UNIVERSITY OF CAPE TOWN
DEPARTMENT OF STATISTICAL SCIENCES
STATISTICS HONOURS MODULE INFORMATION 2018

THEORY OF STATISTICS A

This module is intended for students with undergraduate training in mathematical statistics and comprises of the following three sections (12 lectures each).

1. Stochastic Processes (Mr Melusi Mavuso)

   This module introduces Levy processes. The following topics will be covered:

   Content:

   - Definition and examples of Levy processes
   - Infinitely divisible distributions
   - The Levy-Khintchine formula
   - The Levy-Ito decomposition
   - Simulation of Levy processes
   - Estimation of Levy processes
   - Exponential Levy models

2. Likelihood Theory (Dr Birgit Erni) This module covers likelihood theory. The likelihood function plays an important role in Bayesian and in frequentist statistics. However, here we will treat likelihood as an approach to inference (point and interval estimates, model selection and comparison) in its own right.

   Content:
   Chapter 1 covers the likelihood function, maximum likelihood estimation, invariance property of the likelihood, relative likelihood, and likelihood intervals. Likelihood under reparameterization and transformations of response are also considered.

   Chapter 2 covers asymptotic properties of the likelihood function and maximum likelihood estimates, and includes sections on score, information, deviance, Wilks's likelihood ratio, tests for quadraticity of the likelihood, and Wald tests and intervals.

Chapter 3 considers likelihoods with two or more parameters, profile likelihood, marginal and conditional likelihoods, compatible inference, link to AIC, and introduces predictive likelihoods, and likelihood for random parameters.

In practice, most likelihood methods require substantial computation. Therefore the module will have a strong practical component, and will require the use of R for tutorial questions and the assignment. A tutorial for using R for likelihood is included in the course material.

3. Computational Bayesian Inference (Mr Allan Clark)

   The rise in computing power has allowed for the introduction of several computationally intensive statistical procedures. This has helped to revolutionise the field of statistics over the past few decades. This module will cover a number of standard computational Bayesian techniques used in statistical analysis. This will include Acceptance-Rejection sampling, Gibbs sampling and Markov Chain Monte Carlo methods. The course is intended to be practical in nature with a large practical component.

   Assessment:

Class assignments (tutorial / coding)	40%
Final examination	60%

THEORY OF STATISTICS B

This module is intended for students with undergraduate training in applied statistics and comprises of the following four sections (12 lectures each).

1. Generalised Linear Models (Assoc Prof Francesca Little)

   An introduction to GLMs and the underlying theoretical concepts, including:

   - The exponential family of distributions
   - Estimation
   - Inference

2. Introduction to Stochastic Processes (Mr Dominique Katshunga)

   (a) Chains and Markov Chains.
   - Transition possibility matrix.
   - Time homogeneity assumption.

- Classification of states.
- Probability models for transitions.

(b) Stationarity and Reversibility.

3. Likelihood Theory (Dr Birgit Erni)
   See the detail for the likelihood theory module above under theory of statistics (A).

4. Introduction to Bayesian Inference (Mr Allan Clark)

   This module gives a basic overview of the philosophy and methods of Bayesian analysis. Key Bayesian concepts are introduced and it is shown how they may be used to conduct statistical inference under a Bayesian framework. The Bayesian analysis is contrasted with the traditional frequentist approach.
   Content:

   - Philosophy
   - Bayesian concepts
   - Hypothesis Testing: the Bayesian approach
   - Bayesian Estimation
   - The Acceptance/Rejection algorithm

   Assessment:

Class assignments (tutorial / coding)	40%
Final examination	60%

OPERATIONS RESEARCH A

This module is intended for students who have not completed STA3036S or equivalent.

Lecturers: Dr Sheetal Silal, Emeritus Prof Theo Stewart

Content: Definitions of operations research, linear programming (formulation, solving in Excel, sensitivity, network and project management models), queuing theory, simulation.

Assessment: Two class assignments (1/3) and one exam (2/3).

OPERATIONS RESEARCH B

This module is intended for students who have completed STA3036S or Operations Research A.

Lecturers: Dr Sheetal Silal, Dr Juwa Nyirenda

Content: Linear programming (multi-objective and goal programming, duality, data envelopment analysis, integer programming), deterministic and stochastic dynamic programming, simulation.

Assessment: Two class assignments (1/3) and one exam (2/3).

MATRIX METHODS

Lecturer: Mr Stefan Britz

Content: Basic matrix operations, determinants, inverses, rank, generalised inverses and linear equations, eigenvalues and eigenvectors

---

STATISTICAL COMPUTING

- The R programming language

  Lecturer: Mr Allan Clark – 24 lectures

  Content:
  The R programming language will be introduced in the course. We start with the foundations of the course and cover topics such as the handling of matrices, vectors, loops, conditional statements, importing data and general programming logic. We then progress to topics such as high performance computing, "the bootstrap", Monte Carlo simulations and optimisation procedures in R. The aim of the course is to walk away with a general understanding of how to code in R as well as be able to logically attempt a programming problem. e.g. in previous years, students were required to code up the workings of a basic snakes and ladders game.

- Visual Basic for Applications

  Lecturer: Mr Neil Watson – 12 lectures

  Content:
  This module provides a broad overview of two of the fundamental uses of the Visual Basic for Applications (VBA) language:

a) Automating Excel spreadsheet calculations and tasks, and

b) Building Userforms (i.e. Excel applications) to enable users to perform a variety of tasks in Excel in a fraction of the time that it would take them to do manually.

BIOSTATISTICS

This course is an introduction to the analysis of data from medical research. Students will be prepared for entry positions as biostatisticians and for postgraduate research degrees in medical statistics. The course could be called "Applied generalized mixed effect models" and though the applications taught in this course will refer to medical data, the methods have much wider application. The emphasis of the course will be on the application and interpretation of the different statistical methods.

Lecturers: Dr Freedom Gumedze and Mr William Msemburi

Content:

- Design of Studies
- Introduction to programming in STATA/ R for biostatistics
- Measures of disease frequency and effect
- GLMs, with specific emphasis on Logistic Regression
- Analysis of Longitudinal Data, including GEE models and mixed effect models
- Survival Analysis

Assessment:

Two class assignments	50%
Final examination	50%

A minimum of 40% is required for the class assignment component and for the exam in order to pass the module.

ECONOMETRICS

This is a course in econometrics with an applied orientation and makes intensive utilisation of econometric computer software (Eviews). Forms part of the triad of honours modules (together with Portfolio Theory and Time Series Analysis) intended for those wanting to pursue a financial career.

Lecturer: Prof Graham Barr

Content: Causality. Passive observation versus controlled experiments. What we can do (and can't do) with statistical analysis. The role of regression and the theory of the regression model and its attendant assumptions. Heteroscedasticity and autocorrelation. Multicollinearity and mis-specification. Time series regression. Econometric modelling. Single-equation and simultaneous equation modelling. Applications to the SA financial and economic sector.

Assessment:

One class assignment	30%
Final examination	70%

PORTFOLIO THEORY

This module is intended for students with undergraduate training in mathematical statistics and does not require prior finance knowledge.

Lecturer: Assoc Prof Tim Gebbie

The course aims to familiarise the student with the foundations of portfolio theory as related to the quadratic optimal asset allocation problem and its use near to financial equilibrium. The course will be data-driven and consider daily sampled equity, bond and money market returns. The student will be introduced to the pre-processing of data using R within the RStudio IDE and will be expected to become proficient with R markdown for assignments and the exam.

The student will be introduced to a history of stock markets, the gambling problem, the Kelly criterion, and the theory of lotteries that leads to the Von Neumann-Morgenstein utility theory. This will then be used to introduce the inter-temporal portfolio selection problem as presented by Merton using the theory of stochastic dynamic control. The mean-variance theory and the related equilibrium asset pricing models will then be demonstrated. This theory is used to explore and derive both CAPM and then the less

restrictive APT from the perspective of neo-classical finance and its equilibrium relationship to mean-variance portfolio theory. The mean-variance theory is then extended to the Bayesian setting using the Black-Litterman model. This is applied to practical portfolio selection problems.

The implicit equilibrium nature of the framework is a key element of the course. An introduction to the realities and statistics of historical back-testing and strategy simulation will be included in the course. The course data will be used to explore various performance, attribution and risk measures. The course considers benchmark relative performance, the tactical and strategic asset allocation problems, in the context of the Roll criticism.

The course is a pre-requisite for courses on data-analysis for algorithmic and quantitative trading.

Content:

- South African markets and decision making under uncertainty
- Data Wrangling in R for Portfolio Selection
- Single Period Mean-Variance Portfolio Selection
- Market Efficiency, Equilibrium and Asset Pricing
- Asset Allocation and Multiperiod Portfolio Selection
- Tactical Asset Allocation and Active Management
- Performance Measures
- Performance Attribution
- The Fundamental Law and Information
- Portfolio Management and Roll's Criticism

Assessment:

Class mark	30%
Examinations	70%

TIME SERIES ANALYSIS

This module may not be taken by students who have completed STA3041F.

Time series data differs from data used when undertaking regression analysis since the data is not collected at one point in time. The data is collected at different time points and thus requires specialised techniques to understand the underlying generating process governing it. In this course, we will specifically focus on the analysis of time series. The techniques learnt are particularly useful for the analysis of financial and economic data. We will cover topics such as model building, estimation, prediction/forecasting, volatility modelling (useful when undertaking risk management and options trading) and cointegration (the technique can be used to uncover long-term economic relationships that may be useful for traders (equity readers specifically) and asset managers). The course is not intended to be overly theoretical and practical applications will be emphasised. Eviews (an econometrics and time series package) will be used during the practical sessions and for assignments.

Lecturer: Mr Dominique Katshunga

Content:

- Stationarity, simple time series models, random walks
- Unit root tests
- The Box Jenkins methodology – building simple time series models
- Forecasting
- Volatility models – ARCH and GARCH models
- Bivariate and Multivariate Cointegration

Assessment:

One class assignment	30%
Final examination	70%

DECISION MODELLING

The aim of this module is to develop an understanding of human preferences and subjective judgements, for purposes of constructing models for decision support, within which we can apply quantitative tools studied elsewhere. The module thus provides a bridge

between technical statistical and operational research tools on the one hand, and interaction with our clients on the other hand (especially in the context of decision support). We study both how to represent human preferences in mathematical models, and what cognitive models and errors may be exhibited by clients when we elicit subjective estimates of underlying parameters.

A substantial portion of the course relates to multiple criteria decision analysis (MCDA), in which we specifically aim to capture preferences and value tradeoffs between conflicting objectives and goals (e.g., social, environmental and economic goals).

In the course, we seek to develop an understanding of the following processes:

1. How can we assist decision makers in defining and structuring decision problems in such a way that we are able to apply our quantitative analytical models to them?
2. How should we model the preferences and value judgements which are expressed by decision makers or their advisors, in order to generate maximum insight into which courses of action best satisfy these preferences?
3. How do people form and express subjective value judgements, for example regarding the relative importance of different goals or of the likelihood of certain events, which we may need to use in our models?

Decision Modeling is a multi-disciplinary field and in addressing the above questions, we touch on issues in psychology, economics, information systems, and operational research. The course consists of two main parts. In the first part, we look at how we can take a general decision problem and structure it in a way that is useful for later modeling. We introduce the three main schools of decision modelling, which offer different ways of modeling how people can and should make multicriteria decisions. We then focus on one of these schools, value function modelling. In the second part of the course, we extend the models of the first section into more advanced decision contexts: those involving uncertainty (where measurements cannot be made precisely), preferences over time, or groups of decision makers. As part of this section we also consider problems and biases that people commonly experience when they are asked to think about information or make a judgement.

Lecturers: Assoc Prof Leanne Scott and Dr Ian Durbach

Assessment:

Two class assignments	40%
Final examination	60%

In most years the assignments are based on the material in Part 1, with the exam consisting mostly of material from Part 2.

MULTIVARIATE STATISTICS

Multivariate analysis is broadly concerned with the analysis of data comprising observations on large numbers of variates for each of a sample of individuals. Such data are generated everywhere, industry, biology, engineering and so on, and the attendant methods of analysis have become a vital part of the statistician's tool-kit. The aim of the course is to introduce the theory and practice of multivariate analysis. The R programming language will be used in this course.

Lecturer: Assoc Prof Francesca Little and Mr Stefan Britz

Content:

- The Multivariate Normal Distribution (SB)
- Multivariate Maximum Likelihood Estimation (SB)
- Multivariate Inference (SB)
- MANOVA and Multivariate Regression Analysis (SB)
- Principal Component Analysis and Factor Analysis (FL)
- Canonical Correlation Analysis (FL)
- Discriminant Analysis (FL)

Assessment:

Two assignments	50%
Final examination	50%

ANALYTICS

This course will cover computationally-intensive statistical methods for analysing datasets of various sizes. The course will cover three broad sections: (1) Parallel and high performance computing in R, (2) Supervised Learning and (3) Unsupervised Learning. In the first section, students will learn how to use R to analyse large datasets on multiple computer processors, and UCT's own HPC cluster. The second section will expose students to machine learning techniques that are used to infer a regression or classification rule based on labelled training data. The last section will cover statistical methods for classifying observations into groups where the group memberships of the training data are not known in advance.

Lecturers: Dr Juwa Nyirenda, Dr Etienne Pienaar, Mr Stefan Britz

Content:

1. High Performance Computing
    - Parallel computing in R
    - Cluster computing
2. Supervised Learning
    - Regression and classification trees
    - Bagging and random forests
    - Boosting
    - Neural networks
3. Unsupervised Learning
    - Self-organising maps
    - Association rule mining
    - Cluster analysis

Assessment:

Assignments	40%
Final examination	60%

ANALYSIS OF HOUSEHOLD SURVEY DATA (ECO4027S)

Course description:
This course covers the basics of how survey data are collected and presents methods to compensate for design features of complex sample survey data. These techniques are then applied to poverty, inequality and mobility measurement using South African household surveys. We spend the last three weeks looking at firm survey data and what it can tell us about big picture debates in development economics.

Section 1: Analysis of complex sample surveys
Standard courses on statistical analysis assume that survey data are from a simple random sample of the target population. Little attention is given to characteristics often associated with survey data, including missing data, unequal probabilities of observation, stratified multistage sample designs, and measurement errors. Failure to take most of these properties of survey data into account can have an important impact on the results of all types of analysis, ranging from simple descriptive statistics to estimates of

parameters of multivariate models. The course will cover sampling design and weights, variance estimation for complex sample surveys and consequences of non-response and missing data and methods for dealing with missing data.

Section 2: Poverty and Inequality measurement and the analysis of household and firm survey data. Household surveys provide valuable information about household economic circumstances and socio-economic behaviour. The course investigates a range of policy relevant issues through the analysis of South African household surveys. In 2017 the topics will include: the measurement of poverty, inequality and social mobility; the evaluation of social policy with a particular focus on the state old age pension's impact on poverty, and understanding development through the lens of Ghanaian manufacturing firm survey data.

The course includes a computer practical component built around statistical analysis of household and firm survey data using Stata. The emphasis is on the application of a range of statistical analysis techniques to socio-economic analysis rather than on the theoretical underpinnings of such techniques.

Problem Sets are to be handed in to Yasmina Charles on the 4th floor of the School of Economics building by 4pm on the day they are due. You must sign a plagiarism declaration for each problem set - **plagiarism is prohibited**. Cheating is also prohibited and cheating in a test or problem set will result in the student(s) getting zero and being referred to the relevant bodies in the Economics Department.

Lectures and computer labs: Lectures will be held on Mondays from 14h00 15h45 in Lecture Theatre 1 in the School of Economics Building. Computer pracs will be held on Wednesdays from 14h00 15h45 in the Teaching Lab in the School of Economics Building.

Readings: The book "Sampling, Design and Analysis" by Sharon Lohr will be useful for part 1 but is not required. There are two copies in the main library and one in health sciences. A reader for Inequality, Poverty and Social Mobility will be given out in the second half of the course.

Pre-requisites: students are expected to be familiar with basic econometric/statistical methods through multivariate linear regression. A basic knowledge of Stata is required.

Lecturers: Dr. Andrew Kerr (DataFirst), Dr Muna Shifa (SALDRU)