ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS

# DATA-SCIENCE STREAM

## *Statistical problem-solving*

*Skills course*

Data is at the heart of the modern world; survey data, data from designed experiments, or simply data to be mined and understood. Across Africa there is a great need for people who can solve problems using data. Research and development activities in many fields of application create data but in doing so also create a great need for statistical problem solvers.

Statistics is the discipline which develops the tools and methods to extract information from data. A solid understanding of the basic statistical concepts is sufficient for many data problems. Others require “high-powered” statistics and data science tools. This course builds student experience working with data from many fields of applications.

Many problems in statistics concern data analysis. However the subject is broader than this and the course includes activities on a range of tasks that involve statisticians. Data is fundamental! Studies must be designed, and data collected and organised. Results have to be presented and data reported and archived as well as being analysed.

Examples of consultancy work are presented and linked to key concepts in statistics including cases when there is no ‘textbook’ answer. Students will be exposed to various datasets along with real and realistic problems which they will be challenged to address.

*R Statistical Software.* Students are introduced to R through the R-Instat software. This has largely been developed within Africa. It adds a menu-driven front-end to R, designed so students can concentrate on the data analysis, while viewing and being exposed to R commands at the same time. One important feature of R-Instat that is needed in many data analysis problems, is the ease with which data at multiple levels can be processed. R-Instat keeps a log file of the R-commands, so analyses can, where needed, continue using R commands directly.

**Topics**

- Data Flow
- Examples of consultancy work
- Experimental design game leading to Options by Context analysis
- R-Instat statistical software
- CAST electronic textbook
- Organising data
- Managing multilevel data
- Exploratory and Presentation Graphs
- Good tables
- Objectives and significance tests. They both matter!

- Case Studies for group work (one from)
  - UNICEF Survey data
  - PICSA (Participatory Integrated Climate services for Agriculture).
  - Monitoring and Evaluation Study in Agricultural Climatology
  - Farmer experimentation in women's fields in Niger
  - Climate change – examining African rainfall and temperature data.
  - Sunshine data, comparing satellite and station data
  - And others

**Textbook**

The main textbook is the CAST series of electronic textbooks. This also includes a set of interactive exercises that can be used as the basis for a mastery testing system.

## *Statistical inference*

*Review course*

This course will address probability as a foundation for statistics, and then the main concepts of statistics with regard to data collection, estimation of parameters and statistical modelling. For discovery statistics, one requires an important question about the world, in order to choose a sensible survey or experimental design so that there is good quality data. The use of initial data analysis in choosing random variables to model the data will be discussed. The likelihood principle for estimation of the parameters of the models will be introduced, focusing on statistical inference relevant to the initial questions. The R software will be used to apply ideas to the core data sets for e.g. Binomial, Poisson, Normal, and standard linear regression.

### Topics

- *Probability.* Basic laws of probability; random variables (discrete and continuous); moments; change-of-variable; Bayes theorem; distributions.
- *Statistics.* Descriptive statistics and choice of probability distribution; likelihood theory, including:
  - definition of the likelihood function;
  - likelihood principle for inference;
  - graphical approach to likelihood inference: maximum likelihood estimates and confidence intervals;
  - properties of maximum likelihood estimators (including proof of asymptotic normality for one parameter)
  - log-likelihood ratios, asymptotic distribution (including proof for one parameter case)

**Text books**

James G, Witten D, Hastie T and Tibshirani R (2013) *An Introduction to Statistical Learning: with Applications in R.* http://www-bcf.usc.edu/~gareth/ISL/

An Introduction to Statistical Learning: with Applications in R
Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Second Edition.
https://web.stanford.edu/~hastie/ElemStatLearn/

*Other books, with different approaches to statistics.*

Crawley MJ (2014) *Statistics: An Introduction Using R.* Wiley.
Hogg RV, McKean JW, Craig AT (2005) *Introduction to Mathematical Statistics.* Pearson Prentice Hall.
Rice JA (1995) *Mathematical Statistics and Data Analysis.*

*Some useful basic introductory books if students want to revise concepts for probability and standard distributions.*
Crawshaw DJ, Chambers JS (2001) *Concise Course in A-level Statistics with worked examples*. 4th edition [Basic, A-level, pages 134-259 good worked probability examples]

Hodge SE, Seed ML. (1972) *Statistics and Probability* [Basic, A-level, recommended, uses probability trees]

## *Statistical modelling*

*Review course*

This course will focus on generalized linear regression, its strengths and limitations, as well as introducing basics stochastic processes to model correlated data. Making extensive use of real data examples, the course will emphasize the role statistical models in addressing scientific questions and how these are translated into relevant statistical questions. The student will learn to distinguish between problems of parameter estimation, hypothesis testing and prediction. The course should also introduce some examples of non-independent data while highlighting the generality of the likelihood principle introduced in the “Statistical Inference” course.

### Topics

- *Linear regression.* Assumptions; Simple and multiple linear regression; Interpretation; Least-squares estimation and the Guassian-Markov theorem.
- *Generalized linear models.* Assumptions; Estimation; Examples, including at least Binomial and Poisson distributions.
- *Stochastic processes.* Introducing models for modelling correlation, e.g. fitting a bivariate Normal model, a Markov chain model to a binary sequence, a first-order autoregressive model to a time series.

**Text books**

Weisberg S (2005) *Applied linear regression.* Wiley
Dobson AJ (2002) *Introduction to Generalized Linear Models.* Chapman & Hall.
Faraway JJ (2006) *Extending the Linear Model with R.* Chapman & Hall.

## *Geospatial methods for public health*

*Review course*

The course provides an introduction to geostatistical models and their application to public health problems. By building on the previous courses of the data-science stream, the students will be introduced to approaches for modelling temporally and spatially correlated data through examples drawn from tropical disease epidemiology. The course will focus on the use of geostatistical models for disease prevalence mapping and how these can be used to effectively convey uncertainty to better inform policy decisions. The PrevMap package will be used to illustrate examples and its main functionalities for disease mapping in the R software environment

### Topics

- *Correlation in one dimension.* Time-series data; Harmonic regression; ARIMA models.
- *Correlation in two dimensions.* The variogram; Linear geostatistical models; Spatial prediction.
- *Disease mapping.* Binomial geostatistical models for prevalence data; Monte Carlo maximum likelihood; Exceedance probabilities.

**Text books**

Diggle PJ, Ribeiro PJ (2007) *Model-based geostatistics.* Springer Series in Statistics.
Giorgi E, Diggle PJ (2017) *PrevMap: An R package for prevalence mapping.* Journal of statistical software. 78:1-29