SASA 2014

56th ANNUAL CONFERENCE OF THE SOUTH AFRICAN STATISTICAL ASSOCIATION

27 — 30 October 2014

Rhodes University, Grahamstown, South Africa

PROGRAMME & ABSTRACTS

RHODES UNIVERSITY
Where leaders learn

sas THE POWER TO KNOW.

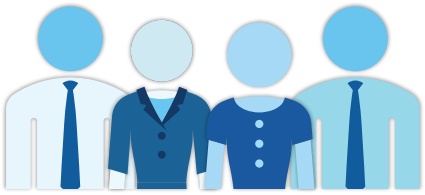
Get With the Program!

Join the thousands of students and professors who are already using free SAS® software and resources.

SOFTWARE ACCESS

58,000+
downloads of free SAS University Edition software
since its launch on May 28

Number of professors trained to teach SAS grows every year



24%
ANNUALLY

CERTIFICATION

SAS Certifications	SAS Exams Attempted
⬆ 17%	⬆ 14.5%

E-LEARNING AND TUTORIALS

FREE E-LEARNING COURSES

9,000+
enrolled

SAS TRAINING TUTORIALS

250,000
visits

Join the U!

sas.com/programu

SAS® ANALYTICS U

sas
THE POWER TO KNOW®

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are trademarks of their respective companies. Copyright © 2014 SAS Institute Inc. Cary, NC, USA. All rights reserved. 107274_S129054US.0814

56^{th} ANNUAL CONFERENCE
OF THE
SOUTH AFRICAN
STATISTICAL ASSOCIATION

RHODES UNIVERSITY
GRAHAMSTOWN

2014

PROGRAMME
and
ABSTRACTS

Contents

1 Introduction **1**

2 General Information **1**

3 Social Events **2**

4 Venues **2**

5 Meetings **2**

6 SASA 2014 Organising Committee **2**

7 Guidelines to Speakers and Chairpersons **3**

8 Programme at-a-glance **4**

9 Detailed Programme **5**

10 List of Poster Presentations **15**

11 Workshops **16**

12 Abstracts **17**
12.1 Keynotes 17
12.2 Oral Presentations (Surnames A - J) 19
12.3 Oral Presentations (Surnames K - O) 41
12.4 Oral Presentations (Surnames P - Z) 63
12.5 Poster Presentations 78

13 List of Delegates **85**

14 Maps **90**

1 Introduction

The South African Statistical Association (SASA) and the Department of Statistics at Rhodes University are proud to host the 56^{th} annual SASA conference. The SASA conference will be held from 27 to 30 October 2014 in Grahamstown. The workshops will take place on Monday 27 October and the main conference will take place during the period 28 to 30 October 2014.

We hope that you will find the conference programme interesting. Also, we trust that you will enjoy the planned social events. We welcome you all to Rhodes University which is situated in the picturesque Grahamstown. It is the smallest university in South Africa, with approximately 7 400 students.

2 General Information

Registration
The registration desk is situated in the foyer of the Eden Grove building and will be open:

Monday	27 October	07h30 - 09h00
Tuesday	28 October	07h30 - 09h00
Wednesday	29 October	07h30 - 08h30
Thursday	30 October	07h30 - 08h30

All questions and queries should be directed to the staff at the desk.

Internet
There will be Wi-Fi access available for delegates.
Instructions to access the internet

- Click Wi-Fi on the task bar
- Click Events @Rhodes
- Select your event (SASA 2014)
- Enter the username and password given below

Username	ev-sasa201
Password	gru24megra
Event name	SASA 2014

Delegates can also connect to the eduroam wireless network. If you are visiting Rhodes and your home institution participates in eduroam, you should be able to get free Internet access at Rhodes by simply connecting to the eduroam wireless network at any of our on-campus hotspots. You will need to authenticate with your home institution's credentials.

Parking
Free, secure parking is available next to the venue. You are requested not to park in the reserved parking bays.

Name Tags
Delegates should wear their name tags at all times to gain access to the lecture halls, lunch venue and evening functions.

Tickets
Tickets will be issued for lunches and social functions. Please make sure you have these tickets on hand as they will give you access to all the lunches and social functions.

Emergency Numbers

Campus Protection Unit (CPU)	046 603 8146
Health Care Centre	046 603 8523
Department of Statistics	046 603 8346
SASA 24-hour line	079 510 2453

Poster Session
Mounting materials for the posters will be available at the registration desk. Posters must be up before 18h30 on 28 October, and remain on the boards until 21h00 on 28 October. Alternatively, you can collect your poster from the registration desk on 29 October.

3 Social Events

- **Meet and Greet for Workshop attendees:** 27 October (Monday) at 18h30 at Prime.
- **Cocktail Function and Poster Presentations:** 28 October (Tuesday) from 18h30 - 21h00 in the Foyer at Eden Grove.
- **Young Statisticians' Function:** 28 October (Tuesday) at 21h00 at the Rat and Parrot.
- **Gala Dinner:** 29 October (Wednesday) at 19h00 on the Drostdy Lawns.

4 Venues

- **Conference venue:** Eden Grove Complex.
- **Tea/Coffee:** Foyer Eden Grove Complex.
- **Lunch:** Marquee on Drostdy Lawns.
- **Cocktail function:** Foyer Eden Grove Complex.
- **Gala dinner:** Marquee on Drostdy Lawns.

5 Meetings

- **SASA Executive Committee meeting:** 27 October (Monday) from 16h30 - 18h00 in Seminar room 3 Eden Grove.
- **Multivariate Data Analysis Group (MDAG) Annual General meeting:** 28 October (Tuesday) from 17h50 - 18h20 in Seminar room 3 Eden Grove.
- **SASA Annual General meeting:** 29 October (Wednesday) from 11h50 - 12h50 in Eden Grove Red.

6 SASA 2014 Organising Committee

- Ms Mutsa Chinyamakobvu - M.Chinyamakobvu@ru.ac.za
- Prof Sarah Radloff - S.Radloff@ru.ac.za
- Dr Lizanne Raubenheimer (chair) - L.Raubenheimer@ru.ac.za

7 Guidelines to Speakers and Chairpersons

1. **Speakers**

   - Double check the date and time of your presentation.
   - Load your presentation on the computer **before** the start of the session.
   - Report to the chairperson of the session before the start of the session.
   - Keep to the time allocated for your presentation.
   - You are not allowed to move your presentation to any other time slot.
   - Laser pointers will be available in the venue.
   - **Once the chair indicates the end of your session, you must stop your presentation immediately.**

2. **Chairpersons**

   - Double check the date and time of your session.
   - **Keep to the scheduled times.**
   - No changes are to be made to the programme.
   - Check the attendance of all the speakers, and ensure that all presentations have been loaded on the computer.
   - Welcome delegates and speakers at the beginning of your session.
   - Make the following announcements:
     - All cell phones to be switched off.
     - State the programme for the session.
     - Start with the first lecture.
   - Warn speakers 5 minutes before the end of their allocated time.
   - Allow questions according to time.
   - Thank all speakers and delegates at the end of the session.
   - Report to the front desk if a speaker was absent.
   - Report shortcomings to the session assistant.

These are only guidelines. Please use your own initiative to make sessions a success.

8 Programme at-a-glance

Monday 27 October 2014	
08h00 - 09h00	Registration (EG Foyer)
09h00 - 11h00	Workshop 1 (EG Red) and Workshop 2 (EG Blue)
11h00 - 11h30	Tea (EG Foyer)
11h30 - 13h00	Workshop 1 (EG Red) and Workshop 2 (EG Blue)
13h00 - 14h00	Lunch (Marquee, Drostdy Lawns)
14h00 - 15h30	Workshop 1 (EG Red) and Workshop 2 (EG Blue)
15h30 - 16h00	Tea (EG Foyer)
16h00 - 18h00	Administrative Meetings
18h30 - ☺	Meet & Greet (Prime) for workshop attendees
Tuesday 28 October 2014	
08h00 - 09h00	Registration (EG Foyer)
09h00 - 10h10	Opening Ceremony (EG Red)
10h10 - 11h00	Keynote 1: Prof Ivette Gomes (EG Red)
11h00 - 11h30	Tea (EG Foyer)
11h30 - 12h10	Parallel Session 1
12h10 - 13h00	CoE discussion session: Prof Ebrahim Momoniat (EG Red)
13h00 - 14h00	Lunch (Marquee, Drostdy Lawns)
14h00 - 15h40	Parallel Session 2
15h40 - 16h10	Tea (EG Foyer)
16h10 - 17h50	Parallel Session 3
17h50 - 18h20	MDAG Annual General Meeting (EG Seminar room 3)
18h30 - 21h00	Cocktail Function and Poster Session (EG Foyer)
21h00 - ☺☺	Young Statisticians' Function (Rat and Parrot)
Wednesday 29 October 2014	
08h30 - 10h30	Parallel Session 4
10h30 - 11h00	Tea (EG Foyer)
11h00 - 11h50	Keynote 2: Prof Jan Swanepoel (EG Red)
11h50 - 12h50	SASA Annual General Meeting (EG Red)
12h50 - 13h50	Lunch (Marquee, Drostdy Lawns)
13h50 - 15h50	Parallel Session 5
15h50 - 16h10	Tea (EG Foyer)
16h10 - 17h50	Parallel Session 6
19h00 - ☺	Gala Dinner (Marquee, Drostdy Lawns)
Thursday 30 October 2014	
08h30 - 10h50	Parallel Session 7
10h50 - 11h20	Tea (Eden Grove Foyer)
11h20 - 12h10	Keynote 3: Prof Simos Meintanis (EG Red)
12h10 - 12h30	Closing (EG Red)
12h30	Please remember to take a packed lunch before you leave

9 Detailed Programme

Monday 27 October 2014		
08h00 - 09h00	Registration (Eden Grove Foyer)	
09h00 - 11h00	Workshop 1	Workshop 2
	Venue: Eden Grove Red Statistics of extremes: An introduction **Presenter:** Prof Ivette Gomes	**Venue:** Eden Grove Blue Inference procedures based on the empirical characterstic function **Presenter:** Prof Simos Meintanis
11h00 - 11h30	Tea (Eden Grove Foyer)	
11h30 - 13h00	Workshop 1	Workshop 2
13h00 - 14h00	Lunch (Drostdy Lawns)	
14h00 - 15h30	Workshop 1	Workshop 2
15h30 - 16h00	Tea (Eden Grove Foyer)	
16h00 - 18h00	Administrative meetings	
18h30 - ☺	Meet & Greet (Prime) for workshop attendees	

Tuesday 28 October 2014				
08h00 - 09h00	Registration (Eden Grove Foyer)			
09h00 - 10h10	Opening Ceremony (Eden Grove Red) Chair: Prof Paul Mostert			
	Welcoming: SASA President, Prof Paul Mostert Opening: Deputy Vice-Chancellor: Research and Development, Dr Peter Clayton, Rhodes University Presidential Address: Prof Paul Mostert Awards: SAS awards for best honours projects Awards: Statistics SA awards for best post-graduate papers Award: Sichel Medal Awards: Fellowship and Honorary Members Award: SAS Thought Leader Award Platinum Sponsor Address: SAS, Mr Murray de Villiers			
10h10 - 11h00	Keynote 1 (Eden Grove Red)			
	Title: Penultimate approximations and reliability of large coherent systems: Past, present ... and future? **Presenter:** Prof Ivette Gomes **Chair:** Prof Tertius de Wet			
11h00 - 11h30	Tea (Eden Grove Foyer)			
	Parallel Session 1 (11h30 - 12h10)			
	Stream: Astrostatistics **(1A)** **Venue:** EG Red **Chair:** Dr Melvin Varughese	**Stream:** Statistics in Education **(1B)** **Venue:** EG Blue **Chair:** Dr Tom Berning	**Stream:** Sports Statistics **(1C)** **Venue:** EG Seminar Room 1 **Chair:** Prof Hoffie Lemmer	**Stream:** Insurance **(1D)** **Venue:** EG Seminar Room 2 **Chair:** Dr Franck Adekambi
11h30 - 11h50	A nonparametric approach to classifying supernova lightcurves **M. M. Varughese**, R. von Sachs, M. Stephanou and B. A. Bassett	Real-time educational interpreting in Statistics **T. Berning**	A method to measure strangling, a dramatic form of choking in cricket **H. Lemmer**	Distribution of the aggregate amount of benefit **F. Adekambi**
11h50 - 12h10	Independent component analysis for pulsar classification **A. Hazra** and S. Ray	Estimation of promotion rates, repetition rates and dropout rates for learners in South African schools **D. Uys** and E. Alant	Comparing the accuracy and fairness of three different resource tables for the Duckworth-Lewis method in Twenty20 cricket **W. Oosthuizen**, R. Schall, I. Carney and R. Ramudzuli	Integral equations for moments in multistate life and health insurance models **F. Adekambi** and M. C. Christiansen
12h10 - 13h00	CoE discussion session (Eden Grove Red): Prof Ebrahim Momoniat, University of the Witwatersrand			
13h00 - 14h00	Lunch (Drostdy Lawns)			

	Parallel Session 2 (14h00 - 15h40)			
	Stream: Bayesian Statistics **(2A)** **Venue:** EG Red **Chair:** Dr Andréhette Verster	**Stream:** Business Stream **(2B)** **Venue:** EG Blue **Chair:** Prof Willie Conradie	**Stream:** Young Statisticians **(2C)** **Venue:** EG Seminar Room 1 **Chair:** Mr Jaco Visagie	**Stream:** Young Statisticians **(2D)** **Venue:** EG Seminar Room 2 **Chair:** Mr Greg Distiller
14h00 - 14h20	Comparing peak over threshold models in terms of the extreme value indexes **A. Verster**	Extreme quantile estimation by using limited historical data and scenario assessments **P. J. de Jongh**, T. de Wet, H. Raubenheimer and J. H. Venter	A topic model approach to inferring episodic directional selection in protein coding sequences **H. Sadiq** and M. Lacerda	Jump-diffusion based-simulated expected shortfall (SES) method of correcting Value-at-Risk (VaR) under prediction tendencies **S. Magagula** and J. O. Oloami
14h20 - 14h40	A variational Bayes approach to analysing site-occupancy models **A. Clark**, R. Altwegg and J. Ormerod	Fair value adjustments for smaller financial institutions **C. van der Merwe**	Analysis of extreme rainfall at East London, South Africa **T. Diriba**, L. K. Debusho, J. Botai and A. Hassen	Matrix variate elliptical model: subjective Bayesian inference **J. van Niekerk**, A. Bekker, M. Arashi and D. J. de Waal
14h40 - 15h00	Objective Bayesian priors for the unbalanced one-way random effects model **J. Harvey** and A. J. van der Merwe	A new form of leverage in stochastic volatility models **I. MacDonald**, W. Zucchini and R. Langrock	Online quantiles via Hermite series density estimation over data streams **M. Stephanou** and M. M. Varughese	Efficiency analysis of South African tertiary education institutions using data envelopment analysis **I. Chitekedza**, W. Brettenny and J. Hugo
15h00 - 15h20	Probit and skew robit models for use in sequential regression multiple imputation **M. J. von Maltitz**	The use of point processes and Markov chains to predict delinquency levels for consumers **F. van der Walt** and F. Lombard	Principal points and principal curves **R. Ganey**	Comparing statistical methodologies in an application to assess perceptions associated with bank card fraud susceptibility **M. Mnci**, W. Brettenny and G. D. Sharp
15h20 - 15h40	A family of weighted bivariate beta type I distributions and its generalizations **R. Ehlers**, A. Bekker and M. Arashi	Has oil price predicted stock returns for over a century? **R. Gupta** and P. K. Narayan	Calibrating option pricing models to down-and-out barrier call option prices **J. Visagie**	Chasing shadows: Analysing data from camera trap studies **G. Distiller** and D. Borchers
15h40 - 16h10	Tea (Eden Grove Foyer)			

	Parallel Session 3 (16h10 - 17h50)			
	Stream: Biostatistics **(3A)** **Venue:** EG Red **Chair:** Prof Francesca Little	**Stream:** Applied Statistics & General **(3B)** **Venue:** EG Blue **Chair:** Dr Richard van der Wath	**Stream:** Young Statisticians **(3C)** **Venue:** EG Seminar Room 1 **Chair:** Mr Johan Ferreira	**Stream:** Young Statisticians **(3D)** **Venue:** EG Seminar Room 2 **Chair:** Mr Sean van der Merwe
16h10 - 16h30	Growth curve comparisons to assess the impact of different ARV treatment regimes on Head Circumference profiles of HIV+ infants **F. Little** and B. Laughton	Quantitative challenges in building an agent-based model of a colonic crypt **R. van der Wath**	Comparative assessment of radiation data for photovoltaic plant applications **C. Clohessy**, G. D. Sharp, E. van Dyk and F. Vorster	Bayesian adaptation of the parametric bootstrap **S. van der Merwe**
16h30 - 16h50	On the use of a semi-parametric estimator for censored data, with applications in food safety exposure assessment **M. Aerts**, R. Nysen and C. Faes	Coverage probabilities and average of confidence intervals for the lengths parameters and quantiles of the Weibull distribution **P. T. Iiyambo**	Forecasting models for possible impact of climate change on the electricity demand in South Africa: Empirical analysis **T. Nyulu** and I. Litvine	Comparison of the genetic algorithm and incremental optimisation routines for a network design problem **A. Nickless**
16h50 - 17h10	Estimating HIV incidence from prevalence data obtained from pregnant women in rural KwaZulu-Natal **K. Leask** and A. Kharsany	Properties and applications of the negative hypergeometric distribution **H. Moolman**	Balanced modified random sampling in the presence of linear trend **L. R. Naidoo**	Modelling residential electricity usage within the eThekwini Municipal area **S. Reade**
17h10 - 17h30	Statistical methodology to measure the HIV serodiscordance among couples: The case of Mozambique **A. Juga**, M. Aerts and N. Hens	Modeling and forecasting South African CPI using ARIMA models **S. Masimula**	Mixed effects modeling in conjoint analysis **T. Cronje**, F. H. J. Kanfer, S. M. Millard and M. Arashi	Causality testing for regime-switching processes **F. Mlambo** and I. Litvine
17h30 - 17h50	New insights into the HIV epidemic in South Africa **K. Zuma**, O. Shisana, T. Rehle, L. Simbayi and S. Jooste	Exploratory analysis of functional data **S. Mangisa**, S. Das and G. D. Sharp	Compounding as method in bivariate extended noncentral distributional development **J. Ferreira**, A. Bekker and M. Arashi	Bayesian non-linear mixed effects regression model for the character-isation of early bactericidal activity of tuberculosis drugs **D. A. Burger** and R. Schall
17h50 - 18h20	Multivariate Data Analysis Group (MDAG) Annual General Meeting (Eden Grove Seminar Room 3)			
18h30 - 21h00	Cocktail Function and Poster Session (Eden Grove Foyer)			
21h00 - ☺☺	Young Statisticians' Function (Rat and Parrot)			

Wednesday 29 October 2014				
	Parallel Session 4 (08h30 - 10h30)			
	Stream: Multivariate Data Analysis Group **(4A)** **Venue:** EG Red **Chair:** Prof Sugnet Lubbe	**Stream:** Statistics in Education **(4B)** **Venue:** EG Blue **Chair:** Prof Delia North	**Stream:** Official Statistics **(4C)** **Venue:** EG Seminar Room 1 **Chair:** Dr Patrick Naidoo	**Stream:** Biostatistics **(4D)** **Venue:** EG Seminar Room 2 **Chair:** Dr Gaëtan Kabera
08h30 - 08h50	Time series outlier detection using singular spectrum analysis and robust principal component analysis **J. de Klerk**	On-line teaching of Statistics D. North, **J. S. Allison**, **S. Britz** and **M. Jankowitz**	Exploring supply-side municipal service delivery in South Africa using non-financial census of municipalities M. Nthangeni and **P. Naidoo**	Some statistical methods for detecting drug interaction **G. Kabera**, P. Ndlovu and L. M. Haines
08h50 - 09h10	The consolidation of forecasts with linear regression models **D. Venter**	On-line teaching of Statistics D. North, **J. S. Allison**, **S. Britz** and **M. Jankowitz**	Factors associated with youth unemployment in South Africa **G. Simangwe**	Immunological responder definition: Tuberculosis clinical research **E. van Zyl**
09h10 - 09h30	Size and shape of multivariate data **M. Greenacre**	Laying the foundation for a Statistically literate society in South Africa – lessons learnt in KZN, way forward **D. North**, T. Zewotir and I. Gal	Comparison of socio-economic status index with the income variable **M. S. Nkwinika**	The estimation of age-specific reference intervals using fractional polynomials and an exponential transformation **T. Reddy** and I. Bhorat
09h30 - 09h50	Stepwise latent class analysis with continuous distal outcome variables **Z. Bakk** and J. K. Vermunt	The affective impact of computer-assisted teaching and learning in statistics **L. van der Merwe**	Exploring inequalities in utilization of maternal health services in South Africa **L. Matizirofa**	Estimation of high dimensional longitudinal and recurrent events data joint models using marginal likelihood techniques **L. Masenyetse**, S. O. M. Manda and H. Mwambi
09h50 - 10h10	A biplot perspective on market-based valuations in an emerging market **N. le Roux**, S. Nel and W. Bruwer	Repeating service modules - reasons and remedies **E. Rapoo**	The determinants of divorce **C. Sikhosana**	Building biostatistics capacity in sub-Saharan Africa **R. Machekano**
10h10 - 10h30	Diagnosing the variability in gas production using support vector machines **A. Mostert**, R. Coetzer and S. Lubbe	Blended learning in large, diverse groups **I. Contardo-Berning**	The marital status of South Africans, a spatial analysis **G. Phakedi**	Bayesian approach towards a better understanding of anemia at pregnancy in Africa **S. Lougue**
10h30 - 11h00	Tea (Eden Grove Foyer)			

11h00 - 11h50	Keynote 2 (Eden Grove Red)			
	Title: Solving some open problems on Brownian areas by applying a new generalization of Euler's Theorem **Presenter:** Prof Jan Swanepoel **Chair:** Prof Sarah Radloff			
11h50 - 12h50	SASA AGM (Eden Grove Red)			
12h50 - 13h50	Lunch (Drostdy Lawns)			
	Parallel Session 5 (13h50 - 15h50)			
	Stream: Distribution Theory **(5A)** **Venue:** EG Red **Chair:** Dr Gretel Crafford	**Stream:** Sampling Theory **(5B)** **Venue:** EG Blue **Chair:** Dr Ariane Neethling	**Stream:** Applied Statistics **(5C)** **Venue:** EG Seminar Room 1 **Chair:** Dr Melisa Koorsse	**Stream:** Young Statisticians **(5D)** **Venue:** EG Seminar Room 2 **Chair:** Dr WD Schutte
13h50 - 14h10	Estimating multivariate normal distributions when relatively few observations are available **N. A. S. Crowther** and N. Strydom	Regression analysis for complex sampling data **A. Neethling**, R. Luus and T. de Wet	Assessing joint spatial autocorrelations between mortality rates due to cardiovascular conditions **T. Darikwa**, M. Lesaoana and S. O. M. Manda	A variability analysis of nanoscale image measurements **A. Haywood**, I. Fabris-Rotelli, S. Das and J. Wesley-Smith
14h10 - 14h30	Parameter estimation by the method of arc lengths with application to the normal distribution **T. Loots** and A. Bekker	Methods in the practice of sample allocation in stratified sampling: Review and perspectives **S. Er** and M. Kozak	Factors affecting the survival of small to medium enterprises in South Africa **I. Mudhombo**, P. Ndlovu and M. A. Managa	The LogNIG distribution and some of its properties **K. Panman** and T. de Wet
14h30 - 14h50	On marginal fiducial inference based on conditional generalized fiducial pivotal quantities **R. Schall**	A study on the apparent randomness of an animal sample **I. Fabris-Rotelli**, C. Kraamwinkel, G. Fosgate, D. Knobel and K. Hampson	Things are getting hotter – climate change in South Africa's National Parks **V. Goodall**, N. van Wilgen, S. Holness, S. Chown and M. McGeoch	CUSUM procedures based on signed sequential ranks **C. van Zy**l and F. Lombard
14h50 - 15h10	On the theory of opportunistic condition-based maintenance (with application) M Shafiee, **M. Finkelstein** and C. Bérenguer	Small area estimation methods for household survey data analysis **M. Kisaka-Lwayo**, N. Mashamba, N. Mogkerepi and C. Mashaba	Preference mapping methods and applications **M. van der Rijst**, T. Næs and N. Muller	Bias reduction studies in non-parametric regression with applications: An empirical approach **M. Cockeran** and C. J. Swanepoel

15h10 - 15h30	Length-biased distributions and their properties **D. Funke**, R. Ehlers, A. Bekker and J. J. J. Roux	Statistical sampling in auditing **L. P. Fatti**	Data visualisation and communication **T. Magadla**, N. Dudeni-Tlhone, J. Holloway, S. Khuluse and R. Koen	Impact of climate change on renewable energy generation in South Africa: Wind energy **E. Ahame** and I. Litvine
15h30 - 15h50	Estimation of the bivariate normal distribution to a two-way contingency table **G. Crafford** and N. A. S. Crowther	A multiple imputation approach for missing air quality measurements **S. Khuluse**, A. Stein and P. Debba	Forecasting sensor alarm and trip values in coal-fired power plants using artificial neural networks **M. Koorsse**, M. C. du Plessis and I. Litvine	Generalized additive modelling of age at first sex in South Africa **M. A. Managa**, O. O. Alaba, J. O. Olaomi and K. Zuma
15h50 - 16h10	Tea (Eden Grove Foyer)			
	Parallel Session 6 (16h10 - 17h50)			
	Stream: Modelling **(6A)** **Venue:** EG Red **Chair:** Dr Morné Lamont	**Stream:** Bootstrap & General **(6B)** **Venue:** EG Blue **Chair:** Dr Leonard Santana	**Stream:** Time Series & Extreme Value Theory **(6C)** **Venue:** EG Seminar Room 1 **Chair:** Dr Morné Sjölander	**Stream:** Statistical Process Control **(6D)** **Venue:** EG Seminar Room 2 **Chair:** Mrs Jenny Holloway
16h10 - 16h30	Comparison of basis function selection and model fitting for functional regression models **S. Lubbe** and R. Essomba	On a data-dependent choice of the tuning parameter appearing in certain goodness-of-fit tests **L. Santana** and J. S. Allison	Time series Thurstonian paired comparisons models **M. Sjölander**	Phase I and Phase II control charts for the variance and general variance **R. van Zyl** and A. J. van der Merwe
16h30 - 16h50	Modelling Likert scale data using the beta distribution **M. Kidd** and N. Laubscher	Multiple imputation in the presence of a detection limit, with applications: An empirical approach **C. J. Swanepoel** and S. C. Liebenberg	A random walk hypothesis test of the South African Rand – US Dollar exchange rate **J. M. Batidzirai**, K. Chinhamu and R. Chifurira	An ARL-unbiased t_r-chart for monitoring times between failures **N. Kumar** and S. Chakraborti
16h50 - 17h10	Using first passage time densities to perform inference on non-linear, multivariate diffusion models **E. A. D. Pienaar**	Efficiency analysis of South African water service authorities using DEA **W. Brettenny**, G. D. Sharp and S. Hosking	Time series modeling of South African GDP **R. Makhwiting**	Optimal design of CUSUM-SIGN chart **N. Chakraborty**, P. Castagliola and S. Chakraborti

17h10 - 17h30	On the application of the CPC model in discriminant analysis **T. Pepler**	Alternative methods to parametric significance testing in linear regression **N. Makhanya**, F. E. Steffens and L. Fletcher	Risk management with generalized hyperbolic distributions: Application to precious metals **K. Chinhamu**, D. Chikobvu and C-K. Huang	Effect of parameter estimation on the performance of the t-chart and t_r-chart **N. Molata**
17h30 - 17h50	Estimating kernel hyper-parameters in kernel Fisher discriminant analysis **M. Lamont**	Hurst exponent for linear regression processes **I. Litvine**	Modeling extreme maximum annual rainfall for Zimbabwe **R. Chifurira** and D. Chikobvu	Identifying clusters and outliers within electricity load profiles **J. Holloway**, R. Koen and R. van der Wath
19h00 - ☺	Gala Dinner (Drosdty Lawns)			

Thursday 30 October 2014				
	Parallel Session 7 (08h30 - 10h50)			
	Stream: Estimation, Biometry & Competition Winners **(7A)** **Venue:** EG Red **Chair:** Mr Frikkie Calitz	**Stream:** Applied Statistics **(7B)** **Venue:** EG Blue **Chair:** Dr Sonali Das	**Stream:** Official Statistics **(7C)** **Venue:** EG Seminar Room 1 **Chair:** Dr Arulsivanathan Naidoo	**Stream:** Modelling & Applied Statistics **(7D)** **Venue:** EG Seminar Room 2 **Chair:** Dr Legesse Debusho
08h30 - 08h50	Fleiss's kappa modified **L. Fletcher**	Analysing groundwater regimes using a Bayesian change point detection approach: Case study from Limpopo **S. Das** and **S. Khuluse**	Under-five mortality in South Africa: The roles of geographic area and socioeconomic status **C. Motsepa**	Multivariate cointegration analysis of employment, inflation and output for short- and long-run linkages **S. Pillay**
08h50 - 09h10	Parameter estimation for a mixture of two univariate Gaussian distributions: A comparative analysis of the proposed and maximum likelihood methods **C. R. Kikawa** and M. Y. Shatalov	Improved confidence intervals for small area estimation under the Fay-Herriot model: Application to food insecurity in Ethiopia **Y. A. Shiferaw**	Local variation of socio-economic inequality in South Africa according to different disparity indices **J-M. V. Hakizimana**	Comparing logistic regression methods for a sparse data set when complete separation is present **M. Botes** and L. Fletcher
09h10 - 09h30	Estimation from ranked set sampling **J. O. Olaomi** and R. Arnab	Analysing key leaf characteristics for spectro-temporal discrimination **N. Dudeni-Tlhone**	Access to basic services in South Africa from 1996 to 2011 **G. Kekana** and A. Naidoo	Geo-additive models for malaria rapid diagnosis test in Ethiopia **D. Ayele**
09h30 - 09h50	Analysis of spatial and temporal patterns of trees affected by guava wilt disease in South African guava orchards **M. Ngwenya** and M. Schoeman	Use of modified stability analysis and best linear unbiased predictor in formulation of recommendation domains for sugarcane varieties **P. M. Njuho** and C. N. Sewpersad	Survival analysis of children under-five: A comparative study between Gauteng and the Rest of South Africa (ROSA) **T. Maremba** and Z. Nukeri	Attitude survey analysis **N. Zondo**
09h50 - 10h10	Incorrect application of replications in agricultural experiments **F. J. Calitz**	Dealing with non-response in the multivariate analysis of categorical data **G. Hendry**	Inequalities in South Africa: A Statistical analysis **L. Masenya**	Mixed effects model for bivariate categorical outcomes using local and continuation ratio logits for the marginal distributions **O. Loquiha**, M. Aerts, N. Hens, E. Martins-Fonteyn and H. Meulemans

10h10 - 10h30	Point-in-time – Through-the-cycle dual calibration **C. Campher** and H. van Rooy *SAS honours project winner*	Longitudinal cohort analysis of student progression in the South African higher education system: The case of University of Limpopo **N. Yibas**	Using geographically weighted regression to explore local crime patterns in South Africa **A. Naidoo**	*A*-optimal designs for two-colour cDNA microarray experiments using the linear mixed effects model **D. B. Gemechu**, L. K. Debusho and L. M. Haines
10h30 - 10h50	Regularised iterative multiple correspondence analysis in multiple imputation **J. Nienkemper-Swanepoel** and M. J. von Maltitz *Statistics SA paper competition winner*	Is "sitting too much" a risk factor for obesity in rural South Africa? **T. Darikwa**, M. Alberts, I. Cook and A. Boateng	Crime statistics in South Africa, a hot spot analysis **Z. Mazibuko**	Optimal factorial designs for two-colour microarray experiments: Properties of admissible designs, *A*-, *D*- and *E*-optimality criteria **L. K. Debusho**, L. M. Haines and D. B. Gemechu
10h50 - 11h20	Tea (Eden Grove Foyer)			
11h20 - 12h10	Keynote 3 (Eden Grove Red)			
	Title: Testing procedures for integer-valued time series **Presenter:** Prof Simos Meintanis **Chair:** Prof James Allison			
12h10 - 12h30	SASA 2014 Closing Ceremony (Eden Grove Red) Chair: Prof James Allison			
12h30	Please remember to take a packed lunch before you leave			

10 List of Poster Presentations

- **Model selection uncertainty and parameter estimation of nonlinear growth models**
  R. Adeyemi
- **Estimating the force of infection from prevalence data: Infectious disease modelling**
  Y. Balakrishna and H. Mwambi
- **R Convenience functions for service course teachers**
  J. S. Baxter
- **Comparison of sampling methods for use with Kriging**
  M. Beckley and S. Kok
- **A study of synthetic Phase II Shewhart-type control charts for monitoring process location**
  T. Chimbwa, M. Graham and S. Chakraborti
- **Eliciting and combining expert opinion – An overview and comparison of methods**
  M. Chinyamakobvu and I. Garisch
- **Using the double-Poisson distribution to analyse Manchester City's 2011/2012 premier league winning season**
  A. Gqwaka and W. Brettenny
- **Techniques for background modelling in image analysis**
  K. Lau and I. Fabris-Rotelli
- **Modelling of multi-state panel data: The importance of the model assumptions**
  T. Mafu and C. Muller
- **Data envelopment analysis as a tool for assessing operations in Eastern Cape ports**
  B. Mienie, G. D. Sharp and W. Brettenny
- **A new bivariate beta model for multiple shifts in a sequential normal process**
  A. Mijburgh, K. Adamksi, A. Bekker and S. Human
- **Modelling summer daily peak loads in South Africa using discrete time Markov chain analysis**
  M. Mokhele and C. Sigauke
- **The South African yellow maize price: A statistical analysis**
  M. Muhl and J. van Niekerk
- **The bootstrapped mean time to survival**
  A. Oppel and T. Loots
- **Biplots for sparse partial least squares**
  O. F. Oyedele and S. Lubbe
- **Matching priors for linear functions of Poisson parameters**
  L. Raubenheimer and A. J. van der Merwe
- **Statistical forecasting modelling for JSE stock prices**
  X. Xotyeni

11 Workshops

Workshop 1

Statistics of extremes: An introduction

M. Ivette Gomes

Department of Statistics and Operations Research, University of Lisbon

---

In *Extreme Value Theory* (EVT) the ordering of a sample is of primordial relevance. Indeed, and more generally in almost all areas of Statistics, the ordering of a univariate sample, as a basis for a clear representation of the sample's content, has been considered highly relevant for a long time. Such an importance enabled us to have nowadays access to a broad statistical methodology related to ordered samples together with associated distributional theory. After a brief reference to a few concepts related to the exact distributional theory of order statistics, we provide some motivation for the need of EVT in the analysis of rare events, in fields as diverse as Environment, Finance, Insurance and Sports, among others. We next provide a few details on the asymptotic behaviour of order statistics. The *general extreme value* (GEV) and the generalized Pareto (GP) distributions are introduced, together with the concepts of *extreme value index* (EVI) and the notion of tail-heaviness, directly related to the Regular Variation Theory. Finally, we shall deal with several topics in the field of *Statistics of Univariate Extremes*, a highly useful area in applications, whenever we want to make inference on the tail (either right or left), estimating rare events' parameters, like high (low) quantiles or return periods of high (low) levels, among others. We shall first review most of the parametric approaches in the area, like the blocks method (BM) or annual maxima method (AMM) and the *peaks over threshold'* (POT) method, and the statistical choice among extremal models and max-domains of attraction. We shall further refer to a few semi-parametric approaches, with the analysis of a few case-studies in the fields of Hydrology, Insurance and Finance, performed essentially through the use of a few R-*Packages for Extreme Values*, like the `evd`, `evdbayes`, `evir`, `ismev`, `extRemes`, `extremevalues`, `fExtremes`, `POT`, and `SpatialExtremes`, among others.

References (basic)

[1] Arnold, B., Balakhrishnan, N. and Nagaraja, H. N. (1992; 2008). *A First Course in Order Statistics.* 1st Ed., Wiley; 2nd Ed., SIAM.

[2] Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004). *Statistics of Extremes: Theory and Applications.* Wiley, England.

[3] Gomes, M. I., Fraga Alves, M. I. and Neves, C. (2013). *Análise de Valores Extremos: uma Introdução.* Edições S.P.E. & I.N.E. (in Portuguese)

Workshop 2

Inference procedures based on the empirical characteristic function

Simos G. Meintanis

Department of Economics, National and Kapodistrian University of Athens

---

1. Introduction, History and Basic Properties of the Empirical Characteristic Function (ECF)
2. The ECF with Simple I.I.D. Data
3. The ECF with Structured I.I.D. Data
4. The ECF with Dependent Data
5. The Probability Weighted ECF (PWECF)

---

12 Abstracts

12.1 Keynotes

Keynote 1

Penultimate approximations and reliability of large coherent systems: Past, present ... and future?

M. Ivette Gomes

Department of Statistics and Operations Research, University of Lisbon

---

The rate of convergence of the sequence of linearly normalized maxima or minima to the corresponding non-degenerate *extreme value* (EV) limiting distribution for maxima ($\mathrm{EVD_M}$) or for minima ($\mathrm{EVD_m}$) is a relevant problem in the field of *extreme value theory*. Fisher and Tippett (1928) observed that, for normal underlying parents, if we approximate the distribution of the suitably linearly normalized sequence of maxima not by the limiting Gumbel distribution EV_0, but by a sequence of other EV distributions, EV_ξ, $\xi = \xi_n = o(1) < 0$, converging thus to EV_0, the approximation is asymptotically improved. Note that ξ is the so-called EV index for maxima. Such approximations are usually called penultimate approximations and have been theoretically studied from different perspectives. For a recent but short overview of the subject see Beirlant *et al.* (2012). Quite recently, this same topic has been revisited in the field of reliability, where any coherent system can be represented as either a series-parallel, a series structure with components connected in parallel, or a parallel-series system, a parallel structure with components connected in series (see Barlow and Proschan, 1975). Its lifetime can thus be written as the minimum of maxima or the maximum of minima. For large-scale coherent systems it can be sensible to assume that the number of system components goes to infinity. Then, the possible non-degenerate EV laws, $\mathrm{EVD_M}$ and $\mathrm{EVD_m}$, are eligible candidates for the finding of adequate lower and upper bounds for such a reliability, a topic developed in Reis and Canto e Castro (2009), among others. However, just as mentioned above, such non-degenerate limit laws are better approximated by an adequate penultimate distribution in most situations. It is thus sensible to assess both theoretically and through Monte-Carlo simulations the gain in accuracy when a penultimate approximation is used instead of the ultimate one, as performed in Reis *et al.* (2013+). Moreover, researchers have essentially considered penultimate approximations in the class of $\mathrm{EVD_M}$ or $\mathrm{EVD_m}$, but we can easily consider a much broader scope for that type of approximations, and such a type of models surely deserves a deeper consideration under statistical backgrounds. Penultimate models seem to be possible and interesting alternatives to the classical models but have never been deeply used in the literature.

References

Barlow A. A. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing: Probability Models.* Holt, Rinehart and Winston, Inc., USA.

Beirlant, J., Caeiro, F. and Gomes, M. I. (2012). An overview and open research topics in the field of statistics of univariate extremes. *Revstat* 10:1, 1-31.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distributions of the largest or smallest member of a sample. *Proc. Camb. Phil. Soc.* 24, 180-190.

Reis, P. and Canto e Castro, L. (2009). Limit model for the reliability of a regular and homogeneous series-parallel system. *Revstat* 7:3, 227-243.

Reis, P., Canto e Castro, L., Dias, S. and Gomes, M. I. (2013+). Penultimate Approximations in Statistics of Extremes and Reliability of Large Coherent Systems. *Methodology and Computing in Applied Probability*, DOI: 10.1007/s11009-013-9338-7

Keynote 2

Solving some open problems on Brownian areas by applying a new generalization of Euler's Theorem

Jan W. H. Swanepoel

Department of Statistics, North-West University, Potchefstroom Campus

In this talk we restrict ourselves to versions of a standard Brownian motion process $\{B(t), 0 \leq t \leq T\}$ and a standard Brownian bridge process $\{B^0(t), 0 \leq t \leq T\}$ defined on a finite interval $[0, T]$. Consider the random Riemann integrals, which often occur in practice, $A(t) := \int_0^t h(s)B(s)ds$ and $A^0(t) := \int_0^t h(s)B^0(s)ds$, for some continuous deterministic function $h : [0, T] \to \mathbb{R}$. It is shown that for certain choices of $h(s)$, closed-form expressions can be derived for these integrals by applying the Karhunen-Loéve expansions of Brownian motion and Brownian bridge processes. This enables one to study the nature of the sample paths of $A(t)$ and $A^0(t)$. More importantly, the exact distributions of $A(t)$ and $A^0(t)$ can be determined without resorting to Itô stochastic integral calculus. This is accomplished by using a newly derived generalization of a theorem by Euler. Interesting conclusions are made, for example, if $h(s) \equiv 1$, then $\mathrm{Var}(A(T))/\mathrm{Var}(A^0(T)) = 4$ for all T, and if $h(s) = s$, then $\mathrm{Var}(A(T))/\mathrm{Var}(A^0(T)) = 6$ for all T. Finally, some general formulas regarding Brownian areas are discussed.

Keynote 3

Testing procedures for integer-valued time series

Simos G. Meintanis

Department of Economics, National and Kapodistrian University of Athens

We consider goodness-of-fit tests and change-point detection procedures in the context of integer-valued time series. Our focus is on Integer Autoregressive (INAR) models and Integer Autoregressive Conditionally Heteroskedastic (INARCH) models. The procedures are constructed by making use of the empirical probability generating function.

12.2 Oral Presentations (Surnames A - J)

Distribution of the aggregate amount of benefit

Franck Adekambi

University of Johannesburg

Abstract: Ramsay (1984) considered a model where the policyholder pays a premium at a constant rate while healthy and upon sickness, the premium ceases and if the duration of the sickness exceeds a certain waiting period , the insurance company pays out a benefit continuously at a constant rate throughout the remaining sickness duration. He used an alternating renewal process to model the length of the periods of health and sickness and derive the asymptotic ruin probability.
Adékambi (2011) consider the same model as Ramsay (1984) with some extensions: first in the contrary of Ramsay (1984) the effect of interest rate is taking into account. Second, while Ramsay (1984) considers a constant rate for the benefit paid, we will allow random rates. Also, the assumption made by Ramsay of the premium ceasing upon sickness is dropped as not only it is not necessary here but also it does not correspond to the practice in health insurance. Adékambi (2011) then, derive an expression for the aggregate amount of benefit paid out up to the end of the sickness period, and even for a given time as well as the first two moments. In this paper, the expression of the cumulative function and the moment generating function of the aggregate amount of benefit paid out up to the end of the sickness period, are given. But from practical point of view these two expressions are difficult to evaluate. Then, an approximation of the distribution of the aggregate amount of benefit paid out up to the end of the sickness period, is given for the case of constant force of interest.
Some justifications of the approximation are given. Illustrations are presented for a constant force of interest and when both the length of healthy and sickness periods have exponential distributions with different parameters. There is no doubt that knowing the distribution of the aggregate amount of benefit paid out is useful for pricing, valuation, solvency and reinsurance.
Key words: Alternating renewal process, Constant force of interest, Discounted aggregate amount of benefit, Discounted aggregate sums of the premium received, Moments
Session: 1D, Tuesday 11h30 - 12h10, Eden Grove Seminar Room 2

Integral equations for moments in multistate life and health insurance models

Franck Adekambi[1,⋆] and Marcus C. Christiansen[2]

[1]University of Johannesburg, [2]University of Ulm

Abstract: For the risk management in life and health insurance, a mathematical key quantity is the probability distribution of the random future liabilities of an insurance contract. We focus on modeling the randomness of the future health status of individual policyholders by Semi-Markovian multistate models. We derive integral equations for the moment generating function and higher order conditional moments of the future liabilities. From the moments we can then construct approximations for the loss distribution. Furthermore, we discuss how to represent the loss distribution directly via an integral equation.
Key words: Counting process, Higher order conditional moments, Integral, Multistate life insurance, Semi-Markov model
Session: 1D, Tuesday 11h30 - 12h10, Eden Grove Seminar Room 2

On the use of a semi-parametric estimator for censored data, with applications in food safety exposure assessment

Marc Aerts*, Ruth Nysen and Christel Faes

Hasselt University

Abstract: Censored data are often encountered in medical and public health studies. In this presentation we focus on modelling concentration data as part of a probabilistic food safety exposure assessment. Such concentration data are censored by the limit of detection and the limit of quantification.
We advocate the use of a semi-parametric model, based on the SNP density estimator of Fenton and Gallant (*Journal of Econometrics*, 1996; see also Gallant and Nychka, *Econometrica*, 1987). More precisely we illustrate how the smooth semi-parametric model of Zhang and Davidian (*Biometrics*, 2008), who built further on this estimator in the context of censored time-to-event data, can also be applied for estimation, model selection and model averaging, as well as for formal lack-of-fit testing in the setting of food safety exposure assessment.
The methodology is illustrated on data of cadmium concentrations, and the performance characteristics of the semi-parametric approaches are illustrated in a variety of simulation settings.
Key words: Censored data, Lack-of-fit test, Model averaging, Model selection, Semi-parametric estimation
Session: 3A, Tuesday 16h10 - 17h50, Eden Grove Red

Impact of climate change on renewable energy generation in South Africa: Wind energy

Edmund Ahame* and Igor Litvine

Nelson Mandela Metropolitan University

Abstract: Renewable energy (RE) is taking centre stage in the world's endeavour to curb climate change, increase clean energy in the mix and sustainable development. It is feared the climate change may alter the usual weather patterns (wind, irradiation, precipitation etc.) on which most renewable energy technologies depend. In the case of wind energy, the energy contained in the wind is proportional to the cube of wind speed, which means that alterations in the later can have significant impacts on the former (Pryor and Barthelmie, 2010, Cradden, Harrison and Chick, 2012). It is therefore imperative to investigate the impact of climate change on renewable energy potential. In literature General Circulation Models (GCMs), used to represent physical processes in the atmosphere, ocean, cryosphere and land surface, are the most advanced tools currently available for simulating the response of the global climate system to increasing greenhouse gas concentrations (IPCC 2007, Nolan *et al.,* 2011, Sailor *et al.,* 2008, Breslow and Sailor 2002). Regional circulation model (RCMs) have been used to downscale the coarse information provided by the GCMs and provide high resolution information for a particular region of interest.
The aim of this study is to investigate the impact of climate change on renewable energy potential in South Africa with emphasis on wind energy. In this project relevant RCMs will be used, and/or new statistical downscaling models, to downscale the GCMs. The simulations will be based on IPCC emission scenarios which best suit the South Africa region. From these results on the effect of climate change on climatology such as wind patterns we will be able to deduce the impact on wind energy generation.
Key words: Climate change, Renewable energy generation, Wind energy
Session: 5D, Wednesday 13h50 - 15h50, Eden Grove Seminar Room 2

Geo-additive models for malaria rapid diagnosis test in Ethiopia

Dawit Ayele

University of KwaZulu-Natal

Abstract: Background
The transmission of malaria is the leading public health problem in Ethiopia. From the total area of Ethiopia, more than 75% is malarious. The aim of this study was to identify socio-economic, geographic and demographic risk factors of malaria based on the rapid diagnosis test (RDT) survey results and produce the prevalence map of the area illustrating variation in malaria risk.
Methods
This study accounts for spatial correlation in assessing the effects of socio- economic, demographic and geographic factors on the prevalence of malaria in Ethiopia. A total of 224 clusters of about 25 households each were selected from the Amhara, Oromiya and Southern Nation Nationalities and People's (SNNP) regions of Ethiopia. A generalized linear mixed model with spatial covariance structure was used to analyse the data where the response variable was the presence or absence of malaria using the RDT.
Results
The results showed that households in the SNNP region were found to be at more risk than Amhara and Oromiya regions. Moreover, households which have toilet facilities, clean drinking water and a greater number of rooms and mosquito nets in the rooms, have less chance of having household members testing positive for RDT. Moreover, from this study, it can be suggested that incorporating spatial variability is necessary for understanding and devising the most appropriate strategies to reduce the risk of malaria.
Key words: Mixed model, Rapid diagnostic test, Spatial statistics
Session: 7D, Thursday 08h30 - 10h50, Eden Grove Seminar Room 2

Stepwise latent class analysis with continuous distal outcome variables

Zsuzsa Bakk* and Jeroen K. Vermunt

Tilburg University

Abstract: Latent class (LC) analysis is an approach used to create a typology or clustering based on a set of observed variables; that is to classify units into a small set of LCs. In most situations the interest lies not only in creating a clustering, but also in using the clustering for prediction of a distal outcome variable. In cases of a continuous distal outcomes this boils down to estimating the class-specific mean of the distal outcome. Until recently estimation of this type of models was very difficult, due to the lack of unbiased stepwise estimators. We provide an overview of recently developed bias-adjusted stepwise approaches to latent class modeling with continuous distal outcomes. We investigate the robustness of these methods to violations of underlying model assumptions by means of a simulation study. While each of the three investigated methods yield unbiased estimates of the class-specific means of distal outcomes when the underlying assumptions hold, some of the methods may fail to different degrees when assumptions are violated. Based on our study, we provide recommendations on which method to use under what circumstances.
Key words: Continuous distal outcome, Latent class analysis, Robustness, Stepwise procedure
Session: 4A, Wednesday 08h30 - 10h30, Eden Grove Red

A random walk hypothesis test of the South African Rand – US Dollar exchange rate

Jesca M. Batidzirai*, Knowledge Chinhamu and Retius Chifurira

University of KwaZulu-Natal

Abstract: This study tests the random walk hypothesis for the log-differenced daily USD-ZAR exchange rate. We utilize the conventional variance ratio test (Chow and Denning, 1993), the multiple variance ratio tests (Lo and MacKinlay, 1988), Wright's ranks and sign test and the Whung and Kim tests to examine the validity of the random walk hypothesis in the South African exchange. Unit root tests, the Augmented Dickey Fuller test (ADF), Phillips Perron (PP) and the KPSS tests are utilized on the daily nominal USD-ZAR exchange rates and exchange rate returns. The empirical results show that random walk hypothesis is rejected and provides support for the violation of the weak-form market-efficiency hypothesis.
Key words: Market-efficiency, Random walk hypothesis, Unit root tests, Variance ratio tests
Session: 6C, Wednesday 16h10 - 17h50, Eden Grove Seminar Room 1

Real-time educational interpreting in Statistics

Tom Berning

Stellenbosch University

Abstract: The issue of language in education has been a contentious issue in most developing countries. This is due to the critical role that language and communication play in the learning process. In order to offer second language learners the opportunity to learn in their language of choice, a project has been launched to promote multilingualism. The institution offers real-time educational interpreting (in the form of simultaneous whisper interpreting between Afrikaans and English) to statistics students in various undergraduate years of study. This offers an additional advantage that lecturers are also able to communicate in their home language. The purpose of this paper is to evaluate educational interpretation from the student and the lecturer's perspectives as well as provide a brief overview of the technology employed.
Key words: Education, Interpreting
Session: 1B, Tuesday 11h30 - 12h10, Eden Grove Blue

Comparing logistic regression methods for a sparse data set when complete separation is present

Michelle Botes[1,2*] and Lizelle Fletcher[2]

[1]Lightstone, [2]University of Pretoria

Abstract: An occurrence which is sometimes observed in a model based on dichotomous dependent variables is separation in the data. Separation in the data occurs when one or more of the independent variables can perfectly predict some binary outcome and it primarily happens in small samples. There are three different mutually exclusive and exhaustive classes into which the data from a logistic regression can be classified: complete separation, quasi-complete separation and overlap.
Separation (either complete or quasi-complete) in the data gives rise to a number of problems since it implies infinite or zero maximum likelihood estimates which are unrealistic and does not happen in practice. In this paper different methods to deal with complete separation will be investigated when only continuous independent variables are considered.
Key words: Complete separation, Continuous independent variables, Logistic regression, Sparse data
Session: 7D, Thursday 08h30 - 10h50, Eden Grove Seminar Room 2

Efficiency analysis of South African water service authorities using DEA

Warren Brettenny*, Gary D. Sharp and Stephen Hosking

Nelson Mandela Metropolitan University

Abstract: In recent years South Africa has experienced numerous service delivery protests from the public. These protests are a result of the lack of delivery of services such as water and sanitation (amongst others). To address this, local governments have taken part in national benchmarking initiatives (Blue Drop, Green Drop) in an effort to improve the quality of potable water and sanitation services. These initiatives focus on the quality of water delivered and not on the efficiency with which this delivery achieved. This study uses data envelopment analysis (DEA), along with related bootstrapping techniques, to evaluate the efficiency with which several South African water service authorities (WSAs), including both metropolitan and local municipalities, provide water services to the public.

Key words: Data envelopment analysis (DEA), Efficiency, South Africa, Water services

Session: 6B, Wednesday 16h10 - 17h50, Eden Grove Blue

Bayesian non-linear mixed effects regression model for the characterisation of early bactericidal activity of tuberculosis drugs

Divan A. Burger[1,2,*] and Robert Schall[1,2]

[1]Quintiles, [2]University of the Free State

Abstract: Trials of the early bactericidal activity (EBA) of tuberculosis (TB) treatments assess the decline, during the first few days to weeks of treatment, in colony forming unit (CFU) count of Mycobacterium tuberculosis in the sputum of patients with smear-microscopy-positive pulmonary TB. Profiles over time of CFU data have conventionally been modelled using linear, bi-linear or bi-exponential regression. We propose a new bi-phasic non-linear regression model for CFU data that comprises linear and bi-linear regression models as special cases, and is more flexible than bi-exponential regression models. A Bayesian non-linear mixed effects (NLME) regression model is fitted jointly to the data of all patients from a trial. The posterior predictive distribution of relevant slope parameters of the model allows one to judge whether treatments are associated with mono-linear or bi-linear decline of log(CFU) count, and whether CFU count initially decreases fast, followed by a slower rate of decrease, or vice versa. Compound Bayes factors are reported for comparing our bi-phasic model with the bi-linear model. In addition, robust models with skew Student t distributed random effects and residuals are investigated.

Key words: Bayesian non-linear mixed effects (NLME) regression model, Bi-phasic, Colony forming unit (CFU) count, Compound Bayes factors, Early bactericidal activity (EBA), Skew distributions, Tuberculosis (TB)

Session: 3D, Tuesday 16h10 - 17h50, Eden Grove Seminar Room 2

Incorrect application of replications in agricultural experiments

Frikkie J. Calitz

Agricultural Research Council

Abstract: The purpose of this paper is to demonstrate how experimental replications are applied incorrectly when performing so called statistical agricultural experiments. The experimental layout is often not statistically defendable as a result of the misunderstanding of the definition of a statistical experimental replication and incorrect application of randomisation.

One of the most common errors made in agricultural experimentation is that experimental units are simply subdivided into sub-plots and then considered as experimental replications (also called random replications). The following experiment demonstrates a typical example.

An investigator wanted to test the effect of high and low protein ration feeding with two sows on the growth rate of the piglets. They measured the growth rate of 10 piglets feeding on each sow as random replications and the data was analysed as a complete randomised design (CRD).
Similar incorrect applications of statistical replication and randomisation occur in agricultural experiments which will be demonstrated with practical examples and the problem with the interpretation of such experiments will be discussed. When analysing data from the above mentioned "so-called statistical randomised trials" statistical analysis programs will accept the data as entered (with the wrong replications). The calculated error is sampling error and not experimental error. It is expected that the sample variation will be much smaller than the experimental variation. Using sampling error to compare treatment means will result in significant differences between treatments that may not be true and thus misinterpretation of results.
It is recommended that consulting statisticians be aware of this incorrect application of replications and of the importance of familiarising themselves with the experimental layout before analysing data.
Key words: Experimental replications
Session: 7A, Thursday 08h30 - 10h50, Eden Grove Red

Point-in-time – Through-the-cycle dual calibration

Cornu Campher [1,2,⋆] and Herrie van Rooy[3]

[1]ABSA, [2]University of the Witwatersrand, [3]Nedbank

Abstract: Given a probability of default of a business which is conditional on various factors such as the quality of management of a business or the profit after tax of a business, this talk gives a manner in which this probability can be converted to a probability that is either stable through the macroeconomic cycle or tracks the cycle closely. A multivariate time series is built in order to predict a credit index which is based on business insolvencies. This model is built using various macroeconomic variables such as GDP Growth, Employment Growth and Debt Service Cost. Predicted values of the credit index are used in order to perform the conversions to a stable probability of default or a cyclical probability of default.
Key words: Credit risk, Distance to default, Multivariate time series, Probabilty of default, Vector autoregressive moving average models
Session: 7A, Thursday 08h30 - 10h50, Eden Grove Red

Optimal design of CUSUM-SIGN chart

Niladri Chakraborty[1⋆], Philippe Castagliola[2] and Subha Chakraborti[3]

[1]University of Pretoria, [2]Université de Nantes, [3]University of Alabama

---

Abstract: CUSUM-Sign (CUSUM-SN) charts are efficient alternatives to Shewhart type Sign charts for detecting small sustained shifts, for individuals or subgroup data. This efficiency, however, depends on a proper design of the CUSUM-SN chart which requires to search for the optimal combination of the chart parameters (the reference parameter, k and the decision interval, H). In this paper we examine the problem of finding the optimal combination of the chart parameters (k, H) of a CUSUM-SN chart given a specified value for the location shift and a nominal value of the in-control average run-length. The optimal (k, H) combinations obtained for the normal distribution are examined for their robustness to non-normality. An extensive study, involving symmetric and asymmetric families of distributions, shows that the optimal (k, H) combinations obtained when assuming normality perform reasonably well for many non-normal symmetric distributions although the kurtosis plays an important role. However, these optimal (k, H) combinations do not seem to work as well for skewed distributions.
Key words: Average run length (ARL), CUSUM control chart , Nonparametric, Optimal design, Robustness
Session: 6D, Wednesday 16h10 - 17h50, Eden Grove Seminar Room 2

Modeling extreme maximum annual rainfall for Zimbabwe

Retius Chifurira[1,⋆] and Delson Chikobvu [2]

[1]University of KwaZulu-Natal, [2]University of the Free State

Abstract: In this paper the generalized extreme value distributions are fitted to the mean annual rainfall to describe the extremes of rainfall. Extreme Value Theory (EVT) is used to estimate the probabilities of meteorological floods. The maxima distribution is used to fit the generalized extreme value distribution to the data and find probabilities of extreme high levels of mean annual rainfall. Anderson-Darling goodness-of-fit test shows the simpler generalized extreme value model, of the Gumbel distribution is of good fit. We explore the possibility of trends in the data; results indicate that there are no such trends. The yearly mean return level estimates are derived and return level of 1 193 mm (recorded high mean annual rainfall amount) is associated with a mean return period of around 300 years. This paper provides the first application of extreme value distributions to mean annual rainfall from a drought prone country such as Zimbabwe.
Key words: Extreme, Generalized extreme value distribution, Maxima, Return level, Time trend
Session: 6C, Wednesday 16h10 - 17h50, Eden Grove Seminar Room 1

Risk management with generalized hyperbolic distributions: Application to precious metals

Knowledge Chinhamu[1,⋆], Delson Chikobvu[2] and Chun-Kai Huang[3]

[1]University of KwaZulu-Natal, [2]University of the Free State, [3]University of Cape Town

Abstract: In the context of financial forecasting and risk management, the accuracy in modelling the underlying returns distribution plays an important role. For instance, risk management tools such as Value-at-risk (VaR) are highly dependent on the underlying distributional assumption, with particular focus being placed at the extreme tails. Hence, identifying a distribution that may best capture all aspects of the given financial data may provide vast advantages to both investors and risk managers. In this paper, we investigate to establish the best Generalized Hyperbolic Distribution (GHD) models to fit the precious metals (gold and platinum) returns. The relative adequacy and goodness-of-fit of these distributions are then assessed through the robustness of their respective value-at-risk (VaR) estimates. The model selection methodologies utilized in our analyses include the Anderson Darling test and the Akaike information criterion. Backtesting on the adequacy of VaR estimates is also performed using the Kupiec likelihood ratio test. The models' performances in VaR estimation and results from the Kupiec Likelihood test show the best GHD model for VaR estimation differ at different levels, for both gold and platinum returns.
Key words: Akaike information criterion, Anderson Darling test, Generalized hyperbolic distributions, Kupiec test, Value-at-Risk
Session: 6C, Wednesday 16h10 - 17h50, Eden Grove Seminar Room 1

Efficiency analysis of South African tertiary education institutions using data envelopment analysis

Ignatious Chitekedza[⋆], Warren Brettenny and Johan Hugo

Nelson Mandela Metropolitan University

Abstract: With an increasing number of students enrolling at higher education institutes institutions, it has become important to investigate whether these institutions are using their resources adequately. This study uses data envelopment analysis (DEA) to estimate the efficiency of 23 South African tertiary education institutions based on both research and teaching outputs. Using DEA we are able to rank South African universities according to their use of resources in these two areas. These rankings can identify institutions which are performing well and also those which require improvement. Owing to the limited sample size,

variable reduction techniques, including the efficiency contribution measure (ECM) and principal components analysis (PCA-DEA), were used to improve the discrimination of this analysis.
Key words: Data envelopment analysis, Efficiency, South African Universities, Variable reduction
Session: 2D, Tuesday 14h00 - 15h40, Eden Grove Seminar Room 2

A variational Bayes approach to analysing site-occupancy models

Allan Clark[1,⋆], Res Altwegg[1] and John Ormerod[2]

[1]University of Cape Town, [2]University of Sydney

Abstract: Presence-absence data is often used to investigate the range and range shifts of animal populations using occupancy models that include covariates assumed to describe the probability of species occurrence as well as covariates assumed to determine the probability of species detection. Bayesian implementations of these models are often analysed using WinBugs or OpenBugs as one requires the use of Markov chain Monte Carlo (MCMC) methods to sample from the posterior distribution of the required parameters. With large data sets, MCMC can be very time consuming. In this article we develop a Variational Bayes (VB) approximation to the posterior distribution of the parameters of a multi site-occupancy model. This is accomplished through the development of a fast iterative algorithm that does not use MCMC methods. Simulations demonstrate the effectiveness of the proposed technique.
Key words: Bayesian modelling, Ecology, Occupancy model, Variational Bayes
Session: 2A, Tuesday 14h00 - 15h40, Eden Grove Red

Comparative assessment of radiation data for photovoltaic plant applications

Chantelle Clohessy⋆, Gary D. Sharp, Ernest van Dyk and Frederik Vorster

Nelson Mandela Metropolitan University

Abstract: Global demand for energy is growing beyond the limits of installed generation capacity (Devabhaktuni *et al.*, 2013). To meet the future demands alternative energy sources need to be investigated. Research investigating Photovoltaic (PV) power generation is one of these alternative energy systems.
This study investigates a solar resource used to obtain data for energy based applications. The resource, Meteonorm is a meteorological database that contains climatological data for solar applications. The results are stochastically generated for a year from observed or interpolated data. These results represent an average year from selected climatological time period based on the users' settings. As such the results do not represent a real historic year but represents a typical statistically generated year at the selected location.
This paper evaluates the accuracy of the database from Meteonorm by comparing the results to actual measured data from the Centre of Energy Research at the Nelson Mandela Metropolitan University.
References
Devabhaktuni, V., Alam, M., Depuru, S., Green, R., Nims, D. and Near, C. (2013). Solar energy: Trends and enabling technologies. *Renewable and Sustainable Energy Reviews*, 19, 555-564.
Key words: Comparative assessment, Energy, Photovoltaic
Session: 3C, Tuesday 16h10 - 17h50, Eden Grove Seminar Room 1

Bias reduction studies in non-parametric regression with applications: An empirical approach

Marike Cockeran* and Cornelia J. Swanepoel

North-West University

Abstract: The effect of three improvement methods on recently developed and well-established nonparametric kernel regression estimators, is investigated. The improvement methods are applied to the Nadaraya-Watson estimator with cross-validation bandwidth selection as well as with plug-in bandwidth selection, a local linear estimator with plug-in bandwidth selection and a bias-corrected nonparametric estimator proposed by Yao (2012), using cross-validation bandwidth selection. The different resulting regression estimators are evaluated by minimising the mean integrated squared error (MISE).

Two improvement methods are bootstrapped-based. Bagging is an acronym for bootstrap aggregating and is primarily a variance reduction tool. Bragging stands for bootstrap robust aggregating. A robust estimator is obtained by using the sample median over bootstrap estimates instead of the sample mean as in bagging.

Boosting is the third general method for improving the accuracy of any given estimator and aims to reduce the bias component of the estimator. The method starts off with a sensible estimator and improves iteratively, based on its performance on a dataset.

A selection of algorithms, involving the various improvement methods, estimators, models to be estimated, parametric set-ups and bandwidth selection methods (also involving various bootstrap strategies) are presented, displaying clear steps in terms of handling the data to obtain the goals of the study. Results are briefly summarized and illustrated, followed by conclusions and remarks regarding computational problems such as grid issues and bandwidth selection issues.

Key words: Bagging, Boosting, Bragging, Cross-validation bandwidth, Kernel regression estimators, Plug-in bandwidth

Session: 5D, Wednesday 13h50 - 15h50, Eden Grove Seminar Room 2

Blended learning in large, diverse groups

Ivona Contardo-Berning

Stellenbosch University

Abstract: Information and communication technologies have the potential to transform education in developing countries from the traditional methodologies to a more modern and flexible learning environment. The application of information and communication technologies can play a significant role in improving the learning experience of students. In our introductory statistics courses we encounter large, diverse, heterogeneous groups of students from various disciplines and backgrounds which create some specific challenges. We would like to challenge and stimulate strong students while providing students with a weaker background with additional learning opportunities and materials. Blended learning offers us a unique opportunity to create a blend of face-to-face and technology enhanced learning opportunities to cater for these diverse groups of students.

Results are presented from a pilot study using blended learning for a large first year service course in statistics.

Key words: Blended learning, Education

Session: 4B, Wednesday 08h30 - 10h30, Eden Grove Blue

Estimation of the bivariate normal distribution to a two-way contingency table

Gretel Crafford* and Nico A. S. Crowther

University of Pretoria

Abstract: A bivariate normal distribution is fitted to a two-way contingency table with an underlying bivariate normal distribution. The maximum likelihood estimators of the bivariate normal distribution are obtained iteratively by making use of the maximum likelihood estimation procedure under constraints of Matthews and Crowther (1995). Since all five parameters of the bivariate normal distribution are estimated, statistical inferences are possible with regard to the marginal as well as the partial distributions. The estimation of the correlation coefficient is used to great effect to describe the relationship between the two variables. The calculation of the correlation coefficient is done by implementing Sheppard's theorem on median dichotomy, which is based on the volumes of the four quadrants of the bivariate normal distribution.
Key words: Bivariate normal distribution, Contingency table, Correlation coefficient, Grouped data, Maximum likelihood estimators, Multinomial distribution
Session: 5A, Wednesday 13h50 - 15h50, Eden Grove Red

Mixed effects modeling in conjoint analysis

Tanita Cronje[1,*], Frans H. J. Kanfer[1], Solly M. Millard[1] and Mohammad Arashi[1,2]

[1]University of Pretoria, [2]Shahrood University of Technology

Abstract: In this business driven world conjoint analysis is a powerful tool that companies can employ to efficiently model consumer choices and preferences. A mixed effects model is proposed to perform a conjoint analysis with normal responses, illustrated by an application of modeling respondent's preferences to different industrial detergents. The proposed model allows for predicting how observed attributes (which describes a product in terms of its characteristics and features) of decision makers and choice options, influence decisions. Inference regarding the parameters of the proposed model with a normal distribution is discussed in the mixed effects conjoint setting. Extensions of this model, regarding Bayesian prior selection, are also studied.
Key words: Conjoint analysis, Hierarchical Bayes, Mixed effect model, Variance-covariance structures
Session: 3C, Tuesday 16h10 - 17h50, Eden Grove Seminar Room 1

Estimating multivariate normal distributions when relatively few observations are available

Nico A. S. Crowther* and Nina Strydom

University of Pretoria

Abstract: The fundamental problem addressed is that of estimating mean vectors and covariance matrices of multivariate normal distributions, if relatively less observations than parameters are available. In such cases constraints on the parameters have to be introduced in a logical and acceptable way. These constraints are usually suggested by the underlying experimental design, as well as the nature of the data. An extreme case is where a number of multivariate populations are sampled with only one observation per population. Specific recommendations for dealing with these scenarios are presented.
Key words: Multivariate normal distributions, Parameter constraints
Session: 5A, Wednesday 13h50 - 15h50, Eden Grove Red

Assessing joint spatial autocorrelations between mortality rates due to cardiovascular conditions

Timotheus Darikwa[1,⋆], Maseka A. Lesaoana[1] and Samuel O. M. Manda[2]

[1]University of Limpopo, [2]Medical Research Council

Abstract: South Africa has a very high burden of non-communicable diseases (NCDs). It is well known that NCDs share the same risk factors and thus may occur together geographically. Researchers tend to focus on estimating disease risks and how these relate to putative risk factors. Often the extent of the geographical autocorrelation in the occurrence of diseases is ignored. This study uses recently developed multivariate spatial clustering measures to determine spatial autocorrelation between rates of mortality due to common cardiovascular diseases at the municipal level in South Africa. Data from the national deaths notification register for 2011 are used. Both global and local measures of spatial associations were derived for each municipality. Local significance maps were used to determine local clusters of joint spatial associations between mortality risks attributable to common cardiovascular diseases in South Africa.

Key words: Cardiovascular diseases, Mortality, Multivariate, Non-communicable diseases, Spatial autocorrelations

Session: 5C, Wednesday 13h50 - 15h50, Eden Grove Seminar Room 1

Is "sitting too much" a risk factor for obesity in rural South Africa?

Timotheus Darikwa⋆, Marianne Alberts, Ian Cook and Alex Boateng

University of Limpopo

Abstract: There is mounting evidence that physical activity alone is not enough to prevent obesity without considering one's sedentary lifestyle. This paper seeks evidence in a South African rural area on the impact, among other factors, of reported waking time spent seated or in an inclined position on obesity. A response dichotomous variable with a value of one for the overweight/obese and zero otherwise was created and used in bivariate and multivariate logistic regression analysis to determine any association between overweight/obese and the individual known associated risk factors of age, gender, educational level, income, employment status, physical activity, marital status and time spent seated or in a reclined position. The bivariate test of independence suggest that obesity is significantly associated with age, gender, occupation, village of origin and marital status (p-value < 0.05) while physical activity and level of income were insignificant. The best multivariate logistic regression model had age, sedentary time, marital status and gender as significant independent explanatory variables. Sedentary time (sitting time) is highly significant and may play a crucial role in obesity control.

Key words: Multivariate, Obesity, Overweight, Rural, Sedentary

Session: 7B, Thursday 08h30 - 10h50, Eden Grove Blue

Analysing groundwater regimes using a Bayesian change point detection approach: Case study from Limpopo

Sonali Das⋆ and Sibusisiwe Khuluse⋆

Council for Scientific and Industrial Research

Abstract: In this paper, we investigate groundwater data from a Limpopo region using measurements from boreholes. We apply the Barry and Hartigan Bayesian change point algorithm to detect possible breaks, or regime shifts. The algorithm not only picked up obvious change points in the data series, but more importantly also picked up subtle change points. Our findings are: first, that there is strong evidence that there existed a partitioned aquifer system, which later seemed to have merged into one aquifer; second, that each of the

regime change locations, whether pronounced or subtle, were preceded by known drought events. This is an important investigation using an objective Bayesian change point detection algorithm to understand aquifer systems in regions where groundwater is a vital and a scarce resource.
Key words: Bayesian, Change point, Groundwater
Session: 7B, Thursday 08h30 - 10h50, Eden Grove Blue

Extreme quantile estimation by using limited historical data and scenario assessments

Riaan de Jongh[1,⋆], Tertius de Wet[2], Helgard Raubenheimer[1] and J. Hennie Venter[1]

[1]North-West University, [2]Stellenbosch University

Abstract: Many banks use the loss distribution approach in their advanced measurement models to estimate economic capital. This boils down to estimating the 99.9% VaR of the aggregate loss distribution and is notoriously difficult to do accurately. Also, it is well-known that the accuracy with which the tail of the loss severity distribution is estimated is the most important driver in determining a reasonable estimate of economic capital. To this end, banks use internal data and external data (jointly referred to as historical data) as well as scenario assessments in their endeavour to improve the accuracy with which the severity distribution is estimated. In this paper we propose a simple new method whereby the severity distribution may be estimated using historical data and experts' scenario assessments jointly. The way in which historical data and scenario assessments are integrated incorporates measures of agreement between these data sources, which can be used to evaluate the quality of both. In particular we show that the procedure has definite advantages over traditional methods where the severity distribution is modelled and fitted separately for the body and tail parts, with the body part based only on historical data and the tail part on scenario assessments.
Key words: Extreme quantile estimation, Historical data, Scenario assessments, Severity distribution
Session: 2B, Tuesday 14h00 - 15h40, Eden Grove Blue

Time series outlier detection using singular spectrum analysis and robust principal component analysis

Jacques de Klerk

North-West University

Abstract: Singular Spectrum Analysis (SSA) is a powerful non-parametric time series technique with wide application in time series analysis. SSA is particularly powerful for time series exhibiting seasonal variation with/without trend components. SSA can be applied to time series found in market research, economics, meteorology and oceanology, to name but a few. SSA places a univariate time series into a multivariate framework by unfolding into a Hankel structured matrix. Outliers that might be present in time series can unduly influence model fitting, forecasting results and confidence intervals constructed using bootstrap methodology. The aim is to compare outlier identification techniques in SSA by simulating time series from the broad spectrum of time series that SSA can handle. Specific attention is paid to modern robust principal component analysis techniques such as ROBPCA which employs projection pursuit combined with estimation of robust covariance matrices. The latter is employed to outlier maps, which represents multivariate data in a two dimensional plot consisting of projected orthogonal distances plotted against score distances, in order to identify outliers. Promising results are obtained by applying robust principal component analysis to SSA. A well-known time series with an additive outlier present is used to illustrate the usefulness of the techniques.
Key words: Outlier maps, Robust principal component analysis, Singular spectrum analysis
Session: 4A, Wednesday 08h30 - 10h30, Eden Grove Red

Optimal factorial designs for two-colour microarray experiments: Properties of admissible designs, *A*-, *D*- and *E*-optimality criteria

Legesse K. Debusho[1,⋆], Linda M. Haines[2] and Dibaba B. Gemechu[1]

[1]University of Pretoria, [2]University of Cape Town

Abstract: In this paper the properties of A-, D-, E-optimal and admissible designs for a factorial structured two-colour microarray experiments are considered. Two type of parameterizations namely, Glonek and Solomon (2004) and standard factorial experiment (orthogonal), are used in the investigation. The numerical results show that A-, E-optimal and admissible designs depend on the parameterizations used in the model. The allocation of the treatment combinations in a 2×2 factorial experiment to the available number of arrays therefore do not necessarily coincide with one another for the two parameterizations.
Key words: Admissible design, A-optimality, Baseline parameterization, D-optimality, E-optimality, Orthogonal parameterization
Session: 7D, Thursday 08h30 - 10h50, Eden Grove Seminar Room 2

Analysis of extreme rainfall at East London, South Africa

Tadele Diriba⋆, Legesse K. Debusho, Joel Botai and Abubeker Hassen

University of Pretoria

Abstract: The aim of the extreme value analysis is to quantify the stochastic behaviour of extreme values. In this paper, the extreme value analysis was done using maximum likelihood techniques for the GEV distribution with block-maxima and the GDP distribution with threshold method for modelling rainfall data of East London. In the GEV case, the distribution was modified to take into account the temporal non-stationary trend in the annual maxima. However, for the GDP case, since the acf value was very small we have assumed independence for the daily maxima. Since the extreme observations are naturally scarce it is expected that the use of Bayesian inference may improve the efficiency of the parameters estimates of the distribution compared to the maximum likelihood method. Therefore, the Bayesian approach was also applied in the paper using the Markov chain Monte Carlo for the GEV distribution using the annual maximum. However this is not fully achieved in this study using the non-informative and informative priors. Therefore, rather than using annual maxim it may be better to consider the daily rainfall data because these data are more representative of the station.
Key words: Bayesian approach, Extreme value analysis, Generalized extreme value distribution, Generalized Pareto distribution, Gumbel distribution
Session: 2C, Tuesday 14h00 - 15h40, Eden Grove Seminar Room 1

Chasing shadows: Analysing data from camera trap studies

Greg Distiller[1,⋆] and David Borchers[2]

[1]University of Cape Town, [2]University of St Andrews

Abstract: Advances in technology have led to new ways of "capturing" animals including DNA analysis and camera traps that are especially useful for monitoring rare and elusive animals. Traditional capture-recapture (CR) and spatially explicit capture-recapture (SECR) models are based on a discrete sampling process that leads to well defined sampling occasions, but camera traps are continuous-time samplers which record the exact time of capture.

I have been working on a continuous formulation for SECR models that is more appropriate for data from camera trap studies. I will present the proposed framework and show how it leads to a likelihood for single-catch traps that has thus far proven elusive. The results from an application on jaguars as well as from simulations will also be presented.

Key words: Camera traps, Density estimation, Spatially explicit capture-recapture

Session: 2D, Tuesday 14h00 - 15h40, Eden Grove Seminar Room 2

Analysing key leaf characteristics for spectro-temporal discrimination

Nontembeko Dudeni-Tlhone

Council for Scientific and Industrial Research

Abstract: Quantification of temporal patterns using remote sensing technology is increasingly becoming important, particularly in monitoring environmental changes and variability in the ecosystems. In support for resource management and monitoring, field or laboratory data is essential in understanding changes and variability in the environment and in examining target objects.

Highly dimensional spectro-temporal measurements were used to understand the mechanism of change in tree leaves and to determine the best possible time for distinguishing between target tree species. These measurements were collected at about two-weekly intervals over an annual growing cycle (from June 2011 to May 2012), using a hyperspectral sensor (ASD Inc. Field Spectroradiometer).

This part of the study involves summarising highly dimensional measurements into key leaf characteristics and assesses their potential in discriminating between the target tree species. These characteristics include leaf pigment, health and structure.

Discriminant analysis was used to distinguish between eight tree species (deciduous and evergreen) and incorporated phenological information. The cross-validated results indicated that leaf characteristics have potential to distinguish between tree species and this can be enhanced by incorporating changes over time. For instance, when looking at seasonal profiles, the highest overall classification (75%) was observed in the winter season. This was supported highest classification accuracy (73%) during a week in August. It should, however, be noted that some of the deciduous trees had lost their leaves during this week. Therefore, this could require careful consideration when identifying these target species from a satellite image, for instance.

Key words: Discriminant analysis, Phenological information, Seasonal variability, Temporal-hyperspectral data

Session: 7B, Thursday 08h30 - 10h50, Eden Grove Blue

A family of weighted bivariate beta type I distributions and its generalizations

René Ehlers[1,⋆], Andriëtte Bekker[1] and Mohammad Arashi[2]

[1]University of Pretoria, [2]Shahrood University of Technology

Abstract: In this paper the family of weighted bivariate beta type I distributions is derived by using Taylor series expansion of the weight function $w(x_1, x_2) = h(\alpha x_1 + \beta x_2)$. We explore specific properties of the distribution and also use the distribution to calculate the Bayes estimator of Shannon entropy under squared error loss when using this distribution as a prior for the multinomial distribution. Several generalizations, including the bimatrix variate case as well as weights of p arguments, are also included.
Key words: Bayes estimator, Bimatrix beta type I, Bivariate beta type I distribution, Shannon entropy, Weighted distribution
Session: 2A, Tuesday 14h00 - 15h40, Eden Grove Red

Methods in the practice of sample allocation in stratified sampling: Review and perspectives

Sebnem Er[1,⋆] and Marcin Kozak[2]

[1]University of Cape Town, [2]University of Information Technology and Management in Rzeszow

Abstract: Stratified sampling is a sampling methodology where a heterogeneous population is divided into homogeneous sub-populations according to one (univariate stratified sampling) or more characteristics (multivariate stratified sampling) in order to gain more precision compared to simple random sampling. There are two main problems in stratified sampling that deserve special attention. One of them is how to construct the subpopulations which are called strata and the other one is how to allocate the total sample size (n) among strata. In this paper, our main concern is how to allocate the total sample size optimally while we assume that the strata are constructed in advance. The aim of our paper is to review the approaches developed in the literature since the optimum allocation method of Neyman's (1934) from different perspectives such as optimality, construction of the problem and the approach to the solution of the problem.
Key words: Multivariate stratified sampling, Sample allocation, Stratified sampling
Session: 5B, Wednesday 13h50 - 15h50, Eden Grove Blue

A study on the apparent randomness of an animal sample

Inger Fabris-Rotelli[1,⋆], Christine Kraamwinkel[1], Geoffrey Fosgate[1], Darryn Knobel[1] and Katie Hampson[2]

[1]University of Pretoria, [2]University of Glasgow

Abstract: It is mostly impossible to design sampling schemes for wildlife, and other animals too, which accurately represent the whole population. Samples are almost always convenience samples as a researcher will obtain data from each animal they can obtain. The representability of the samples need to thus be measured in some way and reported together with the results obtained for a full picture. This paper provides an introduction to this issue in wildlife research and provides some examples of such cases.
Key words: Convenience sampling, Rabies, Random sampling, Wildlife sampling
Session: 5B, Wednesday 13h50 - 15h50, Eden Grove Blue

Statistical sampling in auditing

L. Paul Fatti

University of the Witwatersrand

Abstract: Sampling lies at the heart of Auditing firms' audits of a company's books. However, the methodologies used by auditors for sampling company records can be very simplistic or even downright wrong. Starting with the definitions of control and substantive sampling used by auditors, the paper will comment on some of the recommendations contained in the Guide to the International Standards on Auditing. This will be followed by a proposed approach to sampling and analysis for auditors, aimed at putting their procedures on a firmer statistical basis.

Key words: Auditing, Control sampling, Substantive sampling

Session: 5B, Wednesday 13h50 - 15h50, Eden Grove Blue

Compounding as method in bivariate extended noncentral distributional development

Johan Ferreira[1,⋆], Andriëtte Bekker[1] and Mohammad Arashi[2]

[1]University of Pretoria, [2]Shahrood University of Technology

Abstract: We enrich the existing literature by developing new bivariate extended noncentral chi-square and F distributions, by utilizing the compounding method, using Poisson- and negative binomial probabilities. The method was discussed by Marshall and Olkin (1990, *Topics in Statistical Dependence*, 371-393) and more recently by Yunus and Khan (2011, *Applied Mathematics and Computation* 217, 6237 – 6247).

Some univariate distributions, as functions of the components of these newly proposed bivariate extended noncentral chi-square and F distributions are derived (such as the product, ratio, and proportion).

The benefit of introducing these distributions and the univariate distributions is demonstrated by graphical representations of their probability density functions. An example of rainfall data is used to illustrate the versatility of the proposed models.

Key words: Bivariate chi-square, Bivariate F, Compounding, Extended, Negative binomial, Noncentral, Poisson

Session: 3C, Tuesday 16h10 - 17h50, Eden Grove Seminar Room 1

Fleiss's kappa modified

Lizelle Fletcher

University of Pretoria

Abstract: A plethora of measures to capture the extent of agreement or concordance between two raters giving categorical ratings on a nominal scale to a fixed number of items have been developed over the years. The best-known statistical measure of interrater reliability is probably Cohen's kappa. It is a measure of the degree of agreement that can be expected above chance, i.e. an improvement on a simple percentage agreement calculation. Fleiss's kappa is an extension of Cohen's kappa to evaluate agreement among multiple raters on two or more categories. The statistic takes values between 0 and 1, where a value of 1 means complete agreement. In a study which comprises twenty five independent raters on four categories, some incongruous values were obtained for Fleiss's kappa. In one case, for example, there is 96% total observed agreement among the raters, while Fleiss's kappa is -0.02. The interpretation that there is no agreement beyond chance is not useful since it does not reflect the degree of concordance among the raters. This phenomenon has been documented previously for two observers. This paper explains a modification to Fleiss's kappa to obtain more realistic values for the measure of agreement among multiple raters.

Key words: Cohen's kappa, Fleiss's kappa, Interrater reliability

Session: 7A, Thursday 08h30 - 10h50, Eden Grove Red

Length-biased distributions and their properties

Dorette Funke*, René Ehlers, Andriëtte Bekker and J. J. J. Roux

University of Pretoria

Abstract: The focus is on the length-biased version of the log-normal and gamma distributions and their statistical and geometrical properties. The length-biased version of a distribution is a special case of the concept of weighted distributions introduced by Rao (1965), and is defined as follows: The length-biased version $g(x;q)$ of the original p.d.f. $f(x;q)$ of a non-negative random variable X when $E[X] < \infty$ is given by $g(x;q) = xf(x;q)/E[X]$, where q is a scalar or vector of parameters.
In practical applications the experimenter is forced to use $g(x;q)$ due to the nature of the data even though the interest is in the inference about the original distribution $f(x;q)$. Although the form-invariance of the length-biased version of $f(x;q)$ may be apparent, it is not as clear what happens to the properties of the estimator of q. Therefore, we investigate the differential geometric properties of $f(x;q)$ as well as the properties of $g(x;q)$. The statistical properties that will be investigated for both $f(x;q)$ and $g(x;q)$ are the coefficients of variation, skewness and kurtosis, together with Bowley's quantile measure for skewness and Moor's quantile measure for kurtosis.
Key words: Differential geometry, Form-invariance, Kurtosis, Length-biased, Skewness, Weighted distributions
Session: 5A, Wednesday 13h50 - 15h50, Eden Grove Red

Principal points and principal curves

Raeesa Ganey

University of Cape Town

Abstract: The idea of approximating a distribution is a distinguished problem in statistics. This paper explores the theory of principal points and principal curves as approximation methods to a distribution. Principal points of a distribution have been initially introduced by Flury (1990) who tackled the problem of optimal grouping in multivariate data. In essence, principal points are the theoretical counterparts of cluster means obtained by the k-means algorithm. Principal curves defined by Hastie (1984), are smooth one-dimensional curves that pass through the middle of a p-dimensional data set, providing a nonlinear summary of the data. In this paper, details on the usefulness of principal points and principal curves are reviewed. The application of principal points and principal curves are then extended beyond its purpose to well-known computational methods like Support Vector Machines in machine learning.
Key words: Computational methods, k-means algorithm, Machine learning, Principal component analysis, Principal points
Session: 2C, Tuesday 14h00 - 15h40, Eden Grove Seminar Room 1

A-optimal designs for two-colour cDNA microarray experiments using the linear mixed effects model

Dibaba B. Gemechu[1,*], Legesse K. Debusho[1], and Linda M. Haines[2]

[1]University of Pretoria, [2]University of Cape Town

Abstract: Two-colour cDNA microarray experiments help scientist to study the expression level of thousands of genes simultaneously under different conditions. Microarray experiments have different design challenges, such as for example which mRNA samples should be co-hybridized together and which treatment should be labelled with which dye fluorescent. Therefore, a carefully designed microarray experiment to obtain efficient and reliable data so as to assure the precise estimate of parameters that are of interest is needed. The present

paper is concerned with A-optimal block designs for two-colour microarray experiments, where array is treated as the experimental block. Linear mixed effects models were used to describe the experiments, by taking the arrays as random effects, when comparisons of all possible pairs of treatments are of particular interest. The numerical results show that the optimal block designs under linear fixed effects model are not necessarily optimal under linear mixed effects model setting.

Key words: A-optimal design, Block design, Linear mixed model, Microarray experiments, Robust designs
Session: 7D, Thursday 08h30 - 10h50, Eden Grove Seminar Room 2

Things are getting hotter – climate change in South Africa's National Parks

Victoria Goodall[1,⋆], Nicola van Wilgen[2], Stephen Holness[1], Steven Chown[3] and Melodie McGeoch[3]

[1]Nelson Mandela Metropolitan University, [2]South African National Parks, [3]Monash University

---

Abstract: Air temperatures have increased globally over the past decades, while rainfall has become more variable. In South Africa, substantial climate changes are predicted which means that protected area management agencies will need to actively respond to mitigate impacts. It is critical to understand the specific climate change impacts at a local scale to respond appropriately. Observable changes in temperature and rainfall are quantified over the past five to ten decades across the 19 National Parks in South Africa. Results show significant increases in temperature in most parks, with increases being most rapid in the arid regions of the country. These areas also had clear increases in the frequency of extreme temperature events. A concern is that the recorded increases in several areas were as extreme as or more extreme than predicted for near future scenarios by national models, indicating that these may be too conservative. Rainfall trends were less obvious, but a decrease in rainfall was observed for the southern Cape, while one park showed an increase in rainfall. Rainfall variation and pattern revealed differing trends in different parts of the country. Various statistical techniques and graphical displays are used to tease apart these patterns and create a baseline understanding of the changes that have occurred in the Parks over the past decades. The results also highlight the critical role that individual weather stations play in creating a picture of localised climate change.

Key words: Climate change, Rainfall variation, Temperature increase
Session: 5C, Wednesday 13h50 - 15h50, Eden Grove Seminar Room 1

Size and shape of multivariate data

Michael Greenacre

Universitat Pompeu Fabra

Abstract: Multivariate data are usually positive or, at least, non-negative. The case considered here is when all the data are on the same scale, for example all are counts, or all are in centimetres, or all in rands. Multivariate data vectors have differences between them that are measured by proximity measures, e.g. the Euclidean distance, the chi-square distance or the Jaccard index, the choice depending on the measurement scale of the data. In principal component analysis, which relies on Euclidean distance, the most important principal component often turns out to be a "size" component, lining up the multivariate vectors from ones with lower values to those with higher values. Subsequent components are then mostly capturing differences in "shape", but usually include a certain component of size as well. In correspondence analysis, which relies on the chi-square distance, the size effect has been eliminated from the outset because relative values are analysed, hence all components are those of shape.

In practice, many proximity measures are available and it is not immediately clear whether and how much they are capturing effects of size or shape. For example, the Bray-Curtis dissimilarity is ubiquitous in ecological research, but to what extent does this measure of inter-sample difference include size and shape effects? I investigate this question in the context of several distance and dissimilarity measures and try to come up with a way to assess how much these proximity measures and their subsequent analyses are capturing size, shape or a mixture of both.

Key words: Distance, Multivariate data, Shape, Size, Visualization
Session: 4A, Wednesday 08h30 - 10h30, Eden Grove Red

Has oil price predicted stock returns for over a century?

Rangan Gupta[1,⋆] and Paresh K. Narayan[2]

[1]University of Pretoria, [2]Deakin University

Abstract: This paper contributes to the debate on the role of oil prices in predicting stock returns. The novelty of the paper is that it considers monthly time-series historical data that span over 150 years (1859:10-2013:12) and applies a predictive regression model that accommodates three salient features of the data, namely, a persistent and endogenous oil price, and model heteroskedasticity. Three key findings are unraveled: First, oil price predicts US stock returns. Second, in-sample evidence is corroborated by out-sample evidence of predictability. Third, both positive and negative oil price changes are important predictors of US stock returns, with negative changes relatively more important. Our results are robust to the use of different estimators and choice of in-sample periods.
Key words: Oil price, Predictability, Stock returns
Session: 2B, Tuesday 14h00 - 15h40, Eden Grove Blue

Local variation of socio-economic inequality in South Africa according to different disparity indices

Jean-Marie V. Hakizimana

Statistics South Africa

Abstract: The purpose of the paper is to identify areas in South Africa where social-economic disparity exists using 2011 census data. Different indices are used to measure spatial disparity with the aim of finding the most appropriate approach for measuring disparities under different circumstances. The following measures were used in the study: the multidimensional composite index of deprivation; range ratio; relative mean deviation; standard deviation of logarithms; Gini coefficient; Kuznets ratio; Theil inequality index, mean logarithmic deviation, and the Atkinson index. In the study settlements are either regarded as individual settlements or contiguous settlements as delimited at the main place level in the census. Due to the fragmenting impact of apartheid on the South African society, different measures tend to be the most appropriate in different parts of the country. This implies that different policy interventions are needed to address area-specific challenges.
Key words: Deprivation, Indices, Inequality, Poverty, South Africa, Spatial disparity
Session: 7C, Thursday 08h30 - 10h50, Eden Grove Seminar Room 1

Objective Bayesian priors for the unbalanced one-way random effects model

Justin Harvey[1,⋆] and Abrie J. van der Merwe[2]

[1]Stellenbosch University, [2]University of the Free State

Abstract: In this presentation we present a Bayesian framework for the analysis of data arising from an unbalanced one-way random effects model. The data originates from an occupational exposure setting. Many authors have approached the problem and Krishnamoorthy and Mathew (2002) in particular proposed the use of generalized confidence intervals and a generalized p-value approach for the case of balanced data. Again, Krishnamoorthy and Guo (2005), amongst others, have proposed non-Bayesian methods for modelling unbalanced data as well. In this presentation we will apply an objective Bayesian approach to the problem and evaluate the performance of several non-informative priors.
Key words: Bayes, Non-informative priors, One-way random effects model
Session: 2A, Tuesday 14h00 - 15h40, Eden Grove Red

A variability analysis of nanoscale image measurements

Andries Haywood[1,⋆], Inger Fabris-Rotelli[1], Sonali Das[2] and James Wesley-Smith[2]

[1]University of Pretoria, [2]Council for Scientific and Industrial Research

Abstract: Variability in sample measurements is an important area of investigation in statistical analysis, including in the area of nanotechnology applications, which is the focus of our study. A major challenge faced in nanoresearch is the issue of difference in behaviour of nanoparticles at nanoscale when compared to their corresponding bulk scale. Failure to understand the properties of particles at nanoscale can compromise the way in which they are effectively used. In this paper we focus on investigating the properties of gold nanoparticle images measured using Transmission Electron Microscopy (TEM) under varying imaging conditions. Some observations and recommendations are made based on the results obtained in this study.
Key words: Nanoparticles, Nanoscale, Sample variability, TEM
Session: 5D, Wednesday 13h50 - 15h50, Eden Grove Seminar Room 2

Independent component analysis for pulsar classification

Annapurna Hazra[⋆] and Subharthi Ray

University of KwaZulu-Natal

Abstract: Astrophysical compact objects in the form of collapsed remnants of stars, provide a rich source of knowledge of matter in the extreme states. These compact stars, the pulsars, emit pulsed signals in different wavelengths, like x-ray, radio, gamma-ray, etc., and have been observed and studied vividly in the past two decades and more, thanks to the new age of space based satellites. However, theoretical modeling of these pulsars has always posed a challenge mainly because of the apparent lack of distinct correlation in their observed properties. The primary goal of this study is to find if there exists some correlation among these pulsars from their observed parameters, so as to identify them into distinct categories based on their morphology, using advanced statistical technique, namely, the Independent component Analysis (ICA). ICA is a blind source separation technique and has wide applications in various fields of science and engineering. ICA finds a set of source data that are mutually independent and non-Gaussian (non-normal). Due to non-Gaussian assumption, ICA can identify the original latent components, unlike classical multivariate statistical methods. We assume that data from different physical processes are uncorrelated but this does not always imply that uncorrelated data are coming from different physical processes. This is because uncorrelated variables are only partly independent. In the present case, an objective classification of astrophysical pulsars have been carried out, using the Australian Telescope National Facility (ATNF) Pulsar Catalogue. The set of observable parameters include pulse period, time derivative, surface magnetic field, distance from the galactic plane, radio luminosity, spin down energy loss rate, etc. Components responsible for significant variation have been obtained through ICA and the classification has been done by K-means clustering.
Key words: Classification, Cluster analysis, Independent component analysis
Session: 1A, Tuesday 11h30 - 12h10, Eden Grove Red

Dealing with non-response in the multivariate analysis of categorical data

Gill Hendry

University of KwaZulu-Natal

Abstract: Missing data is commonly encountered in most studies where data is collected by means of surveys. Furthermore, it is highly likely that much of the data will be categorical. Among the methods most frequently found in literature that are used to handle the missing data are complete case analysis and mean substitution. These, and other ad hoc methods of dealing with missing data, however, can lead to biased estimates and are

not, in general, recommended. Recently, there has been a move towards applying multiple imputation to take care of the non-response. This method has been shown to produce reliable parameter estimates under most conditions. Another, less well-known method that takes a very different approach but that effectively deals with the missing data, is subset correspondence analysis. In this presentation, these two diverse methods will be applied to a set of child asthma data that suffers from non-response, in order to identify genetic, environmental and socio-economic factors that affect childhood asthma. Results from the two methods will be compared and advantages of each method will be discussed.

Key words: Categorical data, Missing data, Multiple imputation, Subset correspondence analysis

Session: 7B, Thursday 08h30 - 10h50, Eden Grove Blue

Identifying clusters and outliers within electricity load profiles

Jenny Holloway*, Renee Koen and Richard van der Wath

Council for Scientific and Industrial Research

Abstract: This study relates to electricity load profiles and, in particular, to the peak day load profiles per year. In this context, a load profile can be defined to be the hourly electricity usage (electricity sent out) for a given day. These peak day load profiles are not anticipated to always follow the exact same pattern every year, however, a general overall pattern is expected. For this study it is of interest to know which load profiles deviate from this pattern. Since the load profile is non-linear, a non-parametric approach has to be taken to identify any unusual or atypical patterns.

Within the domain of statistical process control (SPC) there is an existing non-parametric approach that can be used to determine outlier profiles amongst a range of non-linear profiles, using a χ^2 control chart. However, what may be identified as outliers could, in fact, belong to an additional cluster of profiles and therefore should be modelled separately. Consequently, a second method has been identified from SPC literature which can be used to determine whether there are different clusters of profiles amongst a historical set of profiles.

While the electricity load profile data did not emanate from an industrial quality control process, these methods developed within the SPC context were seen as useful in the study of load profiles. This paper will discuss the application of these methods to compare load profiles across different years and will illustrate the findings using appropriate software.

Key words: Clusters, Electricity load profiles, Outlier profiles

Session: 6D, Wednesday 16h10 - 17h50, Eden Grove Seminar Room 2

Coverage probabilities and average lengths of confidence intervals for the parameters and quantiles of the Weibull distribution

Peter T. Iiyambo

University of Namibia

Abstract: Both exact and approximate statistical hypothesis tests and confidence intervals (CIs) for the parameters and the quantiles of location-scale and log-location scale families of distributions can be based on the maximum likelihood (ML) method. However, parameter estimation using ML can be difficult, in particular for censored samples, and the ML method may require intensive programming. This study compares the coverage probabilities (CPs) and average lengths (AVLs) of rank-based CIs for the parameters and quantiles of the two-parameter Weibull distribution, with CPs and AVLs of ML-based CIs. Results from a simulation study suggest that rank-based methods are very competitive with ML-based methods in terms of relative efficiency of CIs.

Key words: Average length (AVL), Coverage probabilities (CP), Rank-based confidence intervals, Weibull distribution

Session: 3B, Tuesday 16h10 - 17h50, Eden Grove Blue

Statistical methodology to measure the HIV serodiscordance among couples: The case of Mozambique

Adelino Juga[1,⋆], Marc Aerts[2] and Niel Hens[2,3]

[1]Eduardo Mondlane University, [2]Hasselt University, [3]University of Antwerp

---

Abstract: The number of people newly infected by Human Immunodeficiency Virus globally is continuing to decline, but national epidemics continue to expand in many parts of the world. In many countries including Mozambique, a substantial proportion of heterosexual couples, often in stable partnerships, is Human Immunodeficiency Virus serodiscordant and some of them are unaware of their partner's status.
Statistical Methodology to measure Human Immunodeficiency Virus 'Serodiscordance' among couples to investigate the effect of Human Immunodeficiency Virus prevalence among Serodiscordance couples is our main focus.
We propose a new measure of Human Immunodeficiency Virus Serodiscordance, the conditional synchrony measure, which is the probability of a couple to be concordant positive (both members within a couple being Human Immunodeficiency Virus positive), given that at least one of the two is already infected.
Cross-sectional data from INISIDA 2009, which is a national survey on prevalence, behavioral risks and information on HIV status in Mozambique was used.
Key words: Concordant positive, Conditional synchrony measure, Human immunodeficiency virus, Odds ratio, Serodiscordance
Session: 3A, Tuesday 16h10 - 17h50, Eden Grove Red

12.3 Oral Presentations (Surnames K - O)

Some statistical methods for detecting drug interaction

Gaëtan Kabera[1,⋆], Principal Ndlovu[2] and Linda M. Haines[3]

[1]Medical Research Council, [2]University of South Africa, [3]University of Cape Town

Abstract: In this talk, we present a number of statistical approaches for detecting additivity, synergy, or antagonism between two or more compounds and derive expressions of the parameters of interest. Historical datasets are used to illustrate the theory.
Key words: Additivity, Antagonism, Combination index, Isobologram, Synergy
Session: 4D, Wednesday 08h30 - 10h30, Eden Grove Seminar Room 2

Access to basic services in South Africa from 1996 to 2011

Gloria Kekana[⋆] and Arulsivanathan Naidoo

Statistics South Africa

Abstract: Many municipalities in the country have inequalities in terms of access to resources, infrastructure and social services. South Africa has one of the highest inequalities in the world, where high income groups and privileged population groups enjoy many basic services. Water is an essential service that should be provided to all South Africans. This study investigates whether more people from lower socio-economic groups are receiving water and other basic services. The data used is the 10% sample from census 1996, 2001 and 2011. Multivariate logistic regression is used to measure the relative odds-ratio of gaining access to formal housing and piped water based on socio-economic factors. Hot spot analysis is also used to identify the locations of statistically significant hot spots and cold spots with access to formal housing and piped water as the variables of interest.
Key words: Access, Housing, Piped water, Socio-economic factors
Session: 7C, Thursday 08h30 - 10h50, Eden Grove Seminar Room 1

A multiple imputation approach for missing air quality measurements

Sibusisiwe Khuluse[1,2,⋆], Alfred Stein[2] and Pravesh Debba[1]

[1]Council for Scientific and Industrial Research, [2]University of Twente

Abstract: Particulate matter pollution occurs when ambient air becomes contaminated with particles (which can be solid or mixtures of solid and liquid matter). Concern is with inhalable particles which have an aerodynamic diameter of less than 10 micrometers and are monitored as PM10. For monitoring annual statistics, such as the number of days an air quality standard is exceeded are of importance while for health impact studies finer temporal statistics are desired. For these purposes PM10 monitoring data pose a challenge because for some time periods the proportion of missing data is high. Simply ignoring such missing data leads to annual statistics with large standard deviations, as well as poor spatial predictions if the intention is to obtain these annual statistics over a region.
In this paper imputation of missing PM10 values is considered for four monitoring location types (i.e. residential, traffic, industrial and background) for a four year period (2008-2012). The imputation approach does not include modelling patterns of incompleteness as PM10 values are assumed to be missing at random (MAR). To exploit relationships between PM10 and meteorological (relative humidity, temperature, rainfall, wind speed and wind direction) variables at each location, a regression-based approach is pursued. Specifically, for each season, multiple fittings of a gamma generalized linear model with a log link are done, obtaining coefficients and reflecting uncertainty about these coefficients. Prediction error is added to reflect uncertainty about predicted values and to reduce the effect of imputed values being biased towards the mean.

Key words: Air quality, Gamma error distribution, Generalized linear models, Missing values, Regression-based multiple imputation
Session: 5B, Wednesday 13h50 - 15h50, Eden Grove Blue

Modelling Likert scale data using the beta distribution

Martin Kidd[1,⋆] and Nico Laubscher[2]

[1]Stellenbosch University, [2]Industat Pro

Abstract: Likert scale data in the form of discrete ordinal data are widely used in the psychometric field and other research where questionnaire surveys are used. For statistical analysis standard statistical methods (parametric or non-parametric) are mostly used. In many cases Likert data are modelled by assuming that the ordinal categories divide an underling normally distributed latent variable into intervals.
Tamhane *et al.* (2002), proposed using the same strategy of dividing the sample space of an underlying latent variable into intervals, but they used the Beta distribution as underlying latent variable. This approach is attractive because the sample space of the beta distribution is bounded, which fits in well with the bounded nature of Likert scale sample spaces.
The approach of Tamhane *et al.* is adapted to provide point probability estimates for each of the Likert scale outcomes, thus moving away from the latent variable approach to directly modelling probabilities for each of the discrete outcomes, but still taking into account the ordinal nature of the data.
A test for comparing average Likert scores from two independent populations will be presented and compared with various other statistical tests e.g. a t-test, the Mann-Whitney U-test, bootstrapping, etc.
Some extensions of the univariate beta to a bivariate beta will be discussed with the aim of estimating correlations between pairs of Likert scale variables.
Key words: Beta distribution, Likert scale
Session: 6A, Wednesday 16h10 - 17h50, Eden Grove Red

Parameter estimation for a mixture of two univariate Gaussian distributions: A comparative analysis of the proposed and maximum likelihood methods

Cliff R. Kikawa⋆ and Michael Y. Shatalov

Tshwane University of Technology

Abstract: Two approaches to the parameter estimation for a mixture of two univariate Gaussian distributions are numerically compared. The proposed method (PM) is based on decomposing a continuous function into its odd and even components and estimating them as polynomials, the other is the usual maximum likelihood (ML) method via the EM algorithm. An overlapped mixture of two univariate Gaussian distributions is simulated. The PM and ML methods are used to re-estimate the known mixture model parameters and the measure of performance is the absolute percentage error. The PM method produces more accurate results compared to the ML method. Given that the PM method produces good estimates, and knowing that the ML method always converges given good initial guess values (IGVs), it is thus recommended that the PM approach be used symbiotically with the ML method to provide IGVs for the EM algorithm.
Key words: EM algorithm, Maximum likelihood, Monte Carlo simulation, Parameter estimation, Univariate Gaussian mixture
Session: 7A, Thursday 08h30 - 10h50, Eden Grove Red

Small area estimation methods for household survey data analysis

Maggie Kisaka-Lwayo*, Neo Mashamba, Ngoako Mogkerepi and Caiphus Mashaba

Statistics South Africa

Abstract: The study of statistical methods for small area estimation has gained importance among several national statistical organisations, as small area techniques become more and more relevant in the production of official statistics. In South Africa, governance and planning at municipal (county) levels requires that statisticians are able to provide data that will assist government to address issues of service delivery and resource allocation. The necessity of small area estimation in survey analysis is considered when the domain sample sizes are insufficient to yield direct parameter estimates with adequate precision and reliability. Through small area estimation, statistical models are used to obtain indirect estimates for small domains using auxiliary information related to the variable of interest. It combines survey sampling and finite population inferences with statistical models. This paper describes the unit-level and area-level models for small area estimation in relation to their application in household surveys in South Africa.
Key words: Auxiliary, Domain, Finite populations, Small area estimation, Statistical model
Session: 5B, Wednesday 13h50 - 15h50, Eden Grove Blue

Forecasting sensor alarm and trip values in coal-fired power plants using artificial neural networks

Melisa Koorsse*, MC du Plessis and Igor Litvine

Nelson Mandela Metropolitan University

Abstract: A modern coal-fired electricity power generating plant is a complex industrial engineering system consisting of a number of mechanical and electrical devices and equipment. Supervisory control and data acquisition (SCADA) systems are in place for the condition monitoring of these plants. The maintenance and service of these plants is important and can result in planned capacity loss. However, the impact of unplanned capacity loss due to moderate or serious failures of energy production equipment, can result in serious problems requiring an alternative energy supply in a short period of time. In this paper an intelligent system to forecast sensor alarm and/or trip states in a coal-fired power plant using Artificial Neural Networks (ANNs) is proposed. A method of comparing actual sensor values with predicted values to forecast alarms is presented. The proposed approach is able to identify alarms earlier than the plant SCADA system. This is beneficial in order for plant management to have more time to mitigate and prevent trips and failures, which may result in unplanned capacity loss. The investigations reveal sensitivity of the ANN accuracy to the training data. The accuracy of the ANN is influenced by different inputs, including other sensor readings and lagged sensor readings. Results indicate that once an accurate model has been trained, it is generalisable to forecast the same sensor's readings in other components.
Key words: Artificial neural networks, Forecasting
Session: 5C, Wednesday 13h50 - 15h50, Eden Grove Seminar Room 1

An ARL-unbiased t_r-chart for monitoring times between failures

Nirpeksh Kumar[1,*] and Subha Chakraborti[2]

[1]University of Pretoria, [2]University of Alabama

Abstract: Xie *et al.* (2002) proposed the t-chart to monitor the times between events (failures) following an exponential distribution. They also generalized the t-chart to improve its performance and proposed the t_r-chart based on the times between the occurrences of r (≥ 1) events. It can be shown that these charts are ARL-biased, that is their ARL curves do not achieve a maximum at the in-control parameter value in known

as well as unknown parameter cases. To overcome this undesirable property, following Zhang *et al.* (2006), we consider an unbiasedt_r-chart ($r \geq 1$) for both the known and unknown parameter cases. A fixed size of Phase I sample is used to estimate the unknown parameter. The performance of the t_r-chart is evaluated in terms of ARL values in the known parameter case. We also investigate the effect of estimation error on the in-control and out-control chart performance in terms of metrics AARL (average of conditional ARL) and SDARL (standard deviation of conditional ARL). The size of the Phase I sample is recommended to have a desired level of chart performance in Phase II on the basis of standard deviation of conditional IC average run length (SDARL0) which reflects the practitioner-to-practitioner (or reference sample to reference sample) variation. An example is given to demonstrate the ARL-unbiased approach of the t_r-charts.

Key words: Average run length, False alarm rate, Gamma distribution, Phase I sample, Time between events

Session: 6D, Wednesday 16h10 - 17h50, Eden Grove Seminar Room 2

Estimating kernel hyper-parameters in kernel Fisher discriminant analysis

Morné M. C. Lamont

Stellenbosch University

Abstract: Statistical classification of observations plays a very important role in many applied research areas. The most well-kown classification (discriminant) technique was proposed by Fisher (1936) and is called Fisher's linear discriminant analysis. Many traditional statistical techniques such as Fisher's linear discriminant analysis have been kernelized (Mika *et al.*, 1999; Shawe-Taylor and Cristianini, 2004). Other kernelized methods include, kernel principal component analysis, kernel ridge regression and kernel clustering. Kernel-based multivariate techniques have gained quite some popularity in statistics over the past few decades. The most well-known kernel-based technique is probably the support vector machine proposed by Vapnik and co-workers (Boser *et al.*, 1992; Schölkopf, 1997).

This paper looks at two algorithms for kernel Fisher discriminant analysis (KFDA), which are kernelized versions of Fisher's linear discriminant analysis. The main aim of this paper is to address the problem of hyper-parameter estimation in KFDA. The popular Gaussian kernel function is used to perform KFDA. A general purpose genetic algorithm is used to find the optimal hyper-parameters by minimizing the error rate. The procedure is applied to real-world data sets using the R programming language. The results of the two KFDA algorithms are compared with linear discriminant analysis and the support vector machine.

Key words: Genetic algorithm, Kernel Fisher discriminant analysis, Kernel hyper-parameters

Session: 6A, Wednesday 16h10 - 17h50, Eden Grove Red

A biplot perspective on market-based valuations in an emerging market

Niel le Roux*, Soon Nel and Wilna Bruwer

Stellenbosch University

Abstract: The benchmark approach is commonly used for analysing the relationship between market prices and accounting information, as contained in financial statements, and to assess the valuation accuracy of market-based valuations. This study investigates the ability of market-based models to predict actual share prices in an emerging market. The actual share prices, as reflected on the South African JSE Securities Exchange, are benchmarked as "correct". The value estimates obtained from several market-based valuations, are subsequently compared to these benchmarked prices. Sixteen market-based valuations are investigated individually and as composites for the period 2001 to 2010. The chief objective is to optimise the weight allocations of the individual market-based models for inclusion in the composite model. Therefore, the minimisation of the median of the valuation errors (MVE) was employed as objective function for optimisation. The local optima problem was addressed using (i) random starts and (ii) employing the solution set for minimising the sum of absolute valuation errors (SAVE) as starting values in the MVE procedure. This two-step procedure produced the most accurate results. Given the multi-dimensional nature of the data, the use of

biplots and correlation monoplots proved to be particularly adept at displaying the behaviour of the various individual and composite models for market-based valuations.
Key words: Biplot, Multivariate analysis, Optimisation, Share price valuation
Session: 4A, Wednesday 08h30 - 10h30, Eden Grove Red

Estimating HIV incidence from prevalence data obtained from pregnant women in rural KwaZulu-Natal

Kerry Leask* and Ayesha Kharsany

CAPRISA

Abstract: Evaluating the rate of new HIV infections is necessary for investigating the dynamics of HIV transmission, for determining the impact of interventions and for identifying targets for further investigation and intervention. Collecting the data required for incidence rate calculations, however, is costly and can be difficult to implement since it necessitates the follow up of uninfected individuals over time. Frequently, annual prevalence data are available from cross sectional sero-prevalence surveys but the information gleaned from these measures is limited.
A method of estimating incidence from prevalence data which considers changes in prevalence over time while accommodating differential mortality in infected and uninfected individuals will be described and implemented on data obtained from an antenatal survey conducted by CAPRISA. Annual, anonymous cross-sectional HIV sero-prevalence surveys were conducted between 2001 and 2013 amongst first visit prenatal clinic attendees. The incidence estimates obtained will be compared with other estimates that have been described in the literature.
Key words: HIV, Incidence, Prevalence
Session: 3A, Tuesday 16h10 - 17h50, Eden Grove Red

A method to measure strangling, a dramatic form of choking in cricket

Hoffie Lemmer

University of Johannesburg

Abstract: Consider the situation where a team batting second in a limited overs cricket match needs to score only a few more runs with many balls still to be bowled, but is bowled out. This phenomenon is called 'strangling' because the bowling team succeeded to bowl their opponents, who were in a strong batting position, all out – they have strangled them. In this presentation a criterion is proposed to measure the severity of strangling. The measure takes into account how strong the batting team was just before this disaster happened. Real cases of strangling are used to illustrate the method.
Key words: Batting strength, Cricket, Sport
Session: 1C, Tuesday 11h30 - 12h10, Eden Grove Seminar Room 1

Growth curve comparisons to assess the impact of different ARV treatment regimes on Head Circumference profiles of HIV+ infants

Francesca Little[1,⋆] and Barbara Laughton[2]

[1]University of Cape Town, [2]Stellenbosch University

Abstract: We present an analysis and comparison of Head Circumference growth profiles for a cohort of 701 South African infants from birth to 36 months of age, including 250 HIV negative infants and 451 HIV infected infants. The HIV+ infants were divided into two groups based on the CD4 percentage counts at baseline. Those with CD4 percentage counts greater than 25% were randomised to one of three arms: (1) ART was delayed until immunological/clinical criteria were met, (2) ART commenced before 12 weeks and was interrupted after 40 weeks and (3) ART commenced before 12 weeks and was interrupted after 96 weeks. Those with CD4% count less than 25% at baseline could only be randomize onto the early treatment arms with no interruption of therapy.

The aim was to look at the impact of HIV infection and ARV treatment on head circumference measurements, as a proxy for brain development. More specifically, we compare the growth profiles for the HIV/ART groups using linear fractional polynomial models, cubic spline as well as nonlinear parametric mixed effect models, including shape invariant (SITAR) models. Preliminary results suggest possibly slower initial growth for HIV+ children in the delayed treatment arm, as well as different growth curves for boys and girls and for the two ethnic groups. In addition we assess the suitability of WHO and CDC reference curves for the South African cohort of HIV negative unexposed infants and consider the possible need for the development of growth reference ranges specific to the South African infant population.

Key words: Growth Curves, HIV+ infants, Shape Invariant Models

Session: 3A, Tuesday 16h10 - 17h50, Eden Grove Red

Hurst exponent for linear regression processes

Igor Litvine

Nelson Mandela Metropolitan University

Abstract: Hurst exponent is widely used in the time-series analysis as a measure of predictability. In this paper we evaluate the Hurst exponent for linear processes described by linear regression models assuming different distributions of the errors. The method uses a combination of analytical approach and computer simulations. We also discuss the feasibility of the Hurst exponent for describing predictability of such processes.

Key words: Hurst exponent, Linear regression, Predictability

Session: 6B, Wednesday 16h10 - 17h50, Eden Grove Blue

Parameter estimation by the method of arc lengths with application to the normal distribution

Theodor Loots[⋆] and Andriëtte Bekker

University of Pretoria

Abstract: A method, for estimating the parameters of normally distributed variables, is presented, based on the arc length of the probability density function. The goodness-of-fit is measured by means of the Anderson Darling test statistic, and the fit is compared with that obtained through maximum likelihood estimates.

Key words: Goodness-of-fit, Normal distribution, Order statistics, Parameter estimation, Spacings

Session: 5A, Wednesday 13h50 - 15h50, Eden Grove Red

Mixed effects model for bivariate categorical outcomes using local and continuation ratio logits for the marginal distributions

Osvaldo Loquiha[1,⋆], Marc Aerts[2], Niel Hens[2], Emilia Martins-Fonteyn[3] and Herman Meulemans[3]

[1]Eduardo Mondlane University, [2]Hasselt University, [3]University of Antwerp

Abstract: For categorical outcomes, often we relate relevant conditional univariate or bivariate marginals through log-linear or multivariate logistic regression models, or the seldom used special class of models introduced by Glonek (1996), based on parameterizations for the joint probabilities that combine both models. Colombi and Forcina (2001) introduced an extension to Glonek's approach which apart from linear constraints, allows for inequality constraints on the marginal parameters and 4 different types of logits to be used in the univariate and bivariate distribution.
We propose a hierarchical model based on a parameterization for the joint probability as in Colombi and Forcina (2001) using local and continuation ratio logits in the context of correlated or clustered data. Different formulations for the variance components that are estimated by maximum likelihood are provided and applied to investigate the association between HIV disease status, perception of risk of HIV and risky sexual behaviors.
Key words: Bivariate distributions, Categorical outcomes, Marginal models, Variance components
Session: 7D, Thursday 08h30 - 10h50, Eden Grove Seminar Room 2

Bayesian approach towards a better understanding of anemia at pregnancy in Africa

Siaka Lougue

University of KwaZulu-Natal

Abstract: In this study, a comparison is made between Bayesian and frequentist results of the analysis of pregnant women's anemia in Burkina Faso, Ghana, Niger and Cote d'Ivoire. The aim of this study is to show that the Bayesian method provides better results and to encourage the use of the Bayesian approach in a setting of enough prior information. In other hand, this study presents a case study of Bayesian approach to demographic and health issues to demystify the method to nonstatisticians and beginners. However, results of this study should have a significant contribution to the understanding and the fight against anemia at pregnancy and maternal mortality in Africa. The four recent Demographic and Health Surveys (DHS) for selected countries is considered in this study. Both frequentist and Bayesian hypothesis testing methods are used to check if there is a significant drop of anemia at pregnancy in the target countries. A multivariate Bayesian linear regression is implemented in comparison with frequentist approach to highlight the significant covariates of anemia in West Africa. The results of both Bayesian and frequentist approaches show that the level of anemia at pregnancy can be monitored if priority is given to some specific factors. When information are available from previous experiments, Bayesian methods provides better results than frequentist approach.
Key words: Africa, Bayesian statistics, Malnutrition, Maternal mortality, Multi-level analysis, Pregnancy
Session: 4D, Wednesday 08h30 - 10h30, Eden Grove Seminar Room 2

Comparison of basis function selection and model fitting for functional regression models

Sugnet Lubbe⋆ and Rene Essomba

University of Cape Town

Abstract: Functional data analysis (FDA) deals with continuous functions, rather than discreetly observed multivariate data. In FDA, the continuous functions are represented as a linear combination of basis functions. Probably the most popular basis functions for fitting functional data, is the b-spline basis, or for periodic data the Fourier basis. FDA is discussed in detail in the seminal work of Ramsey and Silverman (2005), with implementation of the methods in the R package fda. This is probably the most widely used FDA methods

in practice. However, Matsui, Kawano and Konishi (2009) criticize the least squares fitting methodology for functional regression models of Ramsey and Silverman, and suggests and alternative strategy based on penalized likelihood. They also use a different set of basis functions, based on the Gaussian distribution. A crucial issue in constructing functional regression models is the selection of regularization. Although generalized cross validation is widely used, Konishi and Kitagawa (1996) suggest the Generalized Information Criterion. In this paper, a critical evaluation is given of the different types of basis functions, the different methods for fitting functional regression models and a comparison of the selection of regularization parameters.

References

Konishi, S. and Kitagawa, G. (1996). *Information Criteria and Statistical Modeling.* Springer-Verlag: New York.

Matsui, H., Kawano, S. and Konishi, S. (2009). Regularized functional regression modeling for functional response and predictors. *Journal of Math-for-Industry*, 1, 17–25.

Ramsey, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, 2nd edition. Springer-Verlag: New York.

Key words: B-splines, Fourier basis, Functional data analysis, Functional regression models, Gaussian basis functions, Generalized information criteria

Session: 6A, Wednesday 16h10 - 17h50, Eden Grove Red

A new form of leverage in stochastic volatility models

Iain MacDonald[1,⋆], Walter Zucchini[2] and Roland Langrock[3]

[1]University of Cape Town, [2]Georg-August-Universitaet Goettingen, [3]University of St Andrews

---

Abstract: Stochastic volatility (SV) models are an alternative to (G)ARCH models as models for returns on an asset. Both can mimic the 'stylized facts' of share return series. Although there has long been evidence that such SV models outperform GARCH models, it is well known that they are more difficult to fit. But the methods of hidden Markov models (HMMs) can overcome these difficulties, and indeed such methods can be used to extend greatly the range of SV models that can be fitted.

I begin by giving a very brief introduction to SV models (in discrete time, without leverage) and to hidden Markov models. I then describe how SV models can be well approximated by hidden Markov models, which are fairly easy to fit. This technique has been used to fit a number of innovative models of SV type, e.g. one in which the underlying Gaussian autoregressive process is replaced by an independent mixture of two such processes. The technique can also be used to fit SV models with leverage, which appear to be superior to the models without leverage. I show how the HMM technique can be used in particular to fit a nonstandard SV model with leverage, one which provisionally appears often to outperform the standard SV model with leverage.

Key words: Hidden Markov models, Leverage, Stochastic volatility models

Session: 2B, Tuesday 14h00 - 15h40, Eden Grove Blue

Building biostatistics capacity in sub-Saharan Africa

Rhoderick Machekano

Stellenbosch University

Abstract: The creation of a sustainable, multidisciplinary health research enterprise in Sub-Saharan Africa (SSA) is of paramount importance for improving health, promoting development and advancing science in the region. In recent years, there has been a welcome expansion of health research activity in the region. This growth has created a demand for well-trained methodologists who can contribute to the design and analysis of research and provide mentoring and training of the next generation of researchers. Locally trained methodologists, including epidemiologists, social scientists and (to a lesser extent) health economists are now available in many SSA countries. However, there remains a critical shortage of biostatistics expertise in SSA resulting in an overreliance on input from biostatisticians sourced from economically developed countries or the

pharmaceutical industry for writing competitive grants, executing statistical procedures, conducting advanced data analysis, publishing in high profile journals and teaching biostatistics at the postgraduate level.
Good statistics departments do exist at SSA universities. The problem is that they focus almost exclusively on training students for business, accounting and commercial careers, rather than for work in the biomedical field. Those statistics departments that offer teaching in biostatistics generally do not link up with health sciences faculties, which means that students do not develop an understanding of the clinical context, or worse, never have the opportunity to fully appreciate the value of statistics in health and medical research. We propose a Center of Biostatistical Excellence for SSA focused on building biostatistics capacity through formal training programs and short courses.
Key words: Biostatistics, Capacity-building
Session: 4D, Wednesday 08h30 - 10h30, Eden Grove Seminar Room 2

Data visualisation and communication

Thandulwazi Magadla*, Nontembeko Dudeni-Tlhone, Jenny Holloway, Sibusisiwe Khuluse and Renee Koen

Council for Scientific and Industrial Research

Abstract: Data visualisation provides a means of highlighting patterns and making the underlying meaning of data much more apparent. However, statistical graphs are often uninspiring and more often than not, presented in a manner that assumes that the audience is able to tease out what is most important about them. This study explores how data visualisation can be used as a means of conveying statistical information to a general audience in a manner that is clear, concise and easily understood. A graphical user interface (GUI) is developed for this purpose. The results of a cluster analysis that was conducted for the purposes of identifying housing characteristics and household patterns in Ekurhuleni is used as a case study. Because the results are best represented graphically and interactively, using a print medium such as a report or a journal article becomes rather difficult as a large proportion of information has to be discarded in order to focus on the most important information. A digital medium allows for much more information to be retained and for it to be conveyed in a more interactive manner thus allowing the audience to explore it more thoroughly.
Key words: Clustering, Graphical user interface, Visualisation
Session: 5C, Wednesday 13h50 - 15h50, Eden Grove Seminar Room 1

Jump-diffusion based-simulated expected shortfall (SES) method of correcting Value-at-Risk (VaR) under-prediction tendencies

Sibusiso Magagula [1,2,*] and John O. Oloami[2]

[1]Nedbank, [2]University of South Africa

Abstract: Value-at-Risk (VaR) model fails to predict financial risk accurately especially during financial crises. This is mainly due to the model's inability to calibrate new market information and the fact that the risk measure is characterized by poor tail risk quantification. An alternative approach which comprises of the Expected Shortfall measure and the Lognormal Jump-Diffusion (LJD) model has been developed to address the aforementioned shortcomings of VaR. This model is called the Simulated-Expected-Shortfall (SES) model. The Maximum Likelihood Estimation (MLE) approach is used in determining the parameters of the LJD model since it's more reliable and authenticable when compared to other non-conventional parameters estimation approaches mentioned in other literature studies. These parameters are then used into the LJD model, which is simulated 100 times in generating the new loss dataset used in the developed model. This SES model is statistically conservative when compared to its peers and more reliable in predicting financial risk especially during a financial crisis.
Key words: Coherence, ES model, Fat-tailed distribution, HES model and SES model, Historical-simulation VaR model, Jump-diffusion models
Session: 2D, Tuesday 14h00 - 15h40, Eden Grove Seminar Room 2

Alternative methods to parametric significance testing in linear regression

Nhlanhla Makhanya*, Francois E. Steffens and Lizelle Fletcher

University of Pretoria

Abstract: The aim of the study is to survey permutation tests and Bootstrapping methods and their application to significance testing in simple and multiple linear regression analysis. A Monte Carlo simulation study was performed in order to compare the different methods in terms of empirical type 1 error and power of a test as measures of comparison. The empirical type 1 error and power of a test were used to compare permutation tests, Bootstrapping and parametric methods. These comparisons were performed in order to investigate the effect of
(1) sample size
(2) error distribution normal, uniform and lognormal
(3) number of explanatory variables 1, 2 and 5
(4) the correlation coefficient between the explanatory variables 0, 0.5 and 0.9.
The results obtained from the Monte Carlo simulation study showed that Permutation and Bootstrap produced similar type 1 error results while the parametric methods understated type 1 error when errors are lognormally distributed. In the absence of multicollinearity all the methods were almost equally powerful and in presence of multicollinearity all methods suffered equally in terms of power. It was concluded that permutation tests as well as Bootstrap methods are good alternative methods to use in significance testing in regression analysis.
Key words: Bootstrapping, Multicollinearity, Permutation tests, Power, Regression, Type 1 error
Session: 6B, Wednesday 16h10 - 17h50, Eden Grove Blue

Time series modeling of South African GDP

Rhoda Makhwiting

Statistics South Africa

Abstract: The Gross domestic Product (GDP) is the best indicator of the general state of the economy and the underlying trend of the economy, such as whether it is expanding, contracting, recessionary or inflationary. The data for the paper consists of quarterly market price of GDP from Statistics South Africa over the period 1st quarter 1993 to 2nd quarter 2014, generating 86 observations. To model the GDP, a class of ARMA (autoregressive integrated moving average) models is built following Box-Jenkins methodology. The model selection criterion includes AIC, MAPE and RMSE. The empirical result reveals that the most adequate model for the GDP at market price is ARMA(1, 1). The findings will assist the policy makers and managers to formulate economic and business strategies.
Key words: ARMA model, Box-Jenkins methodology, GDP, Time series modeling
Session: 6C, Wednesday 16h10 - 17h50, Eden Grove Seminar Room 1

Generalized additive modelling of age at first sex in South Africa

Mulalo A. Managa[1,*], Oluwayemisi O. Alaba[2], John O. Olaomi[1] and Khangelani Zuma[3]

[1]University of South Africa, [2]University of Ibadan, [3]Human Sciences Research Council

Abstract: We used the 2008 South African National HIV Prevalence, Incidence and Behaviour Survey data to investigate the determinants of age at first sex using the generalized additive model. The model was used to simultaneously measure the fixed and nonlinear effects. The fixed effect of categorical covariates was modelled using the diffuse prior and P-spline with second-order random walk for the nonlinear effect of continuous variable. The dependent variable follow a Gaussian distribution. Inference was fully Bayesian approach. Results showed a declining effect of coming from a middle class, primary and secondary education,

and staying in Eastern Cape, Free State, Mpumalanga and Northern Cape on early sexual intercourse. Early sexual intercourse is positively associated with internet, reading newspaper and magazine, coming from a poor home, female, white and Indian race, no father and living in Gauteng, Kwazulu Natal, North West and Western Cape. The average age at first sex in South Africa is 15.
Key words: Age at first sex, Bayesian inference, Gaussian, Generalized additive model, South Africa
Session: 5D, Wednesday 13h50 - 15h50, Eden Grove Seminar Room 2

Exploratory analysis of functional data

Siphumlile Mangisa[1,⋆], Sonali Das[2] and Gary D. Sharp[1]

[1]Nelson Mandela Metropolitan University, [2]Council for Scientific and Industrial Research

---

Abstract: "Functional data" refers to data where each observation is a curve. Functional data analysis (FDA) displays the data so as to highlight various characteristics and important sources of variation among the data. This presentation is based on a preliminary analysis of data using FDA methods. In particular, we consider the exploratory techniques of FDA (e.g. functional principal component analysis (FPCA)).
Key words: Functional data analysis
Session: 3B, Tuesday 16h10 - 17h50, Eden Grove Blue

Survival analysis of children under-five: A comparative study between Gauteng and the Rest of South Africa (ROSA)

Thanyani Maremba⋆ and Zandile Nukeri

Statistics South Africa

Abstract: South Africa's Childhood Mortality Rates (CMR) and causes of death is generalised, however, a more in-depth study of survival data is necessary to give indication of survival rates in different geographic areas. Studying causes and rates of under-five mortality between rural and urban area can help decision makers to assess programmatic needs and prioritise interventions. This paper analyses the survival times and levels of childhood mortality in Gauteng (a predominantly urban province in South Africa with most metropolitan municipalities) and the Rest of South Africa (ROSA).The 2006-2012 Mortality and Causes of Death data was used including children born from 2006 and either still alive or died after 60 months. The statistical methods applied examine whether there are disparities on the survival rates of children and infants. Analysis uses the Kaplan-Meier survival plot and Cox Proportional Hazard models. Advanced techniques are also discussed, including proportional hazards models to test the statistical differences of the two areas, Gauteng and ROSA, Log rank tests are used to assess the extent of the differences between the two areas. The results show that there is a trend that in earlier months, the survival curves are different and in latter months they tend to be the same, however, the trend differs across provinces. The relevant explanatory variables related to the survival times provide more explanation on the trends revealed by the survival functions, and they are therefore included in the study.
Key words: Child mortality, Kaplan–Meier estimate, Mortality and causes of death, Survival analysis
Session: 7C, Thursday 08h30 - 10h50, Eden Grove Seminar Room 1

Inequalities in South Africa: A Statistical analysis

Lehlogonolo Masenya

Statistics South Africa

Abstract: South Africa is ranked as one of the most unequal societies in the world. As the country celebrates 20 years of democracy in the year 2014, the question remains as to whether the economic inequalities of the apartheid era are beginning to fade. Using income data from the Population Censuses of 1996, 2001 and 2011, this paper first determines whether the income inequality among population groups is statistically significant. Thereafter, the correlation between income inequality and education and unemployment will be analysed. We also look at the policies that the new government have put in place to rectify inequalities. Finally, some conclusions are drawn based on the empirical results of the study.
Key words: Education, Income, Inequality, Race, Unemployment
Session: 7C, Thursday 08h30 - 10h50, Eden Grove Seminar Room 1

Estimation of high dimensional longitudinal and recurrent events data joint models using marginal likelihood techniques

Lieketseng Masenyetse[1,⋆], Samuel O. M. Manda[1] and Henry Mwambi[2]

[1]Medical Research Council, [2]University of KwaZulu-Natal

Abstract: The theory and analysis of longitudinal and recurrent events data separately is now becoming well known. Extensions to analysis of simple joint longitudinal and recurrent events data are fitted and implemented in most standard statistical software. However, estimation of parameters governing high dimensional processes for longitudinal and recurrent events data poses a myriad of problems. For example, in HIV treatment studies, researchers collect a number of longitudinal measurements together with a number of time-to-events outcomes, some of which are recurrent. Estimation in this context must account for correlations within each longitudinal and within each recurrent events processes, in addition to correlation structure between these processes; resulting in very high dimensional algorithms.
We show how the estimation procedure can be carried out using marginal likelihood function using Gaussian quadrature technique to find estimates of model parameters. This was implemented in SAS software and applied to a HIV antiretroviral pharmacovigilance study.
Key words: Gaussian quadrature, Longitudinal, Recurrent events
Session: 4D, Wednesday 08h30 - 10h30, Eden Grove Seminar Room 2

Modeling and forecasting South African CPI using ARIMA models

Sipho Masimula

Statistics South Africa

Abstract: In this article we attempt to model and forecast South African Consumer Price Index (CPI) using ARIMA model on monthly data for the period of February/2010 to May/2014 obtained from Statistics South Africa website. The Box-Jenkins model building procedure was adopted to identify an adequate model for the data series. To test for stationarity; Augmented Dickey-Fuller (ADF) test was employed.
It was found that the CPI data was stationery and that it is ARIMA(1, 0, 1). Using the model to forecast CPI for the month of June and July the following CPI values were obtained, 6.70%,6.60% respectively . Our findings suggest that the CPI will increase by 0.1% in June.
Key words: ACF, ADF test, ARIMA, CPI, PACF
Session: 3B, Tuesday 16h10 - 17h50, Eden Grove Blue

Exploring inequalities in utilization of maternal health services in South Africa

Lyness Matizirofa

University of Johannesburg

Abstract: Background: The maternal mortality rate in South Africa is 625 deaths/100.000 deaths, which is high compared to the Millenium Development Goals of 38 deaths / 100 000 births. Understanding the factors contributing to its high maternal mortality rate, will assist policy makers and stakeholders to formulate interventions that can help to improve barriers to utilization of maternal health services.
Objective: To identify barriers to the utilization of maternal health care services in disadvantaged communities of South Africa.
Methods: A cross-sectional study was carried out in three purposively selected study sites in the three South African provinces. A stratified sampling technique was used to recruit 180 mothers for our study. Descriptive statistics were done to explore the data. A comparison of categorical variables was done using Chi-square test. Kruskal-Wallis test was applied to test differences in barriers to utilization of maternal health care services. Simple logistic regression analysis was further carried out to explore the relationship between barriers to maternal health services and the response variable.
Results: Access to maternal health services was hindered by financial, transport, long distances to the nearest health facility barriers. Lack of money was the main barrier for not attending antenatal care and transport delay was a barrier to delivery care services.
Conclusion: Financial and transport problems are barriers to maternal service utilization. Poverty reduction and economic empowerment of rural women are prerequisites for any tangible improvement in the utilisation of maternal health services.
Key words: Barriers, Health, Maternal, Services, Utilization
Session: 4C, Wednesday 08h30 - 10h30, Eden Grove Seminar Room 1

Crime statistics in South Africa, a hot spot analysis

Zanele Mazibuko

Statistics South Africa

Abstract: This study investigates the Crime statistics reported in South African Police stations. A hot spot analysis is performed on the number of murders, robberies and sexual crimes. A comparison between the various municipalities is conducted and the results are mapped spatially using Arcmap.
Key words: Crime statistics, Hot spot analysis
Session: 7C, Thursday 08h30 - 10h50, Eden Grove Seminar Room 1

Causality testing for regime-switching processes

Farai Mlambo* and Igor Litvine

Nelson Mandela Metropolitan University

Abstract: Causal analysis has always been a significant role playing field in applied sciences such as statistics, econometrics and technometrics. Particularly, probability raising models have warranted significant research interest. Contemporarily, the econometric causality theory developed by C. J. Granger is popular in causal applications. While this type of causality technique has many strong features, it has serious limitations. The processes studied, in particular, should be stationary and causal relationships are restricted to be linear. I. J. Good, proposed a probabilistic event-type explication of causality which circumvents some of the limitations of Granger's methodology. In this paper, the probability raising causality ideology, as postulated by Good is used to propose some causal analysis methodology applicable in a stochastic, non-stationary domain. A proposal is made for a Good's Causality Test, by transforming the originally specified probabilistic causality theory from a random events to a stochastic-process framework. Methodological validation is performed via causality simulations for a Markov Regime Switching model. The proposed test can be used to detect whether one stochastic process can be said to have caused the observed behaviour of another, probabilistically. In particular, the regime switch causality explication proposed herein is pivotal to the results articulated.
Key words: Causality, Methodology, Non-stationarity, Regime-switching
Session: 3D, Tuesday 16h10 - 17h50, Eden Grove Seminar Room 2

Comparing statistical methodologies in an application to assess perceptions associated with bank card fraud susceptibility

Mihlali Mnci*, Warren Brettenny and Gary D. Sharp

Nelson Mandela Metropolitan University

Abstract: This study aims to investigate and identify attitudes, behaviours and perceptions harboured by bank cardholders and merchants, which are associated with a higher susceptibility towards experiencing bank card fraud. Primary data was obtained from bank cardholders and merchants, from various business categories, in both the Nelson Mandela Bay Metropolitan Municipality and the City of Johannesburg Metropolitan Municipality. Following the use of parametric Multinomial Logistic Regression (MLR) and nonparametric conditional density estimation to analyze the data, the results are compared and relevant covariates/perceptions are determined from the more accurate of the 2 techniques.
The results of the analyzed survey data serve as a tool, highlighting areas which require further education and awareness on the part of merchants and bank clients.
Key words: Bank card fraud, Nonparametric, Parametric, Perceptions
Session: 2D, Tuesday 14h00 - 15h40, Eden Grove Seminar Room 2

Effect of parameter estimation on the performance of the t-chart and t_r-chart

Nthabiseng Molata

Statistics South Africa

Abstract: Process control has become necessary in the fast-paced environment that businesses operate under where defects have to be minimized. The objective of the study is to examine the effect of parameter estimation on the performance of the t-chart and t_r-chart. These are time-between-events charts which monitor the time between successive occurrences of events. The study will be done based on a thorough examination of the conditional false alarm rate (CFAR) and the conditional and unconditional in-control Average run length. The Maximum likelihood estimator (MLE) and the uniformly minimum variance unbiased estimator (UMVUE) will be used when estimating parameters for this analysis.

Key words: Conditional and unconditional in-control average run length, Conditional false alarm rate (CFAR), MLE, UMVUE

Session: 6D, Wednesday 16h10 - 17h50, Eden Grove Seminar Room 2

Properties and applications of the negative hypergeometric distribution

Henri Moolman

Walter Sisulu University

Abstract: The negative hypergeometric distribution is the "sampling without replacement" equivalent of the negative binomial distribution. Properties and applications of this distribution as well as its relation to some other distributions will be discussed.

Key words: Finite population, Negative hypergeometric, Waiting times

Session: 3B, Tuesday 16h10 - 17h50, Eden Grove Blue

Diagnosing the variability in gas production using support vector machines

André Mostert[1,⋆], Roelof Coetzer[1] and Sugnet Lubbe[2]

[1]Sasol Technology Research and Development, [2]University of Cape Town

Abstract: The Sasol® FBDB™ coal gasification units of Sasol Synfuels, Secunda South Africa, convert approximately 30 million tons of bituminous coal per year into hydrocarbon synthesis gas. Coal gasification processes are acknowledged to be complex by nature with a large degree of interaction and variation between multitudes of variables. In order to better understand and interpret these complex systems in pursuit of continued improvements and optimization, advanced statistical modelling is required for prediction and trouble shooting. Sasol's coal scientists and engineers have developed a good fundamental understanding of the correlations between coal quality and the amount of gas produced. However, the implementation of a fundamental model in a commercial process is very difficult due to many uncontrollable variables present in the process. Fortunately, modern instrumentation allows for the real time collection of large volumes of process data. Consequently, real time data are employed for multivariate statistical modelling on many process variables, and for the quantification of the effect of the input on the outputs of the process. In this presentation we demonstrate how a support vector machine regression model is used to model the relationship between the amount of gas produced and the quality of the coal, and the implementation of such a model for real time process performance monitoring. Furthermore, we show how directional partial derivatives are used to identify the contribution of specific coal parameters to production deviations.

Key words: Coal gasification processes, Gas production, Multivariate statistical modelling, Process perfomance monitoring, Support vector machines

Session: 4A, Wednesday 08h30 - 10h30, Eden Grove Red

Under-five mortality in South Africa: The roles of geographic area and socioeconomic status

Collen Motsepa

Statistics South Africa

Abstract: This study examined the association between socioeconomic status and under-five mortality (U5M) at sub place (suburb) level in South Africa. Data is utilised from Census 2011. Two provinces were selected: Kwazulu-Natal (KZN) and Western Cape (WC). The two provinces were selected based on their under-five mortality rate (U5MR) as compared to other provinces in South Africa. KZN had the highest U5MR whilst WC had the lowest. U5MR was calculated at a sub place level in each selected province. Predictors of U5MR were determined by multiple regression analysis. Explanatory variables used in model: Unemployment rate, average household size and proportion of Black-headed household. In addition to this, this study looked at the effect of spatial analysis on the data due to the fact that only 50% of the area in KZN is urban whilst 90% of the area in WC is urban. This was conducted using geographically weighted regression. In the results of the regression model that explain 61% of variation across sub places in the two provinces, average household size explained more than the unemployment rate. Unemployment rate was the weakest predictor as compared to the proportion of black-headed households and the average size of the households. The average household size and the type of the population group that dominated the sub place statistically explain much of the variation in U5MR. In households exhibiting the same characteristics with regards to population group of head of household and household size, U5MR was higher in rural areas than in urban area.

Key words: Geographically weighted regression, Multiple regression analysis, Spatial analysis, Under-five mortality

Session: 7C, Thursday 08h30 - 10h50, Eden Grove Seminar Room 1

Factors affecting the survival of small to medium enterprises in South Africa

Innocent Mudhombo*, Principal Ndlovu and Mulalo A. Managa

University of South Africa

Abstract: In this study the eleven factors affecting survival or failure of SMEs in South Africa are investigated. The study analyses the survival times of SMEs using non-parametric methods (Kaplan-Meier and Life Table), semi-parametric method (Cox regression) and parametric methods. Male owned SMEs seem to survive longer than the female owned. In the long run being educated seems to have a better positive effect on the survival time of SMEs especially beyond nine years. The SMEs owned by Whites survive longer as compared to the other racial groups. The survival of SMEs is likely to be influenced by industrial sector and location arrangement. SMEs keeping full accounting records have longer survival times to enterprises that keep part or no records. Although the data fitted well on the Weibull distribution, the Gompertz distribution is the most appropriate model. Parametric methods are found to be superior to their non-parametric and semi-parametric counterparts.

Key words: Cox regression, Kaplan-Meier, Life table, SME survival factors, Survival analysis

Session: 5C, Wednesday 13h50 - 15h50, Eden Grove Seminar Room 1

Using geographically weighted regression to explore local crime patterns in South Africa

Arulsivanathan Naidoo

Statistics South Africa

Abstract: This paper examines a structural model of murder and house robberies in South Africa by exploring spatial patterns of the crimes reported and their covariates. The recently released crime statistics as reported per police stations are mapped onto the Census 2011 data. Spatial clustering techniques are applied and hot spots are identified. An ordinary least squares regression model is fitted to find relationships between the crimes and socio economic variables of the community. Geographically weighted regression is then introduced as an alternative to the traditional approaches to modelling crime. The Geographically weighted regression procedure estimates a local model, producing a set of mappable parameter estimates and t-values of significance that vary over space.
Key words: Crime patterns, Spacial clustering techniques, Structural model
Session: 7C, Thursday 08h30 - 10h50, Eden Grove Seminar Room 1

Balanced modified random sampling in the presence of linear trend

Llewellyn R. Naidoo

University of KwaZulu-Natal

Abstract: A major disadvantage for systematic sampling is that it is impossible to obtain an unbiased estimate of the sampling variance based on a single sample. Various modified systematic sampling designs have been proposed over the years to address this issue. However, in the presence of linear trend, few of these designs provide satisfactory results. In the present paper, we propose a modified systematic sampling design to tackle the above-mentioned shortcoming. The results suggest that the proposed sampling scheme provides optimal results for populations exhibiting linear trend.
Key words: Balanced random sampling, Systematic sampling, Variance
Session: 3C, Tuesday 16h10 - 17h50, Eden Grove Seminar Room 1

Regression analysis for complex sampling data

Ariane Neethling[1,⋆], Retha Luus[2] and Tertius de Wet[1]

[1]Stellenbosch University, [2]University of the Western Cape

Abstract: Regression analysis is a widely applied area in Statistics. In practice, complex survey (CS) data is often used where the information is collected according to a non-simple random sample design and each observation is assigned a design weight which encompasses the design of the survey. This complicates the analysis compared to the case of regression analysis with iid error terms. However, it has been found that when investigators perform regression analysis using CS data, the underlying design and design weights are ignored and ordinary least squares regression is used. Alternatively, it is assumed that using weighted least squares regression instead, where the design weights are then used to compensate for non-homoscedasticity in the model errors, adequately takes the design of the survey into account. Although this will still correctly estimate the regression coefficients, their estimated standard errors could be incorrect.
Survey weighted least squares regression, however, specifies a general variance structure for the model residuals which leads to estimated standard errors of the regression coefficients that differ from the standard errors obtained from weighted least squares regression.
The purpose of this presentation is to investigate the effect of the survey design and the trimming of extreme weights, on the estimated regression coefficients and their estimated standard errors. For this investigation the aim is to model person income from various explanatory variables identified from the Income and Expenditure

Survey of 2005/2006. In the analysis, results from ordinary least squares, weighted least squares, and survey weighted least squares regression will be compared.
Key words: Complex sampling, Extreme weight trimming, Regression analysis, Survey weighted least squares regression
Session: 5B, Wednesday 13h50 - 15h50, Eden Grove Blue

Analysis of spatial and temporal patterns of trees affected by guava wilt disease in South African guava orchards

Mzabalazo Ngwenya* and Maritha Schoeman

Agricultural Research Council

Abstract: In this study spatial and temporal pattern analysis on data collected on the incidence of guava wilt disease in South African guava orchards was carried out. This was in an attempt to discover and characterize patterns in time and space of trees affected by guava wilt disease, a soil borne pathogen in order to develop an integrated control strategy for the disease. Spatial patterns were investigated directly by use of Ripley's K function and the bivariate K function. Analysis of the spatial patterns at successive time points allowed some temporal component into the analysis. From the analysis it was found that spread of the pathogen is predominately over short distances and this is driven by splash dispersal. The results also suggested root contact between neighbouring plants as another mechanism for pathogen spread.
Key words: Pathogen spread, Spatial patterns, Temporal patterns
Session: 7A, Thursday 08h30 - 10h50, Eden Grove Red

Comparison of the genetic algorithm and incremental optimisation routines for a network design problem

Alecia Nickless

Council for Scientific and Industrial Research, University of Cape Town

---

Abstract: The genetic algorithm (GA) is an optimisation routine which takes advantage of the 'survival of the fittest' concept from evolutionary biology in order to obtain the optimal solution based on a cost function. GA's are a class of stochastic optimisation procedures for any numerical algorithm which calculates a score based on a function of inputs. This optimisation routine considers all parameters in a solution simultaneously. Based on the specification of the population size and number of iterations, the computational cost of this routine can be very large. The GA also requires the use of random numbers for the generation of new population members, therefore there is a possibility of a different final solution each time the algorithm is run. An alternative is the incremental optimisation (IO) routine which computes the cost function from the set of possible solutions for each parameter incrementally, selecting the best solution at each iteration before moving onto the next parameter. This method is more efficient in terms of computational resources, but may not be able to find the best solution in the case of a multiple-parameter problem, since the parameters are considered sequentially, rather than simultaneously.
The estimation of fluxes of trace gases requires a network of atmospheric monitoring sites where the concentrations of these gases are measured at high precision. The optimal location of new measurement sites for such a network is a case where both of these methods are applicable. This paper compares the use of the GA and IO routines for the optimisation of a five member network of carbon dioxide monitoring sites in South Africa. The cost function is based on the posterior covariance matrix of the Bayesian inverse modelling solution for the carbon dioxide fluxes estimated from the measurements from this network. The comparison considers the utility of the solution, the similarity of solutions, and computational requirements of these two approaches. This paper found that the IO routine failed to find the solution with the global maximum uncertainty reduction, but that the solution was extremely similar to the best solution found by the GA, at only a fraction of the computational resources. The GA performed inconsistently if the number of iterations or population size were not large enough.

Key words: Genetic algorithm, Incremental optimisation, Inverse modelling, Network design
Session: 3D, Tuesday 16h10 - 17h50, Eden Grove Seminar Room 2

Regularised iterative multiple correspondence analysis in multiple imputation

Johané Nienkemper-Swanepoel[1,2,⋆] and Michael J. von Maltitz[2]

[1]Stellenbosch University, [2]University of the Free State

Abstract: The occurrence of non-response in survey data is a prevalent problem. The application of a regularised iterative multiple correspondence analysis (RIMCA) algorithm in single imputation (SI) has been suggested for the handling of missing data in survey analysis. An adapted version of this algorithm is applied as a multiple imputation (MI) technique in this paper and compared to the published results. The RIMCA algorithm expresses MCA as a weighted principal component analysis (PCA) involving the generalised singular value decomposition of a weighted data matrix. The success of the RIMCA algorithm lies in the fact that all eigenvalues are shrunk and the last components are omitted, thus a 'double shrinkage' occurs which reduces variance and stabilises predictions. A comparison is drawn between the performance of SI and MI making use of RIMCA in both simulated and real data. It was found that the MI procedure provided accurate estimates and wider confidence intervals. Therefore, whilst adding the uncertainty when imputing missing values, sufficient estimates were obtained. Over 1000 simulations smaller bias and mean squared error were obtained from the MI estimates than the SI estimates. The advantages of MI with regards to SI in the context of RIMCA will be presented.
Key words: Incomplete ordinal categorical data, Multiple correspondence analysis, Multiple imputation, Principal component analysis, Regularised iterative multiple correspondence analysis
Session: 7A, Thursday 08h30 - 10h50, Eden Grove Red

Use of modified stability analysis and best linear unbiased predictor in formulation of recommendation domains for sugarcane varieties

Peter M. Njuho[1,⋆] and C. N. Sewpersad[2]

[1]University of South Africa, [2]South African Sugarcane Research Institute

---

Abstract: Intensive breeding of sugarcane varieties sustain production that ensures South Africa maintains the lead as a major sugarcane producer. The South African Sugar Association Research Institute (SASRI) and Swaziland Sugar Association Technical Services (SSATS) are continually extending out-grower services to new growing areas. Evaluation of commercial varieties across a range of sites is conducted with a view to drawing recommendation domains to growers for different agro climatic conditions and management practices. The current climatic conditions and socio-economic factors demand for development of high yielding varieties that utilise the available resources more efficiently. This paper discusses two alternative statistical procedures, the modified stability analysis (MSA) and the best linear unbiased predictor (BLUP) in the formulation of recommendation domains for sugarcane varieties grown across ten sites in two Southern African countries (South Africa and Swaziland). We illustrate these two procedures using yield data on fourteen sugarcane varieties. The complexity of missing data overtime and space is highlighted. Evaluation of treatment effects and partitioning of sites into recommendation domains is easily done using the proposed statistical procedures. We demonstrate the process of analysis and highlight the challenges encountered. Analysis in stages is suggested in order to achieve maximum results
Key words: Best linear unbiased predictor (BLUP), Missing data, Modified stability analysis (MSA), Sugarcane varieties
Session: 7B, Thursday 08h30 - 10h50, Eden Grove Blue

Comparison of socio-economic status index with the income variable

Muthetho S. Nkwinika

Statistics South Africa

Abstract: In the absence of income and expenditure data, some multivariate techniques can be used to develop a proxy for wealth. There are concerns though about the meaning of the wealth proxy in some quarters of multidimensional poverty analysis. This paper takes advantage of the presence of both the income variable and variables on durable goods and related housing characteristics in the Housing and Population Census 2011 datasets of South Africa. Using Principal Component Analysis (PCA) approach as described by Deon Filmer and Lant H. Pritchett (February 2001) an asset index is developed and compared with household income variable. The purpose of comparing the index with income is to show reasonable correspondence between the classification of households based on the asset index and a classification based on income.
Key words: Assets wealth index, Deprivation, Income, Principal component analysis
Session: 4C, Wednesday 08h30 - 10h30, Eden Grove Seminar Room 1

On-line teaching of Statistics

Delia North[1], James S. Allison[2,⋆], Stefan Britz[3,⋆] and Mardi Jankowitz[4,⋆]

[1]University of KwaZulu-Natal, [2]North-West University, [3]University of Cape Town, [4]University of South Africa

Abstract: More and more school leavers in South Africa are emerging from a schooling system with university level passes, resulting in increased pressure on university management systems to accept more students for tertiary education. Resulting pressures on venues and teaching staff have lead some universities to adopt an on-line teaching platform for some selected modules. This session explores the feasibility of on-line teaching of Statistics within the reality of the South African setting, by getting presentations from a software company that facilitates such a process, as well as getting the input from local academics that have experience in this regard.
Key words: On-line teaching, Tertiary education
Session: 4B, Wednesday 08h30 - 10h30, Eden Grove Blue

Laying the foundation for a Statistically literate society in South Africa – lessons learnt in KZN, way forward

Delia North[1,⋆], Temesgen Zewotir[1] and Iddo Gal[2]

[1]University of KwaZulu-Natal, [2]University of Haifa

Abstract: Developing countries face many challenges when aiming to develop statistical literacy. Training in-service teachers to promote statistical literacy in such countries has to be very carefully planned to make the most effective use of limited training resources available.
A team of Statisticians from UKZN have worked collaboratively (with Stats SA, Dept of Basic Education), to offer an annual training program, which aims to empower in-service teachers with skills to teach Statistics topics meaningfully. This program was introduced as a legacy program after the hosting of the ISI conference in Durban in 2009 and has to date empowered over 2000 teachers. The talk has three parts. First, to give a brief overview of the project. Second, to reflect on lessons learnt, so as to gain insight into the adoption of statistics instruction in teachers who previously had little knowledge or exposure to the subject. Third, to offer a possible model to expand on successes on the UKZN campus to serve as a "testing ground" for further research into statistics capacity building at grass roots level.
Key words: In-service teachers, Statistical literacy, Training program
Session: 4B, Wednesday 08h30 - 10h30, Eden Grove Blue

Exploring supply-side municipal service delivery in South Africa using non-financial census of municipalities

Mbulaheni Nthangeni and Patrick Naidoo*

Statistics South Africa

Abstract: The purpose of this presentation is to illustrate the potential use and value of the Non-Financial Census of Municipalities (NFCM) as an alternative measure for evaluating and influencing service delivery in South Africa.
Municipal service-delivery data is normally obtained from irregularly conducted household-based surveys, or from the demand side. The NFCM, however, uses data obtained annually from the municipalities – or from the supply side, using the consumer unit as a measure. The possible application of this data is illustrated by some observations on basic and free basic service delivery trends over the period 2006-2012 for electricity, water, sewerage and sanitation and solid waste management. It also investigates the use of this data to provide a spatial analysis of reasons for changes in service delivery.
The NFCM provide some indication of the problems facing municipalities, and important factors influencing service-delivery decisions from the supply side.
Some statistical, economic, accounting and other applications are indicated in this presentation. From this perspective it could serve as a stepping-stone to various areas of research on service delivery at the local government level.
Key words: Municipalities, Non-financial census, Service delivery, Supply-side
Session: 4C, Wednesday 08h30 - 10h30, Eden Grove Seminar Room 1

Forecasting models for possible impact of climate change on the electricity demand in South Africa: Empirical analysis

Thandekile Nyulu* and Igor Litvine

Nelson Mandela Metropolitan University

Abstract: Forecasting, by nature, is a stochastic problem rather than deterministic. There is no perfect forecasting model. There is always room for improvement, such as reducing the variance, choice of other predictors, etc. A good forecasting model has to capture all the important features of the series or the real process, while utilising minimum variables or parameters. In various publications the following driving factors of electricity consumption are suggested: climate (weather), calendar events (e.g. holidays), economic indicators, human activities (e.g. strikes, elections), and the interactions of the above. In this study, we are looking for a robust and reliable forecasting model, a model that will indicate the impact of each driving factor on electricity demand. Various statistical techniques (i.e. Multiple Regression, Auto-Regression, Categorical Regression etc) are considered to perform this task. Artificial Neural Networks (ANNs) will also be considered in this study since electricity load is nonlinear due to the effects of temperature.
Key words: ANNs, Electricity demand, Forecasting, Forecasting model, Statistical techniques
Session: 3C, Tuesday 16h10 - 17h50, Eden Grove Seminar Room 1

Estimation from ranked set sampling

John O. Olaomi[1,⋆] and Raghunath Arnab[1,2]

[1]University of South Africa, [2]University of Botswana

Abstract: Ranked set sampling is used when the measurement or quantification of units of the variable under study is difficult but the ranking of units of sets of small sizes can be done easily by an inexpensive method. Stokes (1977) considered ranking as an auxiliary variable. Prasad (1989), Kadilar *et al.* (2009) and Singh *et al.* (2012) considered the estimation of the population mean μy assuming the population μx is known. In this paper, we have proposed improved methods of estimation of the population mean using the ranking variable x as an auxiliary variable when the population mean μx is unknown. An empirical investigation based on life data shows all proposed estimators are approximately unbiased and bring gain in efficiency of up to 50 percent.

Key words: Order statistics, Ranked set sampling, Relative efficiency, Sampling with replacement

Session: 7A, Thursday 08h30 - 10h50, Eden Grove Red

Comparing the accuracy and fairness of three different resource tables for the Duckworth-Lewis method in Twenty20 cricket

Wallina Oosthuizen⋆, Robert Schall, Ian Carney and Rilinde Ramudzuli

University of the Free State

Abstract: In recent years Twenty20 cricket has strongly grown in popularity. The question has been asked whether the Duckworth-Lewis method, which is based on the scoring patterns of international 50-over one-day cricket games, is the best "rain rule" for interrupted Twenty20 cricket games.
Recently, two new resource tables specifically for Twenty20 cricket have been introduced by Perera and Swartz. In the present study, therefore, we investigate the relative merits the two rain rules associated with the new resource tables with the Duckworth-Lewis method. In order to do so we consider a sample of completed Twenty20 cricket games, but artificially interrupt the second innings of each game at various stages of the innings. Comparing the actual result of the game with the result predicted by the different rain rules we can assess the accuracy and fairness of those rain rules.

Key words: Accuracy, Fairness, Rain rules, Twenty20 cricket

Session: 1C, Tuesday 11h30 - 12h10, Eden Grove Seminar Room 1

12.4 Oral Presentations (Surnames P - Z)

The LogNIG distribution and some of its properties

Kevin Panman[1,⋆] and Tertius de Wet[2]

[1]North-West University, [2]Stellenbosch University

Abstract: In the financial risk management (FRM) environment, a risk manager is often concerned with calculating various risk measures, such as the value-at-risk and expected shortfall. These quantities are derived from the estimated quantiles of some underlying distribution. A wide range of techniques and distributions exist to estimate the relevant extreme quantiles. A well known distribution used in FRM is the Normal Inverse Gaussian (NIG) distribution.
The traditional parameterization of the NIG distribution uses 4 parameters and is clearly a parameter-rich distribution. This implies that it could be used to model a wide variety of data. It is observed that the NIG distribution has semi-heavy tails by analyzing it using extreme value theory.
Data in the financial risk environment is often very heavy-tailed. Fitting the NIG distribution to such heavy-tailed data can lead to underestimating extreme quantiles that can results in major losses from a financial risk manager's point of view. There is then clearly a need to extend the NIG to a heavier tailed distribution. A natural method of obtaining a heavy-tailed distribution from a semi-heavy tailed distribution is to take exponentials. Applying this to the NIG distribution, one obtains the Log Normal Inverse Gaussian distribution, written as LogNIG. The LogNIG distribution can be used to model financial risk data possessing heavy tails. This talk briefly discusses the NIG distribution, and then focuses on the LogNIG distribution and some of its properties. Applications will also be discussed. Further research possibilities will be pointed out.
Key words: Heavy-tailed distribution, Frechét-Pareto class, Log normal inverse Gaussian
Session: 5D, Wednesday 13h50 - 15h50, Eden Grove Seminar Room 2

On the application of the CPC model in discriminant analysis

Theo Pepler

Stellenbosch University

Abstract: The common principal components (CPC) model provides a way to model the population covariance matrices of several groups by assuming a common eigenvector structure. When the assumption of equality of the covariance matrices of several populations is untenable but the CPC hypothesis is justified, this model can provide covariance matrix estimators of which the elements have smaller standard errors than that of the per group unbiased sample covariance matrix estimators.
In this paper, the CPC covariance matrix estimators and the recently developed regularised CPC covariance matrix estimators are plugged into the ordinary quadratic discriminant function. Simulation results show that, in the CPC situation, discriminant analysis for two groups under the CPC model outperforms both ordinary quadratic discriminant analysis and linear discriminant analysis with regard to misclassification error rates.
Based on the simulation study and an heuristic explanation of the multidimensional shapes defined by the covariance matrix estimators in different circumstances, recommendations are made about which type of covariance matrix estimator to use when performing discriminant analysis for two groups.
Key words: Common principal components, Covariance matrix, Discriminant analysis, Monte Carlo simulation, Shrinkage estimator
Session: 6A, Wednesday 16h10 - 17h50, Eden Grove Red

The marital status of South Africans, a spatial analysis

Gaongalelwe Phakedi

Statistics South Africa

Abstract: This study investigates the marital status of the South African community. The data from Census 2001 is compared with the data from Census 2011. The ages of the married community and the number of unmarried people living together is profiled by population group and other socio economic factors. Esri's ArcMap is used to map the results spatially.
Key words: ArcMap, Marital status
Session: 4C, Wednesday 08h30 - 10h30, Eden Grove Seminar Room 1

Using first passage time densities to perform inference on non-linear, multivariate diffusion models

Etienne A. D. Pienaar

University of Cape Town

Abstract: Diffusion models are useful tools in the modelling of multivariate time series. Although most non-linear systems of stochastic differential equations prove to be analytically intractable, much progress has been made in the development of algorithms that facilitate inference on such diffusion models. Specifically, when sample trajectories may be observed at discrete points in time, a diverse array of methods exist that may be used to infer model parameters. However, in scientific fields such as ecology and biology it often occurs that the direct observation of sample trajectories may be infeasible or impractical. In such cases, where it may be more natural observe the time taken for a trajectory to first move from an initial state into a region of the state space, one may resort to the first passage time density in order to perform inference on a given model. To this end we explore how a special type of first passage time problem may be used to perform inference on a diffusion model. In the present paper we show how one may combine an approximation to the transitional density with integral equations that govern the evolution of the first passage time density, in order perform inference on a proposed diffusion model. The methodology is developed for a class of multivariate diffusion models with quadratic drift and diffusion, thus allowing for non-linear model specifications. We illustrate the methodology with a short simulation study.
Key words: Cumulant truncation procedure, Diffusion process, First passage time, Stochastic differential equation, Volterra equation
Session: 6A, Wednesday 16h10 - 17h50, Eden Grove Red

Multivariate cointegration analysis of employment, inflation and output for short- and long-run linkages

Sagaren Pillay

Statistics South Africa

Abstract: This study empirically examines the short-run and long-run linkages between South Africa's output, inflation and employment using quarterly data from 1994 to 2013. The theoretical framework used for the study is based on Johansen's Maximum Likelihood cointegration technique which tests for both the existence and number of cointegration vectors that exists. The study finds that the three series are integrated of order one and are cointegrated. A statistically significant cointegrating relationship is found to exist between employment, inflation and output. The findings of the study confirm the existence of long run and short run linkages between the variables. This unique linear and lagged relationship is modelled using a Vector Error Correction Model (VECM). The VECM (2) model is found to be a good fit for the data with output and employment a weak exogeneity of inflation.
Key words: Cointegration, Exogeneity, Lagged, Linear, Maximum likelihood, Vector error correction model
Session: 7D, Thursday 08h30 - 10h50, Eden Grove Seminar Room 2

Repeating service modules - reasons and remedies

Eeva Rapoo

University of South Africa

Abstract: UNISA's open distance education setting allows learners to repeat an individual module many times, providing a good opportunity to study how motivation and attitude impact on learning statistics. We illustrate some typical trends in service modules, and the implications they have on the way we need to organise and teach the modules.
Key words: Open distance learning, Statistics education
Session: 4B, Wednesday 08h30 - 10h30, Eden Grove Blue

Modelling residential electricity usage within the eThekwini Municipal area

Samantha Reade

University of KwaZulu-Natal

Abstract: The aim of this research was to model residential electricity consumption within the eThekwini municipal area, the intention being, that the developed model be used to predict future electricity consumption of individual households/dwellings. Using a data set obtained from eThekwini Electricity, a subsidiary of eThekwini Municipality and official supplier of electricity to the municipality, monthly electricity consumption was modelled using both standard and generalized linear mixed models. Key findings of the research included finding evidence of a seasonal pattern in monthly electricity consumption, and, proving that variations in consumption of different households could not be related to physical distance between them. Predictions made using the fitted mixed models were found to be closer to the actual value, than that of the customary eThekwini Electricity predicted values.
Key words: Electricity consumption, Generalized linear mixed models, Linear mixed models, Prediction
Session: 3D, Tuesday 16h10 - 17h50, Eden Grove Seminar Room 2

The estimation of age-specific reference intervals using fractional polynomials and an exponential transformation

Tarylee Reddy[1,⋆] and Ismail Bhorat[2]

[1]Medical Research Council, [2]University of KwaZulu-Natal

Abstract: Age-specific reference intervals are an important screening tool in medicine, allowing one to detect individuals presenting with "abnormal" values of a variable of interest. The objective of this study was to estimate age specific reference intervals of the modified myocardial performance index (Mod-MPI) in fetuses aged between 20 and 38 weeks. Since logarithmic and shifted logarithmic transformations did not sufficiently eliminate residual skewness in the outcome, an exponential transformation, involving an age-specific skewness parameter was applied. Fractional polynomials were used to fit the mean and standard deviation curves. Modeling the skewness as a function of gestational age as opposed to a constant in the model significantly improved the model fit. The goodness of fit of the model was assessed by normal Q-Q plots of z-scores. Centile curves were calculated by back-transformation of the estimated centiles to the original scale.
Key words: Fractional polynomials, Reference intervals, Skewness
Session: 4D, Wednesday 08h30 - 10h30, Eden Grove Seminar Room 2

A topic model approach to inferring episodic directional selection in protein coding sequences

Hassan Sadiq* and Miguel Lacerda

University of Cape Town

Abstract: Pathogens, such as HIV and influenza, evolve in response to the selective pressures of their host environments, accumulating changes in their genomes that offer fitness benefits. This selective pressure is characterized by three properties:
(1) it is episodic, tracking changes in the adaptive immune response and drug therapy;
(2) it is directional in that only particular amino acid substitutions are favoured; and
(3) it varies between genomic loci. Most previous models have ignored or inadequately addressed some of these phenomena.
This work extends recent approaches to modelling episodic, directional selection acting on protein-coding sequences. We use inference techniques within the topic model framework, notably variational Bayesian methods, to identify loci evolving under natural selection. We show that our approach performs well in terms of specificity and power, and demonstrate its utility by applying it to a real dataset of HIV sequences.
Key words: Amino acid, Collapsed Gibbs' algorithm, Collapsed Variational Bayes' algorithm, Directional selection, Episodic selection, Evolution, HIV, Pathogens, Protein sequence, Rate heterogeneity, Selection pressure, Therapy, Topic models, Unconstrained approximation
Session: 2C, Tuesday 14h00 - 15h40, Eden Grove Seminar Room 1

On a data-dependent choice of the tuning parameter appearing in certain goodness-of-fit tests

Leonard Santana* and James S. Allison

North-West University

Abstract: We propose a data-dependent method for choosing the tuning parameter appearing in many recently developed goodness-of-fit test statistics. The new method, based on the bootstrap, is applicable to a class of distributions for which the null distribution of the test statistic is independent of unknown parameters. No data-dependent choice for this parameter exists in the literature; typically, a fixed value for the parameter is chosen which can perform well for some alternatives, but poorly for others. The performance of the new method is investigated by means of a Monte Carlo study, employing three tests for exponentiality. It is found that the Monte Carlo power of these tests, using the data-dependent choice, compares favourably to the maximum achievable power for the tests calculated over a grid of values of the tuning parameter.
Key words: Bootstrap, Empirical characteristic function, Exponential distribution, Goodness-of-fit tests, Tuning parameter
Session: 6B, Wednesday 16h10 - 17h50, Eden Grove Blue

On marginal fiducial inference based on conditional generalized fiducial pivotal quantities

Robert Schall

University of the Free State

Abstract: Fiducial generalized pivotal quantities (FGPQs) can be used for inference in many situations where classical pivotal quantities, and classical exact inference, are not available. We introduce the concept of conditional FGPQs (CFGPQs), and associate their distribution with a conditional fiducial distribution of the parameters of a statistical model. Using the Gibbs sampler the joint and marginal fiducial distributions of the model parameters can be derived, and inference carried out. We apply the concept to the problem of constructing generalized confidence intervals (GCIs) for the mean and variance parameters of a possibly

unbalanced mixed model; a simulation study suggests that the GCIs maintain conservative frequentist coverage while being somewhat shorter on average than CIs calculated using the best competing methods.
Key words: Fiducial generalized pivotal quantities, Fiducial inference, Gibbs sampler, Mixed model
Session: 5A, Wednesday 13h50 - 15h50, Eden Grove Red

On the theory of opportunistic condition-based maintenance (with application)

Mahmood Shafiee[1], Maxim Finkelstein[2,3,⋆] and Christophe Bérenguer[4]

[1]Cranfield University, [2]University of the Free State, [3]University ITMO, [4]CNRS Institut Polytechnique de Grenoble

---

Abstract: Offshore wind turbine blades are subject to multiple types of internal and external damages. Internal damages (such as fatigue, wear and cracks) are generally caused by system degradation, whereas the external damages (such as wave shocks) result from harsh marine environments. In this paper, we investigate an optimal opportunistic condition-based maintenance (OCBM) policy for a deteriorating "multi-bladed" offshore wind turbine system subject to stress corrosion cracking (SCC) and external shocks. When the length of a crack in a blade reaches a critical size D, the blade fractures and it has to be replaced by a new one. An external shock is minor with probability $1-p$ and catastrophic with probability p $(0 \leq p \leq 1)$. A minor shock causes an instant drop in power output without resulting in any system failure, whereas a catastrophic shock stops the wind turbine and requires system replacement. When the length of a crack on one of the blades exceeds a threshold d $(< D)$, it undergoes a major repair, and a preventive maintenance (PM) action is performed on the other blade(s) as well; otherwise, a planned PM task is conducted on the whole system when its operational age attains a value of T (> 0). The problem is to determine the optimal control parameters d^* and T^* such that the average long-run maintenance cost per blade per unit time is minimized. The explicit expression of the objective function is derived and under certain conditions, the existence and uniqueness of the optimal solution are shown for the infinite-horizon case. The proposed maintenance model is tested on field data collected from an offshore wind farm database, and its performance is evaluated over the lifetime of wind turbine using a Monte-Carlo simulation technique.
Key words: External shocks, Monte-Carlo simulation, Offshore wind turbine system, Optimal opportunistic condition-based maintenance (OCBM), Stress corrosion cracking (SCC)
Session: 5A, Wednesday 13h50 - 15h50, Eden Grove Red

Improved confidence intervals for small area estimation under the Fay-Herriot model: Application to food insecurity in Ethiopia

Yegnanew A. Shiferaw

University of the Witwatersrand

Abstract: There is a growing demand for small area estimates for policy and decision making, local planning and fund distribution. Surveys are generally designed to give representative estimates at national or district level, but estimates of variables of interest (such as poverty measures) are often also needed at lower (small area) levels. These cannot be obtained from the survey data as the sample sizes at these levels are too small. Census data is often available, but only gives limited information with respect to the variables of interest. This problem is addressed by small area estimation techniques such as Fay-Herriot model, which combine survey and census results, using a weighted combination of the estimates from the survey data (direct estimates) and that from the census data (synthetic estimates). The main aim of the research is to develop confidence intervals with a higher overage probability than currently available in the literature. These are applied to estimation of food insecurity in Ethiopia using the 2010/11 household consumption expenditure survey and the 2007 census data sets.
Key words: Census data, Confidence intervals, Small area estimation
Session: 7B, Thursday 08h30 - 10h50, Eden Grove Blue

The determinants of divorce

Cleopatra Sikhosana

Statistics South Africa

Abstract: South Africa is one of the many countries with increasing divorce rates. It is an unavoidable fact in today's society, but are some more prone to divorce than others. This paper will take a look at selected variables in an attempt to find a correlation between the variables and divorce rates. Four independent variables have been selected based on the availability of information and theoretical relevance for the equation. The data used is from General Household Surveys conducted by Statistics South Africa annually. Regression analysis will be utilised to help explain why divorce has been on the rise, using the ordinary least squares method.
Key words: Divorce, Education, Employment, Income, Population
Session: 4C, Wednesday 08h30 - 10h30, Eden Grove Seminar Room 1

Factors associated with youth unemployment in South Africa

Glacia Simangwe

Statistics South Africa

Abstract: The aim of the study is to investigate the factors associated with youth unemployment in South Africa. Secondary data from Statistics South Africa were used from Quarterly labour force survey (QLFS) second quarter of 2014. There were 15 815 youth respondents, consisting of 7 836 male and 7 880 female. The data were analysed using Pearson's chi-square and logistic regression. Pearson's chi-square was used to identify the associations between the dependent variable and the independent variables. Logistic regression was used to develop statistical models that can be used to identify the factors associated with youth unemployment. The findings show that gender, province, age, educational level and race are reasons related with youth unemployment.
Key words: Educational level, Logistic regression, Quarterly labour force survey,Youth unemployment
Session: 4C, Wednesday 08h30 - 10h30, Eden Grove Seminar Room 1

Time series Thurstonian paired comparisons models

Morné Sjölander

University of the Free State

Abstract: Our existing Time Series Models for Paired Comparisons have successfully been introduced and applied to various situations in recent years. We will introduce new Time Series Thurstonian Paired Comparisons Models based on the Thurstone-Mosteller Model for Paired Comparisons. In the past, we have used several of our Time Series Models for Paired Comparisons to estimate temperatures in an area in which temperature is missing or where it is no longer measured, using the current temperatures of the surrounding towns, as well as available time period's temperatures of this town, and the corresponding temperatures of the surrounding towns during this time period. Models that we have used include the Linear Bradley-Terry Model and the Sinusoidal Bradley-Terry Model. We now compare our past results with similar results based on our new Thurstonian Time Series Models for Paired Comparisons.
Key words: Paired comparisons, Time series, Thurstone-Mosteller model
Session: 6C, Wednesday 16h10 - 17h50, Eden Grove Seminar Room 1

Online quantiles via Hermite series density estimation over data streams

Michael Stephanou* and Melvin M. Varughese

University of Cape Town

Abstract: Algorithms for elucidating the statistical properties of streams of data in real time and for the efficient one-pass analysis of massive data sets are becoming increasingly pertinent. Certain statistical properties naturally lend themselves to efficient, online calculation such as the mean and standard deviation. Depending on the application, these moments may not be sufficient however. This may be true for skewed data for example. In addition, if the data stream being analysed is prone to outliers then more robust statistics may be required. Quantiles are a natural choice in this setting. Examples of areas in which online quantile estimation is relevant include network traffic and latency analysis, real time fraud detection and high frequency trading. We introduce new techniques for online quantile estimation based on Hermite series estimators in the settings of static and dynamic quantile estimation. In the static quantile estimation setting we apply the existing Gauss-Hermite expansion in a novel manner. To treat dynamic quantile estimation we introduce a new expansion which we term the Exponentially Weighted Gauss-Hermite expansion. The Gauss-Hermite based algorithms allow arbitrary quantiles (as opposed to pre-specified quantiles) to be estimated at any point in time, thus overcoming a limitation of many existing algorithms. Indeed, we solve the more general problem of estimating probability densities and cumulative distribution functions on data streams. We discuss some theoretical properties of the Gauss-Hermite algorithms as well as some simulation and real data results.
Key words: Data streams, Massive data, Orthogonal series estimators, Quantiles, Sequential estimation
Session: 2C, Tuesday 14h00 - 15h40, Eden Grove Seminar Room 1

Multiple imputation in the presence of a detection limit, with applications: An empirical approach

Cornelia J. Swanepoel* and Shawn C. Liebenberg

North-West University

Abstract: Missing measurements that are reported to be below a fixed, known detection limit, is a regular occurrence especially in the environmental sciences. Such censored data are often ignored or "guessed" because measurements were made which were incorrectly reported, usually to be zero or to be equal to the detection limit. However, reliable estimates of the population parameters are required to perform statistical analysis. It becomes a complex task to perform when a large number of observations are below this limit. Rigorous robust estimation procedures are then needed.

This study focuses on density estimation in such scenarios by imputing data to replace the censored data below the detection limit in a sensible way. The maximum likelihood procedure of Cohen (1959) and several variants thereof, are then applied to estimate the parameters of the underlying density function. Estimation of this density function is then attempted by using the completed imputed data set. Various boundary kernel density estimators are applied comparatively.

More specifically, in this study three different Log-normal distributions will be considered. The above-mentioned methods are implemented in combination with four new multiple imputation procedures, to assess which of these nonparametric methods are most effective in imputing data to replace the censored values. Several kernel density estimators are fitted to the complete filled-in data set. Comparative measures are applied to establish which combination of strategies are the best to estimate the underlying density function in the presence of a detection limit. The results of an extensive Monte Carlo simulation study are presented and conclusions and recommendations regarding are made.
Key words: Bootstrap, Detection limit, Kernel density estimation, Maximum likelihood estimation, Multiple imputation
Session: 6B, Wednesday 16h10 - 17h50, Eden Grove Blue

Estimation of promotion rates, repetition rates and dropout rates for learners in South African schools

Danie Uys[1,⋆] and Edward Alant[2]

[1]Stellenbosch University, [2]GRS Actuarial Consulting

Abstract: A new procedure for estimating promotion, repetition and dropout rates for learners in South African schools is proposed. The procedure uses three different data sources: data from the South African General Household survey, data from the Education Management Information Systems and data from yearly reports published by the Department of Basic Education. The data from the General Household survey are utilised to estimate repetition rates for learners in three different age groups. Keeping these repetition rates fixed, the data from the other two sources are used to estimate dropout and promotion rates which are based on a birth year cohort approach for the different age groups. In particular, this procedure involves minimising the difference between actual flow through rates and simulated flow through rates for both the birth year cohorts and age groups. The procedure gives different results when compared to published literature.
Key words: Birth year cohorts, Dropout rates, Promotion rates, Repetition rates, South African schools
Session: 1B, Tuesday 11h30 - 12h10, Eden Grove Blue

Fair value adjustments for smaller financial institutions

Carel van der Merwe

Ernst & Young

Abstract: Given the recent changes in the financial landscape and introduction of IFRS 13 Fair Value Measurement, smaller financial institutions are faced with the difficulty of incorporating fair value adjustments for credit risk in their annual financial statements.
A brief outline of how these adjustments can be simplified will be given, together with different solutions of possibly implementing it within the IFRS 13 framework. Different tests will be discussed to see what the impact of the various methodologies on these approaches could lead to.
Key words: CVA, Derivatives, IFRS
Session: 2B, Tuesday 14h00 - 15h40, Eden Grove Blue

The affective impact of computer-assisted teaching and learning in statistics

Linda van der Merwe

University of the Free State

Abstract: Computer-assisted Teaching and Learning (CAT/L) has produced profound changes to content, pedagogy, assessment and course management in the field of statistics education. However, little is known about how students experience a computer-assisted intervention in a statistics course, in particular how students' attitudes are affected by the use of technology, whether they enjoy statistics when technology is incorporated into the course, and whether the use of technology affects their learning experience in general.
It is important to investigate the affective impact of CAT/L, i.e. the impact on students' feelings, enthusiasm and attitudes since students' feelings plays a crucial role in how they learn (Barkley 2010:33). However, teachers often disregard the power of the affective domain and rather emphasize the cognitive when designing classroom activities. Consequently, "students' emotions have been the least studied and most overlooked aspect of classroom teaching" (Barkley 2010:33).
This study examined the experiences of students in a first-year statistics class who were exposed to CAT/L by using the spreadsheet tool Excel. In addition to continuous classroom assessment, the students completed a post-course questionnaire. Forty-four Likert scale questions provided quantitative data and six open-ended

questions provided qualitative data on the experience of students exposed to CAT/L. Students' experience with Excel, their attitude towards Excel, their enjoyment of learning statistical concepts and analysing data using Excel, and their opinion about Excel's functions and tools will be discussed.
Key words: Affective domain, Computer-assisted intervention, Excel, Students' learning experience
Session: 4B, Wednesday 08h30 - 10h30, Eden Grove Blue

Bayesian adaptation of the parametric bootstrap

Sean van der Merwe

University of the Free State

Abstract: We consider the problem of goodness-of-fit testing for a model that has at least one unknown parameter that cannot be eliminated by transformation. Examples of such problems can be as simple as testing whether a sample consists of independent Gamma observations, or whether a sample consists of independent Generalised Pareto observations given a threshold. Over time the approach to determining the distribution of a test statistic for such a problem has moved towards on-the-fly calculation post observing a sample. Modern approaches include the parametric bootstrap and posterior predictive checks. We argue that these approaches are merely approximations to integrating over the posterior predictive distribution that flows naturally from a given model. Further, we attempt to demonstrate that shortcomings which may be present in the parametric bootstrap, especially in small samples, can be reduced through the use of objective Bayes techniques, in order to more reliably produce a test with the correct size.
Key words: Hypothesis testing, Objective Bayes, p-value, Predictive posterior, Simulation
Session: 3D, Tuesday 16h10 - 17h50, Eden Grove Seminar Room 2

Preference mapping methods and applications

Marieta van der Rijst[1,⋆], Tormod Næs[2] and Nina Muller[3]

[1]Agricultural Research Council, [2]Nofima Norway, [3]Stellenbosch University

---

Abstract: Preference mapping refers to a group of multivariate statistical methods designed to elucidate consumer preferences within a set of products. Descriptive sensory analysis generates quantitative data that allows the profiling of a product on its perceived sensory characteristics. Chemical/instrumental measurements of products serve to characterise perceived sensory characteristics. Consumers, on the other hand, score elementary aspects of products based on their liking and preference. Preference mapping links sensory and chemical/instrumental profiles to consumer preferences. Numerous statistical methods for preference mapping analyses are available. This paper will present the basics of preference mapping using practical applications to illustrate results.
Key words: Consumer preference, Multivariate statistical methods, Preference mapping
Session: 5C, Wednesday 13h50 - 15h50, Eden Grove Seminar Room 1

The use of point processes and Markov chains to predict delinquency levels for consumers

Frederik van der Walt[1,⋆] and Freek Lombard[2]

[1]ABSA, [2]University of Johannesburg

Abstract: During this talk we will extensively make use of counting processes and the so-called "real-time" approach to survival analysis (Arjas and Haara, 1987). We approach the prediction of consumer defaults by modelling the delinquency levels of consumers. To model the delinquency levels we make use of two different models, both of which are dependent on idiosyncratic- as well as systematic covariates.
Firstly, we model the number of missed payments by a real-time logistic regression type point process (Arjas and Haara, 1987). We combine this point process with a marked point process modelling the number of repayments made (marks being the size of the repayments made). In so doing, we obtain a process that seems to be plausible to model delinquency levels of consumers.
Secondly, we use and adapt the Markov chain approach due to Grimshaw and Alexander (2010). This approach uses a non-stationary Markov chain and estimation of parameters is done by the so-called independent logistic regression technique (Begg and Gray, 1984).
We apply these approaches to a mortgage loan portfolio. Thereafter we compare the two models regarding simplicity and prediction accuracy.
References
Arjas, E. and Haara, P. (1987). A Logistic Regression Model for Hazard: Asymptotic Results. *Scandinavian Journal of Statistics*. 14, 1-18.
Begg, C.B. and Gray, R. (1984). Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*. 71, 11-18.
Grimshaw, S.D. and Alexander, W.P. (2010). Markov chain models for delinquency: Transition matrix estimation and forecasting. *Applied Stochastic Models in Business and Industry*. 27, 267-279.
Key words: Counting processes, Credit risk
Session: 2B, Tuesday 14h00 - 15h40, Eden Grove Blue

Quantitative challenges in building an agent-based model of a colonic crypt

Richard van der Wath

Council for Scientific and Industrial Research

Abstract: In this talk I will present an interactive agent-based model of colonic crypts in the mouse. The model was developed by me during a post-doc at the University of Western Australia with the purpose to contribute to the understanding and treatment of colonic cancer. One of the fastest self-renewal processes found in mammals occurs in the intestinal epithelium which is replaced every 2-3 days in mice and 3-5 days in humans. New functional mature cells are continuously produced by stem cells found in millions of crypts (tubular pits) scattered along the intestinal epithelium. The regulation of this cell production process is very important since malfunction in regulation is closely associated with the formation of cancerous tumours.
The model simulates a single colonic crypt as a rectangular sheet of cells with individual cells represented by semi-autonomous agents that interact with each other and with their local environment. In this talk I will discuss the quantitative challenges faced during development and analysis of the model. The first challenge was how to compute cell displacement as a result of inter-cellular forces caused by cellular growth and division. Secondly was how to build a probabilistic cell cycle model that prevents instantaneous cell division times. Lastly was how to compare and measure performance of models so that we could perform parametric studies.
Key words: Agent-based modelling, Biological simulation, Cell cycle model, Cell mechanics
Session: 3B, Tuesday 16h10 - 17h50, Eden Grove Blue

Matrix variate elliptical model: subjective Bayesian inference

Janet van Niekerk[1,⋆], Andriëtte Bekker[1], Mohammad Arashi[2] and Daan J. de Waal[3]

[1]University of Pretoria, [2]Shahrood University of Technology, [3]University of the Free State

Abstract: The matrix variate elliptical model is considered, where under a subjective Bayes viewpoint estimators for the location matrix and the determinant of the characteristic matrix are exactly derived, with the normal-inverse Wishart and normal-Wishart as priors. In addition, the newly developed results will be applied to the matrix variate normal distribution and the matrix variate t-distribution as particular subfamilies of the matrix variate elliptical model. Algorithms for the simulation of the posterior distributions are developed and a simulation study as well as Fisher's Iris data set are used to illustrate these new results and the usefulness of the normal-Wishart prior.

Key words: Bayesian analysis, Bessel function of matrix argument, Characteristic matrix, Matrix variate elliptical model, Normal-inverse Wishart, Normal-Wishart

Session: 2D, Tuesday 14h00 - 15h40, Eden Grove Seminar Room 2

CUSUM procedures based on signed sequential ranks

Corli van Zyl[1,⋆] and Freek Lombard[2]

[1]North-West University, [2]University of Johannesburg

Abstract: We propose non-parametric cumulative sum (CUSUM) procedures to detect changes from known initial values of the location or scale parameters of a symmetric distribution. Firstly, we motivate the necessity of a non-parametric CUSUM procedure by illustrating the non-robustness to deviations from the normality assumption of the standard normal CUSUM procedure. Secondly, we propose a CUSUM procedure based on signed sequential ranks to detect a change in location and we relate the appropriate offset to the target change size in the location parameter. Thirdly, we also propose a non-parametric CUSUM procedure to detect a change in the scale parameter. The use of the proposed procedures is illustrated by application to some data.

Key words: CUSUM, Distribution free, Sequential rank

Session: 5D, Wednesday 13h50 - 15h50, Eden Grove Seminar Room 2

Immunological responder definition: Tuberculosis clinical research

Elzanne van Zyl

Quintiles

Abstract: In clinical research, tuberculosis (TB) vaccines are often assessed in terms of the immunological response induced. Immunological response can be measured by means of intracellular cytokine staining (ICS) assays. In order to define a subject as having an immunological response (being classified as a responder) a number of aspects should be taken into account, such as Baseline immunological response and data structure (multiple levels of response for example T cell or cytokine response). In order to address all aspects, a responder definition has been derived by utilizing various statistical methodologies including Fisher's exact tests, multiplicity adjustment, odds ratios and Breslow-day test.

Key words: Fisher's exact test, Immunology, Odds ratio

Session: 4D, Wednesday 08h30 - 10h30, Eden Grove Seminar Room 2

Phase I and Phase II control charts for the variance and general variance

Ruaan van Zyl[1,⋆] and Abrie J. van der Merwe[2]

[1]Quintiles, [2]University of the Free State

Abstract: By extracting the results of Human, Chakraborti and Smit (2010), Phase I control charts are derived for the generalized variance when the mean vector and covariance matrix of multivariate normally distributed data are unknown and estimated from m independent samples of n. In Phase II a predictive distribution based on a Bayesian approach is used to construct Shewart-type control limits for the variance and generalized variance. The posterior distribution is obtained by combining the observed data in Phase I and the uncertainty of the unknown parameters via the prior distribution. By using the posterior distribution the unconditional predictive density function can be derived.
Key words: Phase I, Phase II, Shewart control charts, Variance
Session: 6D, Wednesday 16h10 - 17h50, Eden Grove Seminar Room 2

A nonparametric approach to classifying supernova lightcurves

Melvin M. Varughese[1,⋆], Rainer von Sachs[2], Michael Stephanou[1] and Bruce A. Bassett[1,3,4]

[1]University of Cape Town, [2]Universite Catholique de Louvain, [3]African Institute for Mathematical Sciences, [4]South African Astronomical Observatory

Abstract: Type Ia Supernovae are one of the most widely used standard candles in astronomy. That is, they can be used to calculate distances to astronomical objects since their peak brightness is very well-determined. In addition to distance, one may use a spectroscopic reading of a type Ia supernova (or its host galaxy) to determine how fast it is receding from us. Consequently, by examining how the speed of recession varies with the distance to the supernova, it is possible to recreate the expansion history of the universe. This history enables us to estimate cosmological parameters for the early universe.
A serious impediment to utilizing type Ia supernova to learn about the early universe is the probable contamination of the dataset with other supernova types. Any supernovae dataset will consist of multiple types of supernova, where (with the exception of type Ia supernova) we are neither able to determine the peak brightness of the supernova nor the distance to the supernova. It is important to account for this contamination as misclassifying a supernova as a type Ia will severely bias our estimates of the cosmological parameters.
This talk studies how to go about determining the probability of a transient object being a type Ia supernova given a set of flux measurements for the object through time. In the process, we fit a cross-validated spline to the flux measurements. An unbalanced de Haar wavelet basis is subsequently applied in order to characterize the spline curves using a small set of wavelet coefficients. These coefficients may be fed into a standard classification algorithm. The algorithm is trained on 1,200 spectroscopically observed supernovae and is subsequently used to predict the classes of 20,000 test supernovae. The predictions prove to be highly competitive compared to traditional parametric approaches to classify supernovae light curves.
Key words: Astrostatistics, Classification, De Haar wavelets, Splines
Session: 1A, Tuesday 11h30 - 12h10, Eden Grove Red

The consolidation of forecasts with linear regression models

Danie Venter

Nelson Mandela Metropolitan University

Abstract: Research has shown that a consolidated forecast typically outperforms any of the constituent individual forecasts. This study investigated multiple linear regression analysis as a consolidation technique. A VBA application, named ConFoRM – acronym for Consolidate Forecasts with Regression Models – was developed on a Microsoft Excel platform to provide an application with a user-friendly interface for the consolidation of forecasts with regression models. ConFoRM applies various loss functions, both symmetrical and asymmetrical, and a range of linear regression models to both the raw historical forecasts data and two transformations thereof aimed at obtaining a stationary time-series with removal of time-period fluctuations . The performance of ConFoRM was tested with both simulated and real data. Based on the results it can be concluded that the ConFoRM consolidated forecasts typically outperform any individual forecast, the mean of the forecasts and the weighted mean of the forecasts where the weights are based on the inverse of the correlation between a forecast and the variable being forecasted.
Key words: Forecasts consolidation, Linear regression, Loss functions
Session: 4A, Wednesday 08h30 - 10h30, Eden Grove Red

Comparing peak over threshold models in terms of the extreme value indexes

Andréhette Verster

University of the Free State

Abstract: The objective here is to investigate the behaviour of the Extreme Value Index (EVI) of Peaks Over Threshold (POT) models through a classical and Bayesian approach. The Generalized Pareto distribution (GPD) is the most common and popular POT model. If the parameters of the GPD are estimated effectively one can use these estimates to estimate tail probabilities and tail quantiles. A different POT model called the Extended Pareto distribution (EPD) is also investigated to see whether it results in more stable EVI estimates than those obtained from GPD modelling, which might result in more reliable high quantile estimations.
Key words: Bayesian approach, Extreme value index, Peaks over threshold
Session: 2A, Tuesday 14h00 - 15h40, Eden Grove Red

Calibrating option pricing models to down-and-out barrier call option prices

Jaco Visagie

North-West University

Abstract: Schoutens *et al.* (2004) calibrates various option pricing models to a set of European call option prices. They demonstrate that the calibrated models are capable of nearly perfectly replicating these European call option prices. However, when the calibrated models are used to calculate exotic option prices the resulting prices differ significantly. The down-and-out barrier call option is a popular example of an exotic option. We show that it is possible to calibrate various option pricing models to realistic down-and-out barrier call option prices so that the prices of the options are replicated nearly perfectly. The models used are the Heston model as well as exponential Lévy models.
References
Schoutens, W., Simons, E. and Tistaert, J. (2004). A perfect calibration! Now what? *Wilmott magazine*, 8, 66-78.
Key words: Calibration, Down-and-out call option, Exponential Lévy models, Heston model
Session: 2C, Tuesday 14h00 - 15h40, Eden Grove Seminar Room 1

Probit and skew robit models for use in sequential regression multiple imputation

Michael J. von Maltitz

University of the Free State

Abstract: This paper examines a Bayesian estimation technique for the traditional probit model and a novel skew robit model, so that both can be incorporated into sequential regression multiple imputation (SRMI). The need for a skew robit model in place of the probit is also addressed.
Key words: Bayesian estimation, Probit, Sequential regression multiple imputation, Skew robit
Session: 2A, Tuesday 14h00 - 15h40, Eden Grove Red

Longitudinal cohort analysis of student progression in South African higher education system: The case of University of Limpopo

Negussie Yibas

University of Limpopo

Abstract: Cohort members are individual students admitted to any higher leaning institutions for the first time at a given point in time, usually at the beginning of each academic year, and an entry point, any South African Higher Learning Institution, with the aim of acquiring a qualification (Tertiary level degree and diploma) in a program of choice. It is a common practice that all cohort members do not progress equally. An individual cohort member progresses uninterrupted to the end and obtain the planned qualification; or remain behind but stays within the system to eventually obtain a qualification; or may drop out of the cohort to join the Higher Education System elsewhere or leaves the system without any qualification. It is of great importance to track students' progress through higher education system. By doing so education authorities and policy makers at institutional and national level will be able to identify and measure the proportion of students who complete their study uninterrupted, remain behind but still progress within the system at the point of entry, dropout completely from the system or exit from the entry point to re-join the system elsewhere. Student progression has a direct implication on student throughput, higher education funding, effectiveness and efficiency which are central concerns for higher education sector of the country.
In broad terms longitudinal cohort analysis procedures help follow up students' performance and progression within the system in general and the entry point in particular. This paper through longitudinal cohort analysis procedure attempts to track the progress of cohort members who enter the higher education system via the University of Limpopo, first entry point. Data on five cohorts from 2008 to 2012 from the University Information Management System will be used to estimate parameters that indicate the rate of progression from one level to another; dropout rate, throughput rate and the effect of overstaying on student admission and higher education participation rate. Comparison of progression rates will also be made across different cohorts and faculties.
Key words: Cohort analysis, Cohort members, Higher education, Progression rates
Session: 7B, Thursday 08h30 - 10h50, Eden Grove Blue

Attitude survey analysis

Nombuso Zondo

University of KwaZulu-Natal

Abstract: In more developed countries, it has been shown that fear of Statistics is a real problem at Higher Education Institutions, resulting in fewer students pursuing studies in Statistics. In South Africa Statistics was only recently introduced into the school syllabus, where it is called "Data Handling", so that many students in South Africa do not make the link between this area of school Mathematics and the choice of pursuing studies in Statistics at university. It may thus be that school leavers in South Africa are relatively unaware of Statistics when they enter university and classically only pursue subjects that they feel positively about from school.

South Africa is experiencing a dire shortage of Statisticians at all levels, resulting in various attempts at building capacity in the discipline. In-coming students to institutions of higher education in South Africa could possibly be far better prepared for, and guided through, courses in Statistics, if more is known about their attitude towards statistics when they enter university.

This study uses the Survey of Attitudes Towards Statistics (SATS-36) scale to investigate the attitudes of students towards Statistics, for all students registered for an introductory module in Statistics at UKZN. An attempt is made to identify the factors associated with the various attitudes of students in this module, which is then compared to the findings from previous similar international studies. Furthermore, an attempt is made to find a relationship between the attitudes of students towards the mentioned course and their exam results.

Key words: Attitudes, Confirmatory, Exploratory, Factors, Statistics

Session: 7D, Thursday 08h30 - 10h50, Eden Grove Seminar Room 2

New insights into the HIV epidemic in South Africa

Khangelani Zuma*, Olive Shisana, Thomas Rehle, Leickness Simbayi and Sean Jooste

Human Sciences Research Council

Abstract: South Africa has conducted four repeated cross-sectional HIV population-based household surveys since 2002. A total of 38 431 respondents were interviewed from a multi-stage stratified cluster sample of 11 079 valid households that agreed to participate in the survey with 28 997 (67.5%) tested for HIV, HIV incidence and antiretroviral treatment (ART) exposure. HIV prevalence increased from 10.6% in 2008 to 12.2% in 2012 with females 14.4% consistently more affected than males 9.9% in 2012. Exposure to ART almost doubled from 16.6% in 2008 to 31.2% in 2012 with an estimated 2 million people on treatment. HIV incidence rate among those aged 2 years and older was 1.07% with the highest incidence among black African females aged 20-34 years at 4.5%. These findings show that while South Africa is on the right track there have been worrying increases in most HIV-related risk behaviours. These findings suggest the need for a combination prevention approach which integrates biomedical, behavioural, social, and structural prevention interventions in order to reverse the tide in the fight against HIV in the country.

Key words: Antiretroviral treatment, HIV incidence, HIV knowledge, HIV prevalence, Sexual behaviour

Session: 3A, Tuesday 16h10 - 17h50, Eden Grove Red

12.5 Poster Presentations

Model selection uncertainty and parameter estimation of nonlinear growth models

Rasheed Adeyemi

University of Cape Town

Abstract: This study is to discuss the application of nonlinear growth models to measure the growth data and the selection of best model for growth prediction among the competing candidate models. Six non-linear growth functions were fitted to the South African population data. The nonlinear distribution functions were first fitted using iterative method, so that the process is repeatedly optimized using a predefined stopping rule. The method requires specification of the starting values of the parameters to be estimated, making it more difficult than the linear models. The second objective is to explain and illustrate a method,which interface information theory and mathematical statistics for selection of an estimated best approximate model. An approximating AIC weight is proposed instead of raw AIC or BIC for model selection for the non-nested candidate models. For the population growth forecasts, it was found that empirical distributions performed well as traditional times series polynomial models. The measure of errors considered are based on the differences between the predicted and the actual annual growth rate. It was found out that that the the forecast inaccuracies of the models differ greatly. The accuracy of the simple time series models is better than the accuracy of more complex models.
Key words: Initial value, Nonlinear models, Parameter estimation, South Africa population

Estimating the force of infection from prevalence data: Infectious disease modelling

Yusentha Balakrishna[1,⋆] and Henry Mwambi[2]

[1]Medical Research Council, [2]University of KwaZulu-Natal

Abstract: By knowing the incidence of an infectious disease, we can ascertain the high risk factors of the disease as well as the effectiveness of awareness programmes and treatment strategies. Seven models formulated to estimate the force of infection were discussed and applied to age specific HIV prevalence data based on antenatal clinic attendees from the Vulindlela district in KwaZulu-Natal. The link between the survivor function, the prevalence and the force of infection was demonstrated and generalized linear model methodology was used to estimate the force of infection. Parametric and nonparametric force of infection models were fitted to data from 2009 to 2010. The best fitting model was thereafter applied to data from 2003 to 2010. Despite the general increase in HIV prevalence (from 54.07% in 2003 to 61.33% in 2010), the rate of new HIV infections was found to be decreasing. The results also showed that the age at which the force of infection peaked for each year increased from 16.5 years in 2003 to 18 years in 2010. Farrington's two parameter model for estimating the force of HIV infection was shown to be the most useful. The results obtained emphasised the importance of HIV awareness campaigns being targeted at the 15 to 19 year old age group. The results also suggested that using only prevalence as a measure of disease can be misleading and should rather be used in conjunction with incidence estimates to determine the success of intervention and control strategies.
Key words: Force of infection, HIV, Incidence, Prevalence data

R Convenience functions for service course teachers

Jeremy S. Baxter

Rhodes University

Abstract: There is a wealth of materials and technologies available to assist lecturers teaching introductory statistics courses, in particular statistics service courses. Students, particularly those students residing in student residences, are particularly good at storing previously used printed or electronic course material. Unfortunately some students prefer to search these libraries or the web for the answer to tutorial questions. As a result service course lecturers spend inordinate amounts of time preparing new, but similar, tutorial questions and solutions, particularly if the tutorials sessions are held on different days of the week. There are a number of computer resources available to assist with the preparation of lecture and tutorial materials, including solutions to exercises. This poster will present a number of convenient R functions that generate figures commonly used in teaching introductory statistics or service courses, for example:

- shaded distribution plots that include appropriate annotation which can be used to demonstrate the calculation of various cumulative probabilities for the normal, t, χ^2 and F distributions;
- graphics used to demonstrate the concept and notation as well as calculate the "area to the right" or inverse of cumulative distribution functions commonly employed in these courses;
- graphics used to explain how statistical tables can be used to estimate probabilities where the relevant values are not available on the statistical tables.

The functions demonstrated on this poster are not, currently, available on the internet or in a R package. Please contact the author for the source code should you wish to make use of these functions.
Key words: Graphics, Service course teaching

Comparison of sampling methods for use with Kriging

Michaela Beckley[1,2,⋆] and Schalk Kok[2]

[1]Council for Scientific and Industrial Research, [2]University of Pretoria

Abstract: Many variables of interest are difficult to sample over the entire area of interest based on the conditions (such as weather or access) in that area and/or difficulty in the sampling processes themselves. Thus, one often wishes to sample at a few locations and perform interpolation to estimate the variable at the remaining locations. However, there are instances of data minimisation where the values are known at every location of interest and in order to save on storage space, we wish only to save the values at a few locations and generate the remaining values by means of interpolation when they are required. This study addresses both scenarios when the chosen interpolation technique is Kriging. This study makes use of simple heuristics, the Updated Kriging Variance Algorithm and genetic algorithms. These methods rely on either the true error or the estimated error variance. The data set used is a model generated data set containing values of carbon dioxide flux for the Southern Ocean over a period of a year. The sampling was performed for all methods in two dimensions and for selected methods in three dimensions. The results show that in the case of data minimisation, one should work with a sampling strategy which makes use of the true errors, while alternatively when no data are available, the sampling strategy may make use of the estimated error variance associated with the Kriging technique. This study also shows that the results obtained by the use of the Updated Kriging Variance Algorithm can be rivalled by those obtained from a more simple heuristic which makes use of the maximum estimated error variance.
Key words: Kriging, Maximum estimated error variance, Sampling strategies

A study of synthetic Phase II Shewhart-type control charts for monitoring process location

Tarisai Chimbwa[1,⋆] Marien Graham[1] and Subha Chakraborti[2]

[1]University of Pretoria, [2]University of Alabama

Abstract: In this research report, we discuss the four distinct types of Phase II Shewhart synthetic charts for monitoring process location. We give a thorough review of the parametric and nonparametric synthetic charts that have already been discussed in the literature to this point to monitor location. Synthetic charts have been shown to be a special case of runs-rules charts; however, there is no study that thoroughly investigates the performance of the four types of runs-rules charts and the four types of synthetic charts.
Thus the aim of this study is to:
(1) give a literature review on the four types of synthetic charts,
(2) discuss the run-length properties of each of the synthetic charts, and
(3) compare the run-length performance of the synthetic charts and the runs-rules charts.
Key words: Runs-rules charts, Synthetic charts

Eliciting and combining expert opinion – An overview and comparison of methods

Mutsa Chinyamakobvu[⋆] and Isabelle Garisch

Rhodes University

Abstract: Decision makers have long relied on experts to inform their decision making. Expert judgment analysis is a way to elicit and combine the opinions of a group of experts to facilitate decision making. The use of expert judgment is most appropriate when there is a lack of data for obtaining reasonable statistical results.
The experts are asked for advice by one or more decision maker who faces a specific real decision problem. The decision makers are outside the group of experts and are jointly responsible and accountable for the decision and committed to finding solutions that everyone can live with. The emphasis is on the decision makers learning from the experts.
The focus of this paper is on an overview and comparison of the various elicitation and combination methods available. These include the traditional committee method, the Delphi method, the paired comparisons method, the negative exponential model, the classical model, the histogram technique, using the Dirichlet distribution in the case of a set of uncertain proportions which must sum to one, and the employment of overfitting. The supra Bayes approach, the determination of weights for the experts, and combining the opinions of experts where each opinion is associated with a confidence level that represents the expert's conviction of his own judgment, are also considered.
In this paper the comparison of the various methods involves real-life data where possible.
Key words: Classical model, Decision making, Dirichlet distribution, Expert judgment analysis, Supra Bayes approach

---

Using the double-Poisson distribution to analyse Manchester City's 2011/2012 premier league winning season

Aviwe Gqwaka* and Warren Brettenny

Nelson Mandela Metropolitan University

Abstract: In the 2011/2012 season of the Barclays Premier League, Manchester City won the title for the first time in 44 years. After trailing their fierce rivals, Manchester United, by 8 points with 6 games remaining, Manchester City erased that deficit to claim the championship on the last day of the season. This study focuses on this seemingly unlikely situation and attempts to determine its likelihood. In particular, the last 6 games will be analysed to determine
(a) the number of ways in which the title could have been decided and
(b) the likelihood of the observed sequence of results.
With the use of the Double-Poisson distribution, the most probable outcomes for both of the Manchester team's last 6 games are determined and also, ultimately, the most likely destination of the Barclays Premier League is ascertained.
Key words: Double-Poisson, Premier league, Score prediction

Techniques for background modelling in image analysis

Kwok H. Lau* and Inger Fabris-Rotelli

University of Pretoria

Abstract: The study of background modelling has been prevalent since digital photography. Its purpose is to sieve image data and extract meaningful information; and, in many ways, has diverged into distinguishable mathematical branches. This paper discusses the mixture of Gaussians as well as the kernel density technique for background modelling in image analysis, both methods capable of dealing with multi-modal data, and provides a comparison of the two methods. Other techniques are also briefly discussed.
Key words: Background modelling, Kernel density estimation, Mixture of Gaussians

Modelling of multi-state panel data: The importance of the model assumptions

Thandile Mafu[1,2*] and Chris Muller[2]

[1]Agricultural Research Council, [2]Stellenbosch University

Abstract: Multi-state modelling has developed as the technique of choice when modelling panel or longitudinal data – data that include units that are observed across two or more points in time. A continuous time stochastic process is assumed to govern the multi-state process through its transition probabilities and transition rates. Estimating these transition probabilities or rates of the stochastic process lies at the heart of multi-state modelling.
Three assumptions that are typically made regarding the transition rates before fitting a multi-state model are:

1. Homogeneity of the transition rates through time.
2. Homogeneity of the transition rates across the subject population.
3. The Markov assumption – the transition rates only depend on the history of the process through the current sate.

Various authors have put forward methods to assess these assumptions before fitting a multi-state model. Unfortunately, as with many statistical techniques that have underlying assumptions, these methods are not always used to assess if these assumptions are valid before fitting a multi-state model. In this presentation the results of a simulation study is presented where the importance of these three assumption are assessed. Simulated panel data sets are generated where these assumption are specifically violated. Standard multi-state model are then fitted to these data sets and the results obtained are discussed.
Key words: Markov process, Multi-state modelling, Panel data

Data envelopment analysis as a tool for assessing operations in Eastern Cape ports

Barend Mienie⋆, Gary D. Sharp and Warren Brettenny

Nelson Mandela Metropolitan University

Abstract: The marine transportation industry is directly influenced by economic growth. As Africa, is experiencing this growth, traffic through ports has increased. Rapid globalisation of the shipping lines means that the shipping lines have a wide variety of ports to choose from when docking. Competition among ports to handle the cargo associated with this traffic is therefore competitive. This competition has made operational efficiency in ports essential to getting an advantage over competing ports. An efficient port will save both money and the time of the respective shipping lines. In this research, the efficiency of the ports will be analysed by Data Envelopment Analysis (DEA) models. The research aims to evaluate the efficiency of the NCT (Ngqura Container Terminals) and PECT (Port Elizabeth Container Terminals) in the Eastern Cape. The analysis hopes to identify the sources of any inefficiencies so that these inefficiencies may be remedied in an attempt to better compete for Africa's growing maritime traffic.
Key words: Competition, Container ports, DEA, Eastern Cape, Efficiency

---

A new bivariate beta model for multiple shifts in a sequential normal process

Albert Mijburgh⋆, Karien Adamski, Andriëtte Bekker, Schalk W. Human

University of Pretoria

Abstract: Practitioners in the quality control environment are mainly concerned with the detection and identification of shifts in the production process. Quesenberry presented Q-charts as a method of monitoring the production process, assuming that the observations from each sample are independent and identically distributed (i.i.d) normal random variables. This paper proposes a bivariate beta distribution that is applicable when the process variance encounters two successive shifts.
Key words: Bivariate beta, Chi-squared , Shift in process variance, Statistical process control

---

Modelling summer daily peak loads in South Africa using discrete time Markov chain analysis

Molete Mokhele⋆ and Caston Sigauke

University of the Witwatersrand

Abstract: Electricity demand exhibit fast changes and a large degree of randomness in South Africa, particularly summer. Its description requires a detailed analysis using stochastic processes. The project presents a discrete Markov chain analysis to determine stationary distributions of large summer daily changes in peak electricity demand. Such large changes pose challenges to system operators in scheduling maintenance of their plants, refurbishments of their power plants and dispatching of electrical energy. Electricity demand in summer is mainly due to meteorological factors and usage of major appliances. Using the discrete time Markov chain analysis on the Eskom data of the period of 9 years (2000-2009), we develop the probability

transition matrix and the steady-state probabilities of the electricity daily peak changes. We adopt the two state discrete time Markov chain problem: the two states are the increases and decreases on the daily peak changes. We extend the two-state problem to the three-state discrete Markov chain problem, where the states are extreme increases, small increases and decreases on the daily peak demands. Empirical results show that the steady state probabilities for an increase are 0.5972239 for the two-state problem with a return period of 1.67 days. For the three-state problem the steady state probability of an extreme increase is 0.05234024 with a return period of 19.1 days, giving approximately 19 days of extreme increases in the summer period. Such an analysis is important for planning, load shifting, load flow analysis and scheduling of electricity particularly during peak periods in summer.

Key words: Daily peak electricity demand, Discrete time Markov chain, Mean return time, Steady state probability, Transition matrix

---

The South African yellow maize price: A statistical analysis

Marion Muhl* and Janet van Niekerk

University of Pretoria

Abstract: The literature on the relationship between oil prices and agricultural commodity prices is scarce for emerging markets and developing countries such as South Africa. The relationship between the two has been found to vary across commodities and time intervals: it was found that the crude oil price and maize and soybean prices were cointegrated during the time periods 2006 and 2007 respectively, but not during the time period between 2003 and 2005. Soytas and Nazlioglu state that no clear consensus on the transmission of oil price shocks to individual agricultural markets and commodities has been found yet. This paper contributes to the scarce literature for the emerging market of South Africa by analyzing the stationarity of the agricultural commodity (yellow maize) and the crude oil price, as well as establishing a significant relationship between the two aforementioned variables.

Key words: Crude oil, Infinite memory, Maize, Non-stationary, Time series, Unit root

---

The bootstrapped mean time to survival

Anel Oppel* and Theodor Loots

University of Pretoria

Abstract: We estimate the mean time to survival of the Gastro-intestinal Tumor study group data using a bootstrapped approach. This enables us to do hypothesis testing about the mean survival time and also obtain the standard deviation of the bootstrap estimate in a way which requires more computation but less analysis of the specific problem.

Key words: Bootstrap, Mean survival time

Biplots for sparse partial least squares

Opeoluwa F. Oyedele* and Sugnet Lubbe

University of Cape Town

Abstract: In multivariate statistics, the graphical display of data sets is often of interest. Now and again these data sets are large, and therefore, it is difficult to describe the structure of the data, as well as make visual inspection of the data. Solutions such as regression analysis and the biplot can be useful tools in this situation. The latter is useful for solving the visual inspection problem, while the former is useful for describing and evaluating the relationships of the variables in the data. Since data sets can vary from large to

very large, many forms of regression analysis have been developed, such as the Multivariate Multiple Linear Regression (MMLR), Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR). The biplot can also be used as a graphical tool for exploring these relationships.
Regression analysis and the biplot have been merged together in the past to form the eminent *regression biplot*, see Greenacre (2010). Oyedele and Lubbe (2013) recently extended the biplot employment to PLSR, in the form of the *Partial Least Squares (PLS) biplot*. This biplot provides a single graphical display for visualizing the PLSR analysis of a data set, among other advantages.
In this paper, the biplot employment is extended to a sparse version of PLS, in the form of the *Sparse Partial Least Squares (SPLS)* biplot. This sparse version is obtained through an application of the Singular Value Decomposition (SVD) to PLS and a sparsity constrained soft-thresholding penalization. In addition, as a visual tool for displaying the SPLS of a data set, the PLS biplot is proposed and an application of the resulting biplot is provided.
References
Greenacre, M. J. (2010). *Biplots in Practice*. Fundación BBVA: Barcelona, Spain.
Oyedele, O. F. and Lubbe, S. (2013). The Construction of a Partial Least Squares Biplot. *Submitted to the Journal of Applied Statistics*.
Key words: Biplot, Partial least squares biplot, Partial least squares regression, Sparse partial least squares

---

Matching priors for linear functions of Poisson parameters

Lizanne Raubenheimer[1,⋆] and Abrie J. van der Merwe[2]

[1]Rhodes University, [2]University of the Free State

Abstract: The probability matching prior for a linear contrast of Poisson parameters is derived. This prior is extended in such a way that it can be used as the probability matching prior for the average of Poisson parameters. A comparison is made between the confidence intervals obtained by Stamey and Hamilton (2006) and the intervals derived by us when using the Jeffreys' and probability matching priors. A simulation study will be considered. A weighted Monte Carlo method is used for the probability matching prior.
References
Stamey, J. and Hamilton, C. (2006). A Note on Confidence Intervals for a Linear Function of Poisson Rates. *Communications in Statistics: Simulation & Computation*, 35, 849 – 856.
Key words: Bayesian intervals, Poisson parameters, Probability matching prior, Weighted Monte Carlo method

---

Statistical forecasting modelling for JSE stock prices

Xhanti Xotyeni

Nelson Mandela Metropolitan University

Abstract: In the field of finance, the efficient market hypothesis proposed by Fama has had a great influence on financial theory and practice. The efficient market hypothesis is based on whether newly generated information is instantaneously and sufficiently reflected in stock prices. Depending on time of availability this information can be classified into three types: historical information, public information, and future (or internal) information. The efficient market hypothesis can be divided into three types: weak form efficiency for historical information, semi-strong form for public information efficiency and strong form efficiency for future information (Mitra, 2012:111; Eom, Choi, Oh and Jung, 2008:2).
Key words: Financial time-series, Hurst exponent, Predictability

---

13 List of Delegates

A

Dr Franck Adekambi	University of Johannesburg	fadekambi@uj.ac.za
Mr Rasheed Adeyemi	University of Cape Town	adeyemira@yahoo.ca
Prof Marc Aerts	Hasselt University	marc.aerts@uhasselt.be
Mr Edmund Ahame	Nelson Mandela Metropolitan University	s207058655@nmmu.ac.za
Prof James Allison	Nort-West University	james.allison@nwu.ac.za
Dr Zerrin Aşan Greenacre	Anadolu University	zasan@anadolu.edu.tr
Dr Dawit Ayele	University of KwaZulu-Natal	ejigmul@yahoo.com

B

Ms Zsuzsa Bakk	Tilburg Universty	z.bakk@uvt.nl
Ms Yusentha Balakrishna	Medical Research Council	4yusen@gmail.com
Ms Princess Ncedisa Bango	Department of Rural Development and Agrarian Reform	ncedisabango1@yahoo.com
Ms Jesca Batidzirai	University of KwaZulu-Natal	Batidzirai@ukzn.ac.za
Mr Jeremy Baxter	Rhodes University	j.baxter@ru.ac.za
Ms Michaela Beckley	Council for Scientific and Industrial Research/ University of Pretoria	michaelabeckley@gmail.com
Prof Andriëtte Bekker	University of Pretoria	andriette.bekker@up.ac.za
Mrs Dorette Bekker	Metropolitan Health Corporate	dbekker@mhg.co.za
Dr Tom Berning	Stellenbosch University	tberning@sun.ac.za
Mr Alexander Boateng	University of Limpopo	siralexboateng@gmail.com
Ms Michelle Botes	Lightstone/ University of Pretoria	michelleb@lightstone.co.za
Mr Warren Brettenny	Nelson Mandela Metropolitan University	warren.brettenny@nmmu.ac.za
Mr Stefan Britz	University of Cape Town	stefan.britz@uct.ac.za
Mrs Erin Bromley-Gans	Uti	Egans@go2uti.com
Mr Divan Aristo Burger	Quintiles	divan.burger@quintiles.com
Mr Jayde Butler	Juta	jbutler@juta.co.za

C

Mr Frikkie Calitz	Agricultural Research Council	CalitzF@arc.agric.za
Mr Cornu Campher	ABSA/ University of the Witwatersrand	Cornu.Campher@absa.co.za
Mr Niladri Chakraborty	University of Pretoria	nldr_chkrbrty@yahoo.co.in
Mr Retius Chifurira	University of KwaZulu-Natal	chifurira@gmail.com
Dr Delson Chikobvu	University of the Free State	chikobvu@ufs.ac.za
Ms Tarisai Chimbwa	University of Pretoria	tariecee@gmail.com
Mr Knowledge Chinhamu	University of KwaZulu-Natal	chinhamu@ukzn.ac.za
Mr Amos Chinomona	Rhodes University	a.chinomona@ru.ac.za
Ms Mutsa Chinyamakobvu	Rhodes University	m.chinyamakobvu@ru.ac.za
Mr Raymond Chiruka	University of Fort Hare	rchiruka@ufh.ac.za
Mr Ignatious Chitekedza	Nelson Mandela Metropolitan University	Ignatious.Chitekedza@nmmu.ac.za
Mr Allan Clark	University of Cape Town	allan.clark@uct.ac.za
Ms Chantelle Clohessy	Nelson Mandela Metropolitan University	chantelle.clohessy2@nmmu.ac.za
Mrs Marike Cockeran	Nort-West University	marike.krugell@nwu.ac.za
Prof Willie Conradie	Stellenbosch University	wjc@sun.ac.za
Ms Ivona Contardo-Berning	Stellenbosch University	ivona@sun.ac.za
Dr Gretel Crafford	University of Pretoria	gretel.crafford@up.ac.za
Ms Tanita Cronje	University of Pretoria	tanita.cronje@up.ac.za
Prof Nico Crowther	University of Pretoria	nico.crowther@up.ac.za

D

Mr Timotheus Darikwa	University of Limpopo	Timotheus.Darikwa@ul.ac.za
Dr Sonali Das	Council for Scientific and Industrial Research	sdas@csir.co.za
Prof Riaan de Jongh	North-West University	riaan.dejongh@nwu.ac.za
Dr Jacques de Klerk	North-West University	deklerkjacques@yahoo.co.uk
Mr Murray de Villiers	SAS	Murray.deVilliers@sas.com

Prof Tertius de Wet	Stellenbosch University	tdewet@sun.ac.za
Dr Pravesh Debba	Council for Scientific and Industrial Research	pdebba@csir.co.za
Dr Legesse Kassa Debusho	University of Pretoria	legesse.debusho@up.ac.za
Mr Tadele Diriba	University of Pretoria	tadthstat@yahoo.com
Mr Greg Distiller	University of Cape Town	Greg.Distiller@uct.ac.za
Mrs Nontembeko Dudeni-Tlhone	Council for Scientific and Industrial Research	ndudenitlhone@csir.co.za
E		
Dr René Ehlers	University of Pretoria	rene.ehlers@up.ac.za
Dr Sebnem Er	University of Cape Town	sebnem.er@uct.ac.za
F		
Dr Inger Fabris-Rotelli	University of Pretoria	inger.fabris-rotelli@up.ac.za
Prof Paul Fatti	University of the Witwatersrand	paulfatti@gmail.com
Mr Johan Ferreira	University of Pretoria	johan.ferreira@up.ac.za
Prof Maxim Finkelstein	University of the Free State	FinkelM@ufs.ac.za
Dr Lizelle Fletcher	University of Pretoria	lizelle.fletcher@up.ac.za
Mrs Dorette Funke	University of Pretoria	dorette.funke@standardbank.co.za
G		
Ms Raeesa Ganey	University of Cape Town	raeesaganey@hotmail.com
Dr Isabelle Garisch	Rhodes University	I.Garisch@ru.ac.za
Mr Dibaba Gemechu	University of Pretoria	diboobayu@gmail.com
Prof Ivette Gomes	University of Lisbon	migomes@fc.ul.pt
Dr Victoria Goodall	Nelson Mandela Metropolitan University	victoriagoodall@gmail.com
Mr Aviwe Gqwaka	Nelson Mandela Metropolitan University	aviwegqwaka@yahoo.com
Prof Michael Greenacre	Universitat Pompeu Fabra	michael.greenacre@gmail.com
Prof Rangan Gupta	University of Pretoria	rangan.gupta@up.ac.za
H		
Ms Jean-Marie Hakizimana	Statistics South Africa	Jean-MarieH@statssa.gov.za
Dr Justin Harvey	Stellenbosch University	jharvey@sun.ac.za
Mr Andries Haywood	University of Pretoria	ashaywood@gmail.com
Ms Annapurna Hazra	University of KwaZulu-Natal	annapurna.hazra@gmail.com
Mrs Gill Hendry	University of KwaZulu-Natal	hendryfam@telkomsa.net
Mrs Jenny Holloway	Council for Scientific and Industrial Research	jhollowa@csir.co.za
I		
Mr Peter Iiyambo	University of Namibia	piiyambo@unam.na
J		
Dr Mardi Jankowitz	University of South Africa	jankomd@unisa.ac.za
Mr Stefan Janse van Rensburg	Rhodes University	s.jansevanrensburg@ru.ac.za
Dr Helen Jansen van Rensburg	North-West University	11672196@nwu.ac.za
Dr Adelino Juga	Eduardo Mondlane University	adelino.juga@gmail.com
K		
Dr Gaëtan Kabera	Medical Research Council	gkabera@mrc.ac.za
Mr Busanga Jerome Kanyama	University of South Africa	kanyabj@unisa.ac.za
Ms Mmanate Gloria Kekana	Statistics South Africa	mmanatek@statssa.gov.za
Mr Magakwe Jan Kgope	University of South Africa	mjkgope@gmail.com
Prof Martin Kidd	Stellenbosch University	mkidd@sun.ac.za
Dr Cliff Richard Kikawa	Tshwane University of Technology	richard.kikawa@gmail.com
Dr Maggie Kisaka-Lwayo	Statistics South Africa	maggiel@statssa.gov.za
Dr Melisa Koorsse	Nelson Mandela Metropolitan University	melisa.koorsse@nmmu.ac.za
Ms Amorette Kruger	Quintiles	Ret.Kruger@gmail.com
Dr Nirpeksh Kumar	University of Pretoria	nirpeksh@gmail.com
Mrs Lungelwa Kutu	Department of Rural Development and Agrarian Reform	Lungelwa.kutu@drdar.gov.za
L		
Dr Miguel Lacerda	University of Cape Town	Miguel.Lacerda@uct.ac.za
Dr Morné Lamont	Stellenbosch University	mmcl@sun.ac.za

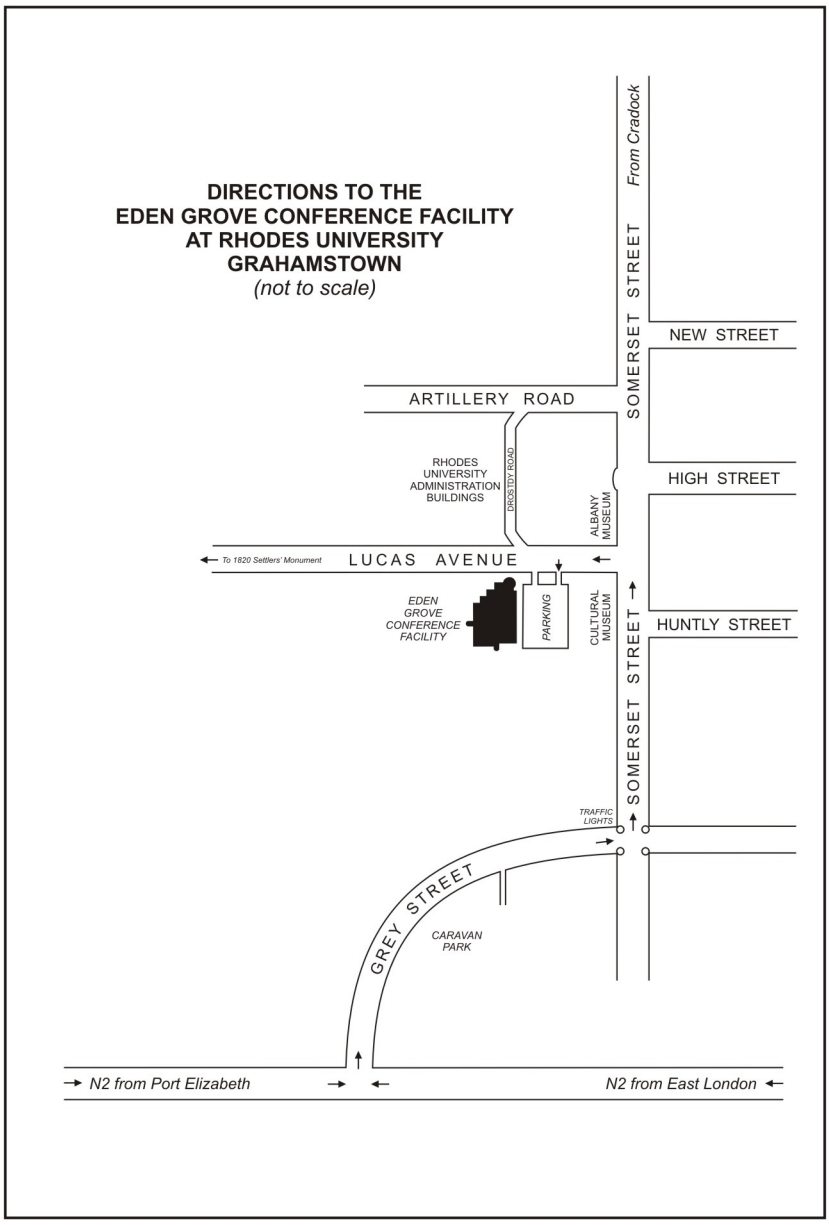
Mr Alex Lau	University of Pretoria	alexlau1712@gmail.com
Prof Niel le Roux	Stellenbosch University	njlr@sun.ac.za
Dr Kerry Leask	CAPRISA	kerry.leask@caprisa.org
Prof Hoffie Lemmer	University of Johannesburg	hoffiel@uj.ac.za
Mr Marothi Peter Letsoalo	University of Limpopo	peter.letsoalo@ul.ac.za
Prof Francesca Little	University of Cape Town	francesca.little@uct.ac.za
Prof Igor Litvine	Nelson Mandela Metropolitan University	igor.litvine@nmmu.ac.za
Mr Ziyaad Loonat	Nedbank	ziyaadl@nedbank.co.za
Mr Theodor Loots	University of Pretoria	theodor.loots@up.ac.za
Mr Osvaldo Loquiha	Universidade Eduardo Mondlane	osvaldo.loquiha@uem.mz
Dr Siaka Lougue	University of KwaZulu-Natal	louguesiaka@yahoo.fr
Prof Sugnet Lubbe	University of Cape Town	Sugnet.Lubbe@uct.ac.za
M		
Dr Iain MacDonald	University of Cape Town	iain.macdonald@uct.ac.za
Dr Rhoderick Machekano	Stellenbosch University	rhoderick@sun.ac.za
Mr Thandile Mafu	Agricultural Research Council	mafut@arc.agric.za
Ms Thandulwazi Magadla	Council for Scientific and Industrial Research	tmagadla@csir.co.za
Mr Sibusiso Vusi Magagula	University of South Africa/ Nedbank	sibusisomag@nedbank.co.za
Mr Thulani Phiwokuhle Magida	Department of Rural Development and Agrarian Reform	Thulani.Magida@drdar.gov.za
Mr Nhlanhla Makhanya	University of Pretoria	nhlanhla2007@gmail.com
Mrs Sibusisiwe Makhanya (Khuluse)	Council for Scientific and Industrial Research	SMakhanya@csir.co.za
Ms Rhoda Makhwiting	Statistics South Africa	RhodaMak@statssa.gov.za
Mr Paul Malan	Quintiles	paul.j.malan@quintiles.com
Mr Happy Maluleke	University of Limpopo	happy.maluleke@ul.ac.za
Mr Michael Mampana	University of South Africa	mampam1@unisa.ac.za
Ms Mulalo Annah Managa	University of South Africa	managma@unisa.ac.za
Mr Siphumlile Mangisa	Nelson Mandela Metropolitan University	Siphumlile.Mangisa2@nmmu.ac.za
Ms Lucia Mans	Quintiles	lucia.mans@quintiles.com
Ms Thina Maqubela	Rhodes University	t.maqubela@ru.ac.za
Mr Thanyani Maremba	Statistics South Africa	thanyanimar@statssa.gov.za
Mr Andries Masenge	University of Pretoria	andries.masenge@up.ac.za
Ms Lehlogonolo Masenya	Statistics South Africa	lehlogonoloma@statssa.gov.za
Ms Lieketseng Masenyetse	Medical Research Council	lieketseng.masenyetse@mrc.ac.za
Ms Neo Mashamba	Statistics South Africa	NeoMa@statssa.gov.za
Mr Sipho Masimula	Statistics South Africa	siphomas@statssa.gov.za
Mrs Kefiloe Masiteng	Statistics South Africa	KefiloeM@statssa.gov.za
Ms Fezeka Matebeni	Department of Rural Development and Agrarian Reform	matebenifn@gmail.com
Mrs Lyness Matizirofa	University of Johannesburg	lmatizi@yahoo.com
Ms Zanele Mazibuko	Statistics South Africa	zanelemaz@statssa.gov.za
Prof Simos Meintanis	National and Kapodistrian University of Athens	simosmei@econ.uoa.gr
Mr Barend Mienie	Nelson Mandela Metropolitan University	mienieb@gmail.com
Mr Albert Mijburgh	University of Pretoria	albert@mijburgh.co.za
Mr Gezani Richman Miyambu	University of Limpopo	miyambugr@yahoo.com
Mrs Unathi Mlaba	First National Bank	umlaba@fnb.co.za
Mr Farai Mlambo	Nelson Mandela Metropolitan University	farai.mlambo@nmmu.ac.za
Mr Mihlali Mnci	Nelson Mandela Metropolitan University	s207027016@nmmu.ac.za
Mr Tshwari Mohlala	University of South Africa	mohlatp@unisa.ac.za
Mr Molete Mokhele	University of Witwatersrand	mokhelemolete01@gmail.com
Ms Promise Molale	University of Limpopo	promisemolale@yahoo.com
Ms Ntokozo Nthabiseng Molata	Statistics South Africa	nthabisengmol@statssa.gov.za
Mr Khehla Moloi	University of Limpopo	khehla.moloi@ul.ac.za
Ms Kirshnee Moodley	Uti	kmoodley1@go2uti.com

Dr Henri Moolman	Walter Sisulu University	moolman.henri@gmail.com
Mr André Mostert	Sasol Technology Research and Development	andre.mostert@sasol.com
Prof Paul Mostert	Stellenbosch University	pjmos@sun.ac.za
Mr Collen Motsepa	Statistics South Africa	collenm@statssa.gov.za
Mr Modupi Peter Mphekgwana	University of Limpopo	vpetoza@gmail.com
Mr Luvolwethu Mqamelo	Department of Rural Development and Agrarian Reform	Luvolwethu.Mqamelo@drdar.gov.za
Ms Suwisa Muchengetwa	University of South Africa	muches@unisa.ac.za
Mr Innocent Mudhombo	University of South Africa	mudhoi@unisa.ac.za
Dr Chris Muller	Stellenbosch University	cmuller@sun.ac.za
N		
Dr Arulsivanathan Naidoo	Statistics South Africa	Aruln@statssa.gov.za
Mr Llewellyn Naidoo	University of KwaZulu-Natal	lewereeve@hotmail.com
Dr Patrick Naidoo	Statistics South Africa	patrickn@statssa.gov.za
Prof Principal Ndlovu	University of South Africa	ndlovp@unisa.ac.za
Dr Ariane Neethling	Stellenbosch University	ariane_neethling@yahoo.com
Mr Mzabalazo Ngwenya	Agricultural Research Council	NgwenyaM@arc.agric.za
Mrs Alecia Nickless	Council for Scientific and Industrial Research	ANickless@csir.co.za
Mrs Johané Nienkemper-Swanepoel	Stellenbosch University	jnienkie@gmail.com
Prof Peter Njuho	University of South Africa	njuhopm@unisa.ac.za
Mrs Efa Nkgabutle	South African Revenue Services	enkgabutle@sars.gov.za
Mr Muthetho Solomon Nkwinika	Statistics South Africa	MuthethoN@statssa.gov.za
Ms Thobeka Nombebe	North-West University	23800364@nwu.ac.za
Prof Delia North	University of KwaZulu-Natal	northd@ukzn.ac.za
Ms Zandile Nukeri	Statistics South Africa/ University of Limpopo	zandilenu@statssa.gov.za
Mr Thandekile Nyulu	Nelson Mandela Metropolitan University	thandekile.nyulu2@nmmu.ac.za
O		
Mr Chinonso Obichili	National Examination Council	chinonsoarchbishop@yahoo.com
Prof John Olaomi	University of South Africa	olaomjo@unisa.ac.za
Ms Wallina Oosthuizen	University of the Free State	oosthuizenw@ufs.ac.za
Ms Anel Oppel	University of Pretoria	anel.oppel@gmail.com
Ms Opeoluwa Oyedele	University of Cape Town	opeoluwaoyedele@gmail.com
P		
Mr Kevin Panman	North-West University	kevinpanman88@gmail.com
Mr Theo Pepler	Stellenbosch University	tpepler@sun.ac.za
Ms Mmalehu Francin Phaho	University of Limpopo	MMALEHU.PHAGO@ul.ac.za
Ms Gaongalelwe Phakedi	Statistics South Africa	PhakediG@statssa.gov.za
Mr Etienne Pienaar	University of Cape Town	etiennead@gmail.com
Mr Marius Pienaar	Standard Bank	marius.pienaar@standardbank.co.za
Dr Sagaren Pillay	Statistics South Africa	sagarenp@statssa.gov.za
Ms Susan Pitts	University of KwaZulu-Natal	pittssue@gmail.com
Mr Charl Pretorius	North-West University	Charl.Pretorius@nwu.ac.za
R		
Prof Sarah Radloff	Rhodes University	S.Radloff@ru.ac.za
Mrs Aluwani Ramalata	University of Limpopo	aluwani.ntsieni@ul.ac.za
Dr Edmore Ranganai	University of South Africa	rangae@unisa.ac.za
Dr Eva Rapoo	University of South Africa	rapooe@unisa.ac.za
Dr Lizanne Raubenheimer	Rhodes University	L.Raubenheimer@ru.ac.za
Ms Samantha Reade	University of KwaZulu-Natal	samreade@gmail.com
Ms Tarylee Reddy	Medical Research Council	tarylee.reddy@mrc.ac.za
S		
Mr Hassan Sadiq	University of Cape Town	SDQHAS001@myuct.ac.za
Dr Leonard Santana	North-West University	leonard.santana@nwu.ac.za
Prof Robert Schall	University of the Free State	schallr@ufs.ac.za

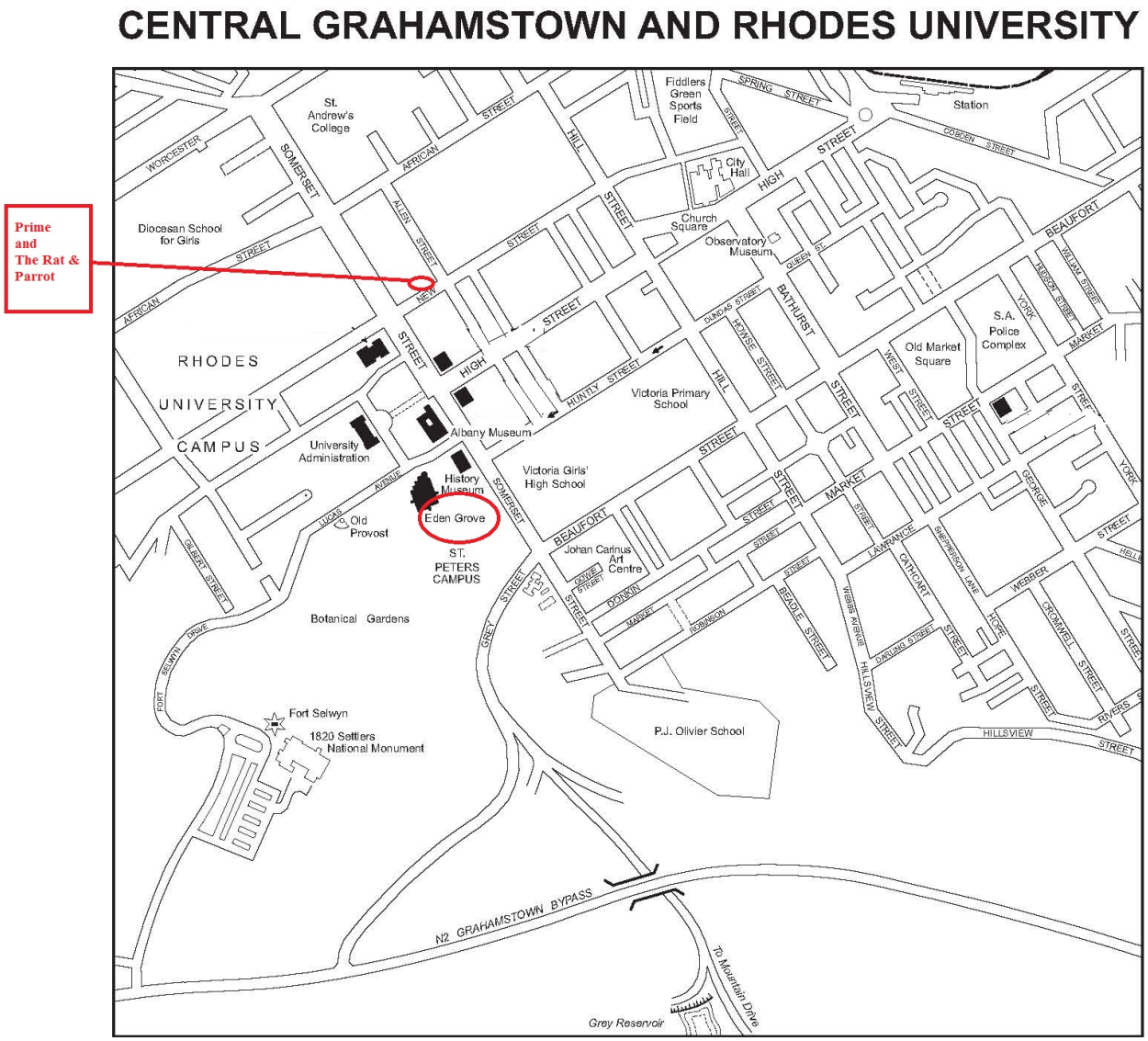
Ms Astrid Scheiber	Casio	astrids@jamesralph.com
Dr WD Schutte	North-West University	wd.schutte@nwu.ac.za
Mr Phuti Sebatjane	Agricultural Research Council	SebatjaneP@arc.agric.za
Prof Solly Seeletse	University of Limpopo	solly.seeletse@ul.ac.za
Prof Gary Sharp	Nelson Mandela Metropolitan University	gary.sharp@nmmu.ac.za
Mr Yegnanew Shiferaw	University of the Witwatersrand	yegna2009@gmail.com
Ms Cleopatra Sikhosana	Statistics South Africa	cleopatras@statssa.gov.za
Ms Glacia Simangwe	Statistics South Africa	glacias@statssa.gov.za
Dr Morné Sjölander	University of the Free State	sjolanderM@ufs.ac.za
Mr Rajab Ssekuma	University of South Africa	ssekur@unisa.ac.za
Dr Michael Stephanou	University of Cape Town	mjsteph@gmail.com
Prof Cornelia Swanepoel	North-West University	cornelia.swanepoel@nwu.ac.za
Prof Jan Swanepoel	North-West University	jan.swanepoel@nwu.ac.za
T		
Mr Kontswi Tshabalala	University of Limpopo	donald.tshabalala@ul.ac.za
U		
Dr Danie Uys	Stellenbosch University	dwu@sun.ac.za
V		
Mr Calven van der Byl	UTi	cvanderbyl@go2uti.com
Prof Abrie van der Merwe	University of the Free State	matheeme@ufs.ac.za
Mr Carel van der Merwe	Ernst & Young	carel.vandermerwe@za.ey.com
Dr Linda van der Merwe	University of the Free State	vdmerwel@ufs.ac.za
Mr Sean van der Merwe	University of the Free State	vandermerwes@ufs.ac.za
Mrs Marieta van der Rijst	Agricultural Research Council	vanderrijstm@arc.agric.za
Dr Frederik van der Walt	ABSA	frederikvdw@gmail.com
Dr Richard van der Wath	Council for Scientific and Industrial Research	rvdwath@csir.co.za
Dr Pieta van Deventer	Stellenbosch University	pjuvd@sun.ac.za
Prof Francois van Graan	North-West University	Francois.VanGraan@nwu.ac.za
Mrs Janet van Niekerk	University of Pretoria	janet.vanniekerk@up.ac.za
Mr Adriaan van Tonder	ABSA	Adriaan.VanTonder@absa.co.za
Ms Corli van Zyl	North-West University	22231609@nwu.ac.za
Mrs Elzanne van Zyl	Quintiles	elzanne.vanzyl@quintiles.com
Mr Ruaan van Zyl	Quintiles	ruaanvz@gmail.com
Dr Melvin Varughese	University of Cape Town	melvin.varughese@uct.ac.za
Dr Danie Venter	Nelson Mandela Metropolitan University	danie.venter@nmmu.ac.za
Dr Andréhette Verster	University of the Free State	verstera@ufs.ac.za
Mr Jaco Visagie	North-West University	20313071@nwu.ac.za
Mr Michael Johan von Maltitz	University of the Free State	vmaltitzmj@ufs.ac.za
X		
Mr Gcobani Xotyeni	Department of Rural Development and Agrarian Reform	Gcobani.xotyeni@drdar.gov.za
Mr Xhanti Mpilo Xotyeni	Nelson Mandela Metropolitan University	s210084774@nmmu.ac.za
Mr Thandolwethu Xozwa	Nelson Mandela Metropolitan University	s208109047@nmmu.ac.za
Y		
Mr Negussie Yibas	University of Limpopo	negussie.yibas@ul.ac.za
Z		
Mr André Zitzke	SAS	Andre.Zitzke@sas.com
Ms Nombuso Zondo	University of KwaZulu-Natal	nombusogm@gmail.com
Dr Khangelani Zuma	Human Sciences Research Council	kzuma@hsrc.ac.za

14 Maps

- **Conference Venue**



- **Grahamstown Street Map**



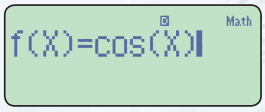
CASIO® FX-82ZA PLUS

The Learning Tool of Choice.

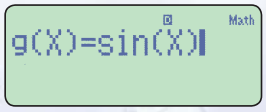
Developed in Collaboration with South African Educators for the SA Syllabus

Natural Textbook Display Scientific Calculator

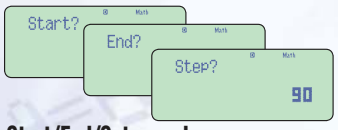
Now with Dual Table Function



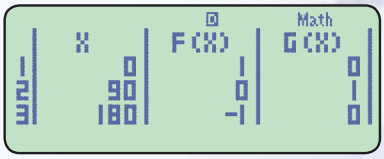
Formula registration



2nd formula registration

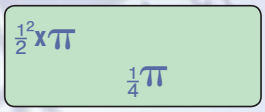


Start/End/Set up values



Resultant dual table

Plus Dual Result Function



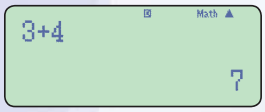
Math result format

Linear result format

Results can be expressed in standard form or decimal

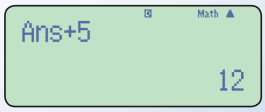
Plus Dual Answer Key to show your previous answer

3 + 4 =



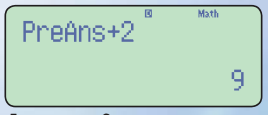
Answer = 7

Answer+5=



Answer = 12
Previous Answer = 7

Previous Answer+2=

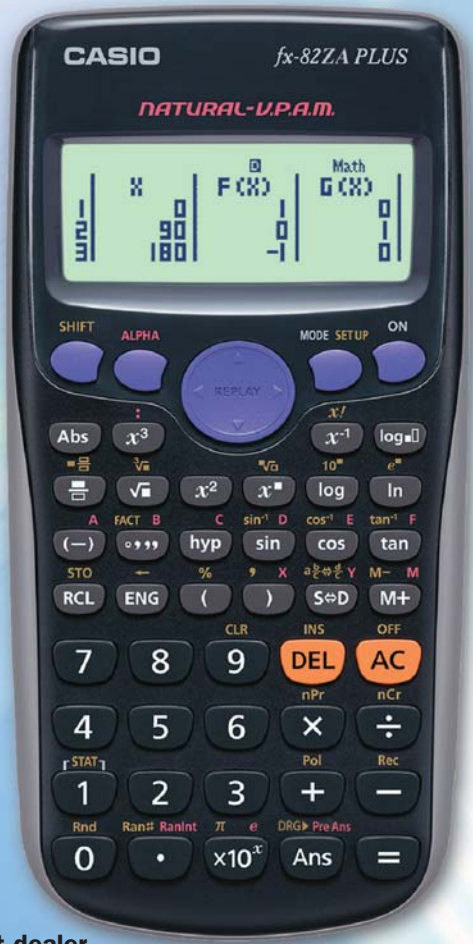


Answer = 9
Previous Answer = 12

Includes These Important Functions

- **Previous answer key for recursive functions**
- **9 memories**
- **Statistics – data handling and regression analysis**
- **Random integers for statistics**
- **Prime factorisation**
- **Time conversion calculations**
- **Co-ordinate transformation – enhanced by dual table function**
- **Selectable auto power off (10 min / 60 min)**

Plus another 246 functions



CASIO means TECHNOLOGY

JAMES RALPH (PTY) LTD

Exclusive distributors for Casio in Southern Africa

See our website for more details: http://www.jamesralph.com

or phone us on 011 314 8888, 0861 CASIO1 or 0861 318 888 for your nearest dealer

PARTNER with us
Realise the POTENTIAL

Potential maps the way forward. It shapes the future. Our aim is to develop it.

Through the funding of scientific and technological research, development of human capacity and provision of research facilities, the National Research Foundation is helping to build South Africa's potential in the global knowledge economy.

As the largest science agency in Africa, we work closely with the private and public sector, universities and other higher education institutions in building the country's research infrastructure through grants and bursaries; funding new and existing scientists; and exploring new and indigenous knowledge for the benefit of all South Africans.

The NRF, Helping realise South Africa's full potential.

TEL: +27 12 481 4124
WEBSITE: WWW.NRF.AC.ZA

NRF National Research Foundation

WOULD YOU **BANK ON YOU?**

To find the extra 0.000034% in a US$10bn Transaction? Of course you would, right? But think about it. Would you really trust yourself to see the detail in a complex deal? If you can honestly say, without looking away, you might just be who we're looking for. Look up **www.standardbank.com/careers** for more information.

Authorised financial services and registered credit provider (NCRCP15).
The Standard Bank of South Africa Limited (Reg. No. 1962/000738/06).
Moving Forward is a trademark of The Standard Bank of South Africa Limited. SBSA 3102 10/14

Standard Bank

Moving Forward™

Sponsors and Exhibitors (*in alphabetical order*)

We would like to thank and acknowledge the following sponsors and exhibitors for their contribution to the SASA 2014 conference.

CASIO

MaSS
DST-NRF CENTRE OF EXCELLENCE IN MATHEMATICAL & STATISTICAL SCIENCES

NRF | National Research Foundation

QUINTILES

RHODES UNIVERSITY
Where leaders learn

sas | THE POWER TO KNOW.

SASA · SASV
South African Statistical Association · Suid-Afrikaanse Statistiese Vereniging

Standard Bank

Statistics South Africa
The South Africa I know, the home I understand