demand attention. This talk explores the definition of Indigenous peoples, the entities responsible for defining them, and the practical implementation of identification processes in official data collections. It also emphasises the significance of ensuring the completeness and accuracy of data on Aboriginal and Torres Strait islander people, as it directly impacts the measurement of health and wellbeing within the nation.

Official national reporting of the health and wellbeing of Aboriginal and Torres Strait Islander people typically relies on data derived from censuses, vital statistics, administrative data collections, and surveys. In accordance with human rights standards, individuals in Australia have the option to self-identify as ‘Indigenous.’ Historical context and limitations in decision making has made Aboriginal and Torres Strait Islander data a fraught discussion. Even as we further develop our national capabilities in data sharing, challenges persist in obtaining high-quality data, which can lead to biases in estimating the characteristics and progress of Aboriginal and Torres Strait Islander communities. The resulting measurement issues arising from incomplete and inaccurate data also require serious consideration, particularly when assessing our developments in ‘Closing the Gap’ within Australian society.

By recognising the historical and contemporary factors that shape data collection practices, policymakers and researchers should be aiming to support developments in Indigenous Data Sovereignty and self-determining governance. These will work towards enhancing the accuracy and completeness of Aboriginal and Torres Strait Islander data, contributing to more informed decision-making processes and the advancement of equitable outcomes for Aboriginal and Torres Strait Islander peoples in Australia.

258

Causal machine learning in health data science

Associate Professor Margarita Moreno-Betancur
Murdoch Children's Research Institute, Melbourne, Australia. University of Melbourne, Melbourne, Australia

Abstract

The ultimate goal of medical and health research is to improve patient outcomes and population health. As a result, the overwhelming majority of clinical and public health research studies ask “causal” questions, concerning the effect of treatments, policies, behaviours and other exposures on health outcomes. In many cases, especially in the current era of data deluge, these studies rely on observational (non-experimental) data to address causal questions. Unfortunately, for a long time the statistics discipline largely shunned the possibility of causal inference beyond randomised trials, and instead focused on the development of tools such as regression models without clarity regarding their usefulness and limitations for addressing the causal questions that substantive areas continued to ask. In recent decades, however, the discipline has seen the rise of a new area focused on determining the settings and approaches that could allow causal inference from observational data.

This talk will first provide an overview of some of the fundamental contributions of this statistical area to enable and improve the study of causality in health research, and then describe the role of machine learning within this causal inference paradigm, including recent methodological advances.

259

“I don’t care about the data …” (said a statistician)

Professor Ian Gordon
Statistical Consulting Centre, University of Melbourne, Melbourne, VIC, Australia

Abstract

Have you ever heard a statistician say this? I have. Why might this be said, and what should be the disposition of a statistical educator to this attitude?

This presentation will explore the many dimensions of “caring about the data”, including: thinking broadly about context, design and research questions; data quality and integrity; representing data well; the coherence of data, models and inferences; data ethics, and ways to encourage student understanding and experience of these issues. I will provide examples from my own experience as a statistician and statistical educator.

260

Trustworthy communication of data-derived evidence: what it is, and how we can get more of it

Professor Sir David Spiegelhalter
Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Cambridge, United Kingdom

Abstract

The recent pandemic has emphasised the key role played by evidence based on data. But how do we decide whether to trust all the claims that are made? Are the numbers being used to manipulate us? Using a wide range of examples, I will look at the way that statistics can be used to try and persuade audiences to think or act in a certain way, and contrast this with efforts to make communication 'trustworthy', by presenting balanced information that seeks to inform rather than persuade. Trustworthy communication should also acknowledge uncertainty and limitations in the quality of the underlying evidence (something sadly missing from ChatGPT).

But if authorities admit their uncertainties and present both the potential benefits and harms of innovations, won't people distrust them? The opposite appears to be the case, and I will discuss results from randomised trials by my colleagues that suggest that current one-sided information actively decreases trust in those sceptical of the innovation who, ironically, are the very people whose trust is being sought.

I will finish by showing some examples of trying to do things properly, in particular relating to the benefits and harms of the AstraZeneca COVID vaccine, and list the questions that anyone should ask whenever you are presented with claims that are based on numbers.

261

The Moderating Effect of Dietary Patterns on the Relationship between Depression, Anxiety, and Cognitive Function

Hilal Salim Said Al Shamsi[1], Samantha Gardener[1,2,3], Hamid Sohrabi[1,3], Stephanie Rainey-Smith[1,2,3], W.M.A.D. Binosha Fernando[1,2], Ralph Martins[1,2,4]
[1]Edith Cowan University, Joondalup, Australia. [2]Ralph and Patricia Sarich Neuroscience Research Institute, Nedlands, Australia. [3]Murdoch University, Murdoch, Australia. [4]Macquarie University, Sydney, Australia

Abstract

Objective: To evaluate whether dietary patterns moderate the relationship between anxiety, depression, and cognitive function.

Method: Cognitively unimpaired participants (n=661) from the Australian Imaging, Biomarkers and Lifestyle (AIBL) study were included in the study. These participants completed a food frequency questionnaire, had depression and anxiety data, and underwent a comprehensive neuropsychological assessment. Individual neuropsychological test scores were used to construct composite scores for six cognitive domains. Scores for three dietary patterns, 1) Mediterranean diet (MeDi), 2) western diet, and 3) Dietary Approaches to Stop Hypertension (DASH) diet were generated for each individual. A moderation analysis and simple slope analysis were used to examine the interactions between dietary patterns, anxiety, depression, and cognition.

Results: Depression was negatively associated with executive function cognitive composite in Apolipoprotein E (APOE) ε4 allele carriers with mean and higher than mean MeDi score adherence.

Depression was negatively associated with episodic recall and AIBL Preclinical Alzheimer Cognitive Composite in APOE ε4 allele carriers with mean and below mean levels of western diet adherence. The association with episodic recall was also observed in males with below mean western diet adherence.

Depression was negatively associated with language cognitive composite in APOE ε4 allele non-carriers with mean and below mean levels of western diet adherence. This association was also observed in females with below mean western diet adherence.

Anxiety level was negatively associated with language composite in APOE ε4 allele non-carriers with mean and below mean levels of western diet adherence.

Conclusion: The study highlights dietary factors as potential moderators for cognitive function and mental health, particularly in higher-risk groups, such as females and APOE ε4 allele carriers. This emphasizes the importance of gender-specific and genotype-specific perspectives in mental health-cognition research and highlights the need for continued investigation.

262

Development and Validation of a composite Healthy Ageing Score (HAS) for middle-to-older aged Australians

Eme John[1], Thomas Astell-Burt[1], Ping Yu[1], Chris Brennan-Horley[1], Xiaoqi Feng[2]
[1]University of Wolloongong, Wollongong, Australia. [2]University of New South Wales, Sydney, Australia

Abstract

We developed and validated a composite healthy ageing score (HAS) to address the absence of a definitive composite score comprising multiple health domains that measure healthy ageing in epidemiology. The HAS is developed from 13 health domains reported to influence healthy ageing. Data to measure these domains was extracted from the 45 and Up Study baseline. We applied best practices for scale validation and development. Physical functioning, cognitive function, mental health, sleep, quality of life, balance, social connections and overall health were retained. Functional capacity and resilience were uncovered as underlying latent structures. The HAS ranges from 0 to 16 with higher scores indicating a better health profile. This research contributes a comprehensive measuring tool, HAS, It enables the examination and comparison of individual or collective health profiles and the investigation of the factors that influence their chances of living healthy for longer.

265

Lunchtime session: Writing for Significance

A/Prof Susanna Cramb

Abstract

Do you consider yourself a budding science communicator? Do you have a compelling data story to tell? Think you can take the jargon out of your stats chat and put your writing skills to the test? If so, then this is the conference session for you! Join the SSA *Significance* Editorial Board members and guests over a bite to eat during the lunchtime session on Thursday 14th December at ASC 2023 to learn about how to write accessible statistical articles for *Significance*: a magazine discussing the role of statistics in life, science, politics and business.

You can learn more about Significance magazine before the session at https://significancemagazine.com/