science
& technology
Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA

SARChI Research Chair
in Biostatistics

NRF National Research Foundation

saMRC advancing life

AD DESTINATUM PERSEQUOR

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

BMW

ZA Hub
BMW digital & global solutions

INTERNATIONAL SYMPOSIUM ON MODERN BIOSTATISTICS AND STATISTICS

3 – 5 JULY 2023

FUTURE AFRICA

UNIVERSITY OF PRETORIA

MONDAY 3 JULY 2023		**TUESDAY 4 JULY 2023**		**WEDNESDAY 5 JULY 2023**	
8:00 - 8:30	Arrival and Registration	**8:00 - 8:30**	Arrival	**8:30 - 9:00**	Arrival
8:30 - 9:00	Opening	**8:30 - 9:15**	Plenary: Prof Din Chen	**9:00 - 10:30**	WORKSHOP: Prof Coelho
9:00 - 9:50	Plenary: Prof Carlos Coelho				
		9:15 - 9:35	Sphiwe Skosana		
		9:35 - 9:55	Iena Derks		
9:50 - 10:10	Innocent Maposa	**9:55 - 10:15**	Johannes Vorster		
10:10 - 10:30	Jarod Smith	**10:15 - 10:35**	Luke Pieters		
10:30 - 11:00	Tea and coffee	**10:30 - 11:00**	Tea and Coffee	**10:30 - 11:00**	Tea and Coffee
11:00 - 11:20	JP Stander	**11:00 - 11:20**	Claudio Jardim	**11:00 - 12:30**	WORKSHOP: Prof Coelho
11:20 - 11:40	Dylan Strydom	**11:20 - 11:40**	Michelle de Klerk		
11:40 - 12:00	Kabelo Mahloromela	**11:40 - 12:00**	Marie Vogel		
12:00 - 12:30	Prof Tahir Pillay	**12:00 - 12:20**	Nomly Ngubeni		
12:30 - 13:30	Lunch	**12:20 - 13:30**	Lunch	**12:30 - 13:30**	Lunch
13:30 - 14:15	Plenary: Prof James Allison	**13:30 - 13:45**	Sponsor Address: BMW	**13:30 - 15:00**	WORKSHOP: Najmeh et al
		13:45 - 14:30	Plenary: Prof Sheetal Silal		
14:15 - 14:35	Salomi Millard	**14:30 - 14:50**	Rene Stander		
14:35 - 15:30	WORKSHOP (NWU)	**14:50 - 15:10**	Farai Mlambo		
		15:10 - 15:30	Frank Heslop		
15:30 - 16:00	Tea and coffee	**15:30 - 16:00**	Tea and Coffee	**15:00 - 15:30**	Tea and coffee
16:00 - 18:00	WORKSHOP (NWU)	**16:00 - 16:20**	Lindo Magagula	15:30 - 17:00	WORKSHOP: Najmeh et al
		16:20 - 16:40	Azam Kheyri		
		16:40 - 17:00	Matthias Wagener		
		17:00 - 17:20	Renate Thiede		

ABSTRACTS

Plenary Speakers

Prof Carlos Coelho: A Likelihood Ratio Test for high-dimensional MANOVA
Mathematics Department and NOVA Math – NOVA School of Science and Technology
NOVA University of Lisbon, Caparica, Portugal

A Likelihood Ratio Test is developed for the one-way high-dimensional MANOVA, which is able to outperform existing tests, displaying an extraordinary behavior even for discrete distributions, extremely skewed distributions as well as heavy tailed distributions, including those with no expected value, in which case it becomes a test for location. It shows a better Type I error control than existing tests and non-inflated power values. Furthermore, the test presented is able to work with samples of size just 1, for all samples, except one of them. Its statistic has a very nice and simple asymptotic Normal distribution, which is asymptotic for the number of variables (opposite to common asymptotic distributions which are asymptotic for increasing sample sizes), and that as such does not require any restrictions on sample sizes in order to hold. Extended simulation results are presented for a wide range of distributions.

Prof James Allison: On a new class of tests for the Pareto distribution using Fourier methods
North West University

We propose new classes of tests for the Pareto type I distribution using the empirical characteristic function. These tests are U and V statistics based on a characterisation of the Pareto distribution involving the distribution of the sample minimum. In addition to deriving simple computational forms for the proposed test statistics, we prove consistency against a wide range of fixed alternatives. A Monte Carlo study is included in which the newly proposed tests are shown to produce high powers. These powers include results relating to fixed alternatives as well as local powers against mixture distributions. The use of the proposed tests is illustrated using an observed data set.

Prof Ding-Geng (Din) Chen: Bayesian Assurance over Statistical Power in Clinical Trial Design University of Pretoria

Any clinical trial should be well-designed and requires an appropriate sample size with adequate statistical power to address trial objectives. Statistical power is traditionally defined as the probability of rejecting the null hypothesis with a pre-specified true clinical treatment effect. This power is a conditional probability conditioned on the true but actually unknown effect. In practice, this true effect is never fixed as a constant so a newly proposed alternative to this conventional statistical power is Bayesian assurance. The Bayesian assurance is a new paradigm in clinical trial design and is defined as the unconditional probability of rejecting the null hypothesis. It can then be obtained as an expected power where the expectation is based on the prior probability distribution of the unknown treatment effect, therefore it is a Bayesian concept. In this talk, we review the transition from conventional statistical power to assurance and discuss the computations of assurance using Monte-Carlo simulation-based approach.

Prof Sheetal Silal: Statistics and Operational Research: Exploring opportunities to contribute to improved health in South Africa University of Cape Town

Statistics and operational research are critical tools in healthcare decision making worldwide. By analyzing data and using statistical methods, healthcare professionals can identify patterns and trends, assess the effectiveness of treatments, and make informed decisions about patient care. Traditionally, statistics has been used to determine the prevalence of a particular disease, identify risk factors, and evaluate the effectiveness of different treatment options. Where data are available, operational research techniques, such as simulation and optimisation, have been used to improve healthcare delivery and resource allocation, such as optimising staffing levels and hospital bed utilisation. But a range of opportunities exist to apply both statistics and operational research techniques in healthcare provision in South Africa. This talk focuses on presenting example cases, suggesting areas of application with urgent need. This talk will demonstrate that statistics and operational research are essential for effective decision making in healthcare allowing for data-driven decisions that improve health outcomes and reduce costs.

Contributed Talks

Innocent Maposa: Heritability of cardiovascular health measures across three generations of families in Soweto, South Africa
Stellenbosch University

An application of random family method, Bayesian MCMC and Hamiltonian Monte Carlo (HMC) approaches.

Jarod Smith: A Data Driven Bayesian Graphical Ridge Estimator
University of Pretoria

Bayesian methodologies prioritising accurate associations above sparsity in Gaussian graphical model (GGM) estimation remain relatively scarce in scientific literature. The Bayesian adaptive graphical lasso prior is used as a departure point in the formulation of a computationally efficient graphical ridge-type prior for events where accurate associations are prioritised over sparse representations. A novel block Gibbs sampler for simulating precision matrices is constructed using a ridge-type penalisation. The Bayesian graphical ridge-type prior is extended to a Bayesian adaptive graphical ridge-type prior. Synthetic experiments indicate that the graphical ridge-type estimators enjoy computational efficiency, in moderate dimensions, and numerical performance, for relatively non-sparse precision matrices, when compared to their lasso counterparts. The adaptive graphical ridge-type estimator is applied to cell signalling data to infer key associations between phosphorylated proteins in human T cell signalling. All computational workloads are carried out using our R package ”baygel”, which is available on CRAN.

Jean-Pierre Stander: The usefulness of level sets for spatial modelling of images
University of Pretoria

Level-sets of an image are a collection of pixels that have the same pixel intensity and each pixel in the collection is the neighbour of at least one other pixel. Level-sets are flexible regions which contour to edges in the image. These set regions are further flexible in terms of shape since they are created based on their values, and being data-driven, therefore, provide the mechanism for the understanding the image content. An image can be represented as a spatial point pattern by using the centroid of each level-set as the coordinate. The response of each point can be seen as the pixel intensity of this level-set. Other than the various level-set measures, such as a shape measure, can be calculated for each level-set. These locations and observed values can be used as input for multiple models such as graphical models.

Dylan Strydom: Panel data regression models: with specific reference to fixed and random error component models
University of Pretoria

In this research, four different error component models are considered, namely: the one- way error component models for fixed and random effects, as well as the two-way error component models for fixed and random effects. Firstly, the theoretical modelling framework is addressed, followed by a practical application of the four error component models. The practical application on weather patterns considers data collected in the UK; this dataset consists of observations for nine weather stations over the period of January 2006 to December 2020. Finally, this mini-dissertation considers the numerical aspects related to error component models through bootstrapping and Monte Carlo simulation exercises.

Kabelo Mahloromela: Window selection in spatial point pattern analysis
University of Pretoria

The analysis of spatial point pattern data is typically done to expand the basic understanding of the first and second order properties of the point process that generated the data. First and second order properties of spatial point patterns are estimated using density and distance based measures. These measures rely implicitly or explicitly on the specification of the window domain. Thus, the correct specification of a window domain and the use of an appropriate distance metric to quantify proximity on the chosen window has an important role in the analysis of spatial point pattern data. Herein, we consider the influence of window choice on the analysis of first and second order properties of spatial point patterns.

Prof Tahir Pillay: The power of big data analysis in laboratory medicine
University of Pretoria, University of Cape Town

The field of laboratory medicine and pathology has witnessed a remarkable transformation with the advent of big data. The exponential growth in data volume, velocity, and variety has created new challenges and opportunities in the analysis and interpretation of laboratory and clinical data. Statistics plays a crucial role in harnessing the potential of big data for improved decision-making, enhanced patient care, and advanced research in this domain.

This presentation will provide an overview of the role of statistics in big data analysis for laboratory medicine and pathology. Statistical methods provide the necessary tools to analyze and interpret large datasets, enabling researchers and clinicians to extract valuable insights and draw meaningful conclusions. These methods encompass a wide range of techniques, including descriptive statistics, inferential statistics, predictive modeling, and machine learning algorithms.

In laboratory medicine, statistics aids in quality control, assay validation, and reference range determination. Through statistical analysis, laboratories can identify trends, assess assay performance, and ensure accurate and reliable test results. Furthermore, statistical techniques help in identifying outliers, detecting patterns, and uncovering hidden associations within complex datasets, facilitating the discovery of novel biomarkers and disease markers.

In the field of pathology, statistics plays a pivotal role in histopathological image analysis, digital pathology, and molecular diagnostics. Statistical modeling enables the development of predictive algorithms for disease diagnosis, prognosis, and therapeutic response prediction. These models leverage big data to identify patterns and risk factors, enabling pathologists to make more accurate and personalized treatment decisions.

Additionally, statistics supports research in laboratory medicine and pathology by enabling hypothesis testing, experimental design, and data visualization. It helps researchers to draw meaningful inferences from large-scale studies, validate research findings, and translate discoveries into clinical practice.

In conclusion, statistics is a fundamental pillar in big data analysis for laboratory medicine and pathology. Its application empowers researchers, clinicians, and laboratory professionals to effectively utilize the vast amount of data available, leading to improved patient outcomes, enhanced disease understanding, and advancements in medical knowledge. The integration of statistics and big data analytics will continue to shape the future of laboratory medicine and pathology, driving innovation and precision in healthcare.

Salomi Millard: The impact of multicollinearity on the linear regression model when applied to big data
University of Pretoria

Multicollinearity refers to the existence of near/perfect linear relationships between the predictor variables in a regression problem. That is, some predictor variables can be linearly predicted from others with a substantial degree of accuracy. When multicollinearity exists within a dataset the assumption of independence is violated. This violation has severe consequences in small or moderate size datasets. This research aims to examine the impact of multicollinearity on the linear regression model when applied to big data by numerically investigating the bias, variance and signs of the estimated regression coefficients. Extensive simulation studies are conducted to examine the impact of multicollinearity on the linear regression model when applied to big data. We will show that in big data analytics, multicollinearity does not substantially alter statistical measures and elaborate on when multicollinearity can be ignored based on the sample size and the number of variables included in the model.

Sphiwe Skhosana: A new approach to estimate semi-parametric mixtures of partially linear models University of Pretoria

Semi- and non-parametric mixture of normal regression models are a flexible class of mixture of regression models. These models assume that the component mixing proportions, regression functions and/or variances are non-parametric functions of the covariates. Among this class of models, the semi-parametric mixture of partially linear models (SPMPLMs) combine the desirable simplicity of a parametric model and the flexibility of a non-parametric model. However, local-likelihood estimation of the non-parametric component regression functions (CRFs) poses a computational challenge. Traditional EM optimisation of the local-likelihood functions is not appropriate due to the label-switching problem. Separately applying the EM algorithm on each local-likelihood function will likely result in wiggly and non-smooth CRFs. This is because the posterior probabilities calculated at the E-step of each local EM are not guaranteed to be aligned.

In this presentation, we propose a unified approach to address the label-switching problem and also obtain efficient estimates of both the parametric and non-parametric terms. To address label-switching, we propose a novel EM-type algorithm in which the same responsibilities are used at each local M-step so as to guarantee that the former are aligned. For a given set of local posterior probabilities, we simultaneously maximize all the local likelihood functions to estimate the non-parametric CRFs. We repeat this for all the other sets. As our final estimate of the non-parametric CRFs, we select the ones that are the smoothest. To achieve efficient estimation of both the parametric and non-parametric terms, we propose one-step backfitting estimates. We also propose a corresponding onestep backfitting algorithm to obtain the latter estimates. The performance and practical usefulness of the proposed methods are evaluated using a simulated dataset and a real dataset.

Iena Derks: A Glasso-Forward Search for Solving Most Relevant Explanation in Bayesian Networks University of Pretoria

Abductive inference plays a critical role in solving problems involving uncertain information and multiple competing hypotheses. In the context of Bayesian networks, abductive inference involves generating the best configuration of target variables as an explanation for observed evidence. However, the search space for potential explanations can be vast, requiring efficient search algorithms to find optimal configurations. This aligns with the challenges in Explainable Artificial Intelligence, where researchers aim to provide meaningful explanations in an efficient manner. To address this challenge, we propose a Glasso-Forward search algorithm that offers the Most Relevant Explanation for the observed evidence. To manage the large search space, the graphical lasso is applied to shrink the set of potential explanations. By reducing the set of instantiations, the algorithm provides explanations that focus on the most relevant dependencies. Combining a forward search algorithm with the graphical lasso enables simplified explanations, identification of key features, and quantification of uncertainties. We evaluate the performance of the Glasso-Forward search algorithm compared to the forward search algorithm using a collection of benchmark Bayesian networks.

Johannes Vorster: A robust simulation to compare meaningful batting averages in cricket
University of Pretoria

In cricket, the traditional batting average is the most common measure of a cricket player's batting performance. However, the batting average can easily be inflated by a high number of not-out innings. Therefore, in this research eight alternative methods are used and compared to the traditional batting average to estimate the true batting average. It is also known that there is a range of different batters within a cricket team, namely first order, middle order, tailenders and a special class of players who can both bat and bowl known as allrounders. There are also different formats of international cricket, namely Test, One-Day International (ODI), and Twenty20 International (T20I) cricket, where Test cricket has unlimited overs compared to the limited overs of ODI and T20I cricket. A method for estimating the batting average should be able to account for all of this variability. By using the smoothed bootstrap in this study, the variability of each estimation method is compared. Despite increasing levels of participation and popularity in the cricket fraternity, women's cricket has not received the same exposure compared to men's cricket in terms of sports data analytics. Therefore, this study looks at measuring the batting ability of female batters. However, the methods considered and the conclusions reached in this study are also valid for men's cricket.

Luke Pieters: Design, application and implementation of new multivariate memory-type monitoring schemes integrated with machine learning for monitoring simple and general linear profiles
University of Pretoria

Modern statistical process monitoring (SPM) has many tools at its disposal to determine whether a process is in or out-of-control. One of the commonly used tools is the monitoring scheme (or control chart), with the exponentially weighted moving average (EWMA) scheme being one of the most popular ones in the literature when it comes to the fast detection of small and moderate shifts in the process parameter. Improved versions for the EWMA scheme have been introduced recently, such as the extended EWMA (EEWMA) scheme, and the modified EWMA (MEWMA) scheme. A new scheme known as the homogeneously weighted moving average (HWMA) was also introduced into the literature recently and was shown to be superior over the EWMA scheme at detecting small shifts in a process parameters. In a similar way, this study will aim to improve the existing HWMA scheme by introducing new univariate and multivariate extended HWMA (EHWMA) and modified HWMA (MHWMA) schemes for monitoring sample means and individual observations. When there is a functional relationship between the quality characteristic (dependent variable) and one or more independent variables, classical (or standard) monitoring schemes are not recommended. In this instance, regression schemes also known as profile monitoring schemes are used. Thus, this study will also demonstrate how the EHWMA and MHWMA schemes can be used for linear and general profile monitoring. In addition, these schemes will be integrated with machine learning to facilitate the identification of the magnitude of the shift (i.e. small, moderate or large shift) and variable that caused the out-of-control situation. The application and implementation of the proposed scheme will be given using simulated and real-life data.

Claudio Jardim: Feature engineered embeddings for machine learning on molecular data
University of Pretoria

The classification of molecules and the prediction of protein-protein interactions are of particular importance to the drug discovery process and several other use cases. Traditional methods of molecule classification rely on structural and sequence/text data. Several methods such as deep learning are able to classify molecules and predict their protein-protein interactions using both types of data. The use of structural data combined with deep learning requires substantial computational power and extensive training time.

In this study, we present a different approach to molecule classification and protein-protein interaction prediction that addresses the limitations of other techniques. Our approach utilises natural language processing methods such as count vectorisation, term frequency-inverse document frequency, word2vec, and latent Dirichlet allocation, to feature engineer molecular text data. Through this approach, we aim to make a robust and easily reproducible embedding that is solely dependent on chemical (text) data such as the sequence of a protein. Further, we investigate the usefulness of these explainable embeddings for machine learning models, for representing a corpus of data in vector space and for protein-protein interaction prediction using embedding similarity.

We apply our approach to FASTA sequence data and Simplified Molecular Input Line Entry Specification data. Through comprehensive experiments, we show that these embeddings provide excellent performance for classification. Moreover, we extend the application to the challenging and time-consuming process of protein-protein interaction prediction. Our method is able to make large-scale predictions in a short amount of time with exceptional performance in predicting protein-protein interactions through embedding similarity. Overall, our approach demonstrates the use of feature engineered embeddings for reducing computational requirements, efficient reproducibility and improving prediction capabilities.

Michelle de Klerk: Spatial prediction on disjoint spatial lattice data
University of Pretoria

Modeling on spatially disjoint lattice data presents challenges in the determination of appropriate spatial dependency. When considering the feeder areas of points-of-interest based on drive-time to residential areas, spatially overlapping areas will typically be observed in metro areas. Spatially disjoint areas will be identified for households which fall outside of metro drive-time catchment areas. We present an approach for spatial regression making use of covariates, to model on such a spatial data set. The methodology is applied to healthcare points-of-interest considering distance and location as well as sociodemographic covariates (population density, income etc.) and environmental covariates (rainfall, temperature, and proportion of healthcare services in an area). Current applications being investigated include retail sales of over-the-counter medication at pharmaceutical stores and identifying service areas of public hospitals and laboratories.

Marie Vogel: A high resolution human movement model using small area estimation
University of Pretoria

Spatial movement models are important for the improvement of epidemiological models of contagious diseases, such as covid-19. High resolution models are needed for accurate predictive capabilities, but there is often a lack of sufficiently detailed data, or access to it. Small area estimation may come to the rescue, by using higher resolution covariate data together with lower resolution mobile network data, to predict a more accurate picture of human movement patterns.

Nomly Ngubeni: Spatial analysis of the South African rail system
University of Pretoria

This research presents spatial analysis of the South African rail system with a focus on the descriptive aspects of the rail network. The rail system plays a vital role in modern transportation, offering a sustainable and efficient alternative to road travel for goods freight. We look at network size and coverage, spatial distribution and accessibility, travel times and connectivity between network nodes and rail network capacity variables.

By combining the descriptive analysis of the rail network with an examination of measures related to train stations, train routes, and capacity variables, this research offers a comprehensive overview of the state of the rail system. The findings contribute to a better understanding of the network's strengths and weaknesses, providing a foundation for targeted improvements and strategic decision-making in rail transportation planning.

The insights and recommendations derived from this analysis have the potential to inform policymakers, Transnet, and industry stakeholders in their efforts to enhance the efficiency, connectivity, and overall performance of the rail system. Furthermore, this study serves as a valuable reference for future research and analysis on rail network development and optimisation.

Rene Stander: Estimation of variogram on spatial lattice data
University of Pretoria

Geostatistical data is observed from a continuous spatial process. A variogram is used to quantify the spatial variability of such data. Although spatial lattice data is observed from a discrete spatial process, the variogram theory can easily be extended. In this work, we explore different methods for estimating a variogram for spatial lattice data.

Farai Mlambo: Applications of Probabilistic (Bayesian) Machine Learning in Healthcare: Past, Present and Future
University of the Witwatersrand

The advent of machine learning has brought about remarkable transformations in various sectors, with healthcare being one of the most prominent beneficiaries. Particularly, the integration of Bayesian methods with machine learning, forming Probabilistic Machine Learning, has presented an innovative approach to address the inherent uncertainties in healthcare data. This presentation seeks to traverse the journey of Probabilistic Machine Learning applications in healthcare, elucidating its past accomplishments, current applications, and potential future prospects. We begin with a retrospect of Probabilistic Machine Learning's historical development and its initial incursion into the healthcare domain. Tracing its roots from Bayesian statistics, we will discuss how this probabilistic approach has tackled past healthcare challenges and improved decision-making. Subsequently, we delve into contemporary applications of these techniques in various facets of healthcare, including disease diagnosis, genetic research, drug discovery, and patient monitoring. By elucidating current methodologies and outcomes, we aim to provide a comprehensive perspective of the prevailing landscape. Looking forward, the presentation concludes with a speculative gaze into the future. We highlight ongoing research in the field and predict how emerging technologies may shape the application of Probabilistic Machine Learning in healthcare. We discuss potential opportunities and challenges that lie ahead in further harnessing this powerful tool in the quest for improved health outcomes. The audience is invited to join this explorative discourse, traversing the evolutionary timeline of Probabilistic Machine Learning in healthcare. We believe this presentation will foster a better understanding of the potentiality of these methods, contributing to the overall objective of the symposium to unify modern methodological research in biostatistics and statistics. Keywords: Probabilistic Machine Learning, Bayesian Methods, Biostatistics, Healthcare Applications, Data Science.

Frank Heslop: Theory and application of mixture of single index models
University of Pretoria

Since the origin of the mixture of regression models (Goldfield and Quandt, 1973) there have been various attempts to address the flexibility of standard mixture regression models. In the process semi-parametric and non-parametric mixture of regression models have been developed. Due to the presence of kernel regression techniques used in the estimation of these models, many proposed techniques suffer from the "curse of dimensionality". This presentation considers a semiparametric mixture of regressions model that can be applied to cases with high dimensional predictors using a single index approach (Xiang and Yao, 2020). To estimate the parameters of the model, a backfitting algorithm and modified EM algorithm is suggested. Simulated examples and practical applications for these models are considered.

Lindo Magagula: Design, implementation and application of distribution-free monitoring schemes based on order statistics. Case: Closed-form expressions of the run-length distribution of the nonparametric double sampling precedence monitoring scheme
University of Pretoria

A major challenge in statistical process monitoring (SPM) is to find exact and closed-form expressions (CFEs) for the run-length properties such as the average run-length (ARL), the standard deviation of the run-length (SDRL), and the percentiles of the run-length (PRL) of nonparametric monitoring schemes. Most of the properties of these schemes are usually evaluated using simulation techniques. Although simulation techniques are useful when the expression for the run-length is complicated, their shortfall is that they require a high number of replications to reach reasonable accurate answers. Consequently, they take too much computational time compared to other methods such as the Markov chain method or integration techniques, and even with many replications, the results are always affected by simulation error. In this paper, closed-form expressions of the run-length properties for the nonparametric double sampling precedence monitoring scheme are derived and used to evaluate its ability to detect shifts in the location parameter. The computational times of the run-length properties for both the proposed and the simulation approach are compared under different scenarios. It is found that the proposed approach requires less computational time compared to the simulation approach.

Azam Kheyri: Biological network structure learning through graphical models
University of Pretoria

Understanding the relationships within biological networks is essential for understanding complex biological systems. This study introduces a novel approach for structure learning in graphical models. By incorporating sparsity-inducing penalties, our method promotes representation of the network structure, enabling the identification of key interactions and meaningful biological relationships.

Matthias Wagener: The perfect fit: interpretation, flexible modelling, and the existing generalisations of the normal distribution
University of Pretoria

Many generalised distributions exist for modelling data with vastly diverse characteristics. However, very few of these generalisations of the normal distribution have shape parameters with clear roles that determine, for instance, skewness and tail shape. In this chapter, we review existing skew mechanisms and their properties in detail. Using the knowledge acquired, we add a skewness parameter to the body-tail generalised normal distribution, which yields the flexible interpretable normal distribution (FIN) with parameters for location, scale, body-shape, skewness, and tail weight. Basic statistical properties of the FIN are provided, such as the PDF, CDF, moments, and likelihood equations. Additionally, the FIN PDF is extended to a multivariate setting using a student t-copula, yielding the multivariate FIN distribution MFIN. The MFIN is applied to stock returns data, where it outperforms the t-copula multivariate generalised hyperbolic, Azzalini skew-t, hyperbolic, and normal inverse Gaussian distributions.

Renate Thiede: Measuring homogeneity of spatial line patterns
University of Pretoria

A spatial line pattern is a collection of lines in geographic space, and can be used to model spatial phenomena such as road networks. Homogeneous spatial line patterns exhibit consistent characteristics such as length and density across geographic space, while the characteristics of inhomogeneous line patterns vary. Although there exists a variety of tests for the homogeneity of point patterns, no equivalent statistical tests currently exist to quantify the homogeneity of line patterns. This research shows that existing tests for the homogeneity of point patterns can be extended to define tests for homogeneity of line patterns. Line patterns are represented by point patterns, where each line is represented by its midpoint. We investigate the power of the proposed tests and show that the tests are successfully able to identify homogeneity and inhomogeneity of line patterns. Finally, we apply the tests to formal and informal road networks in Mamelodi, South Africa, and show that the tests are informative when applied to road networks. This paper is the first to investigate the homogeneity of spatial line patterns, and provides a computationally efficient and mathematically simple way to test for such homogeneity.

Workshops

Workshop 1: Introduction to Copulas

Prof Leonard Santana, Prof James Allison, Prof Jaco Visagie

North West University

The original concept of a copula was first introduced by Sklar in the late 1950s but, in recent years, has become a popular topic in practical applications, notably finance, which has led to a renewal in the interest in theoretical research into this topic. In this workshop, we explore basic concepts relating to modelling dependencies between random variables using copulas. Specifically, the scope of the workshop includes introductory concepts, different classes of copulas and examples, random number generation, measures of dependency, and estimation. Practical examples from finance will be included to illustrate these concepts.

Workshop 2: Likelihood ratio test (LRT) statistics used in Multivariate Analysis

- **as products of Beta distributed random variables**
- **development of near-exact distributions**

Prof Carlos Coelho

We will start with very simple examples involving Beta distributed r.v.'s (random variables), namely going through a very simple example that will show the great usefulness of m.g.f.'s (moment generating functions) and c.f.'s (characteristic functions) as tools to handle the distributions of LRT statistics used in Multivariate Analysis. Although c.f.'s are commonly seen as quite hard to tackle, they are actually quite easy to use in this setting, and we want to show this. The illustration that the distribution of most of the LRT statistics used in Multivariate Analysis have the same distribution as that of the product of a number of independent Beta distributed r.v.'s and how we can sometimes obtain this distribution in a finite closed form

and how we can in the other situations, where this is not possible to be done, obtain very sharp approximations to their distributions, will be done using
– a rather well-known statistic, that is the Wilks Lambda statistic to test the independence of two sets of variables, and
– a not so well-known statistic, that is the statistic used to test the equality of several mean vectors, when the covariance matrices are assumed to be circular.
The techniques illustrated can be used on a very wide range of statistics used in Multivariate Analysis.

Workshop 3: Survival analysis and machine learning

Dr Albert Whata, Sol Plaatje University, Dr Justine Nasejje, University of the Witwatersrand, and Dr Najmeh Nakhaeirad, University of Pretoria

Survival analysis techniques, such as the Kaplan-Meier estimator and Cox proportional hazards model, are commonly used to analyze censored data and estimate survival probabilities or hazard rates. These methods provide valuable insights into the relationships between covariates and survival outcomes. However, they often make assumptions about the underlying data distribution and linear relationships, which may limit their flexibility and predictive accuracy. The combination of survival analysis and machine learning allows for more accurate prediction of survival outcomes, identification of important risk factors, and better understanding of the underlying mechanisms influencing the event of interest. It enables researchers to leverage the strengths of both fields and extract meaningful insights from complex and heterogeneous datasets. This interdisciplinary approach has applications in various domains, including healthcare, finance, biology, and social sciences, where understanding and predicting time-to-event outcomes are important.