

Data Analysis Bootcamp Using Open-Access Program R (Intermediate Level)

About the course:

R is broad and powerful, with many analytic and graphic functions available (more than 50,000). With guidelines and instructions, you can navigate the tremendous resources available in R thereby accomplishing your work with style, elegance and efficiency.

Many people and researchers despise statistics, mainly due to non-mathematical background. This makes understanding complex statistical equations very difficult. The advent of computer programs such as R and the like provides a unique opportunity to teach statistics at a conceptual level without getting too bogged down in equations. However, the downside of the computer is that it makes it easy to make complete fools of ourselves if we do not really understand what we are doing. Running analysis using a computer without any statistical knowledge can be totally misleading. Hence this course could be called *Unearthing the Statistician in You Using R*.

This course is the second of 3 levels, i.e. essential, intermediate, and advanced topics. This level builds upon the essential level to solidify the slow organic tutoring, beginning with the basic structures of the R codes and key statistical concepts leading to a discovery journey that will allow you to quickly access the power of this great open-source endeavor. This course attempts to strike a good balance between theory and practice by using the computer as a tool for learning statistical concepts with the hope that you will gain a better understanding of both theory and practice. A **Certificate of Completion** will be issued upon the successful completion of the course.

Course Content:

Level	Sessions	Modules
Intermediate	Session 1	1 Basic statistics 1.1 Descriptive statistics - Descriptive statistics by group - Visualizing results 1.2 Frequency and contingency tables - Generating frequency tables - Tests of independence - Measures of association - Visualizing results - Converting tables into flat files 1.3 t-tests - Independent T-test - Dependent t-test - When there are more than two groups 1.4 Nonparametric tests of group differences - Comparing two groups - Comparing more than two groups 1.5 Visualizing group differences
	Session 2	2. Regression

		2.1. What will this session include? 2.2. An introduction to regression 2.2.1. Some important information about straight lines 2.2.2. The method of least squares 2.2.3. Assessing the goodness of fit: sums of squares, R and R^2 2.2.4. Assessing individual predictors 2.3. Packages used in this session 2.4. General procedure for regression in R 2.5. Interpreting a simple regression 2.5.1. Overall fit of the object model 2.5.2. Model parameters 2.5.3. Using the model 2.6. Multiple regression: the basics 2.6.1. An example of a multiple regression model 2.6.2. Sums of squares, R and R^2 2.6.3. Parsimony-adjusted measures of fit 2.6.4. Methods of regression 2.7. How accurate is my regression model? 2.7.1. Assessing the regression model I: diagnostics 2.7.2. Assessing the regression model II: generalization 2.8. How to do multiple regression using R Commander and R 2.8.1. Some things to think about before the analysis 2.8.2. Multiple regression: running the basic model 2.8.3. Interpreting the basic multiple regression 2.8.4. Comparing models 2.9. Testing the accuracy of your regression model 2.9.1. Diagnostic tests 2.9.2. Outliers and influential cases 2.9.3. Assessing the assumption of independence 2.9.4. Assessing the assumption of no multicollinearity 2.9.5. Checking assumptions about residuals 2.9.6. What if an assumption is violated? 2.10. Robust regression: bootstrapping 2.11. How to report multiple regression 2.12. Categorical predictors and multiple regression 2.12.1. Dummy coding 2.12.2. Regression with dummy variables Interesting real research
	Session 3	3 Logistic regression 3.1. What will this session include? 3.2. Background to logistic regression 3.3. What are the principles behind logistic regression?

		3.3.1. Assessing the model: the log-likelihood statistic 3.3.2. Assessing the model: the deviance statistic 3.3.3. Assessing the model: R and R² 3.3.4. Assessing the model: information criteria 3.3.5. Assessing the contribution of predictors: the z-statistic 3.3.6. The odds ratio 3.3.7. Methods of logistic regression 3.4. Assumptions and things that can go wrong 3.4.1. Assumptions 3.4.2. Incomplete information from the predictors 3.4.3. Complete separation 3.5. Packages used in this session 3.6. Binary logistic regression: an example 3.6.1. Preparing the data 3.6.2. The main logistic regression analysis 3.6.3. Basic logistic regression analysis using R 3.6.4. Interpreting a basic logistic regression 3.6.5. Model 1: Intervention only 3.6.6. Model 2: Intervention and Duration as predictors 3.6.7. Casewise diagnostics in logistic regression 3.6.8. Calculating the effect size 3.7. How to report logistic regression 3.8. Testing assumptions: another example 3.8.1. Testing for multicollinearity 3.8.2. Testing for linearity of the logit 3.9. Predicting several categories: multinomial logistic regression 3.9.1. Running multinomial logistic regression in R 3.9.2. Interpreting the multinomial logistic regression output 3.9.3. Reporting the results
	Session 4	4 Comparing several means: ANOVA (GLM 1) 4.1. What will this session include? 4.2. The theory behind ANOVA 4.2.1 Inflated error rates 4.2.2. Interpreting F 4.2.3. ANOVA as regression 4.2.4. Logic of the F-ratio 4.2.5. Total sum of squares (SST) 4.2.6. Model sum of squares (SSM) 4.2.7. Residual sum of squares (SSR) 4.2.8. Mean squares 4.2.9. The F-ratio 4.3. Assumptions of ANOVA 4.3.1. Homogeneity of variance

		4.3.2. Is ANOVA robust? 4.4. Planned contrasts 4.4.1. Choosing which contrasts to do 4.4.2. Defining contrasts using weights 4.4.3. Non-orthogonal comparisons 4.4.4. Standard contrasts 4.4.5. Polynomial contrasts: trend analysis 4.5. Post hoc procedures 4.5.1. Post hoc procedures and Type I (a) and Type II error rates 4.5.2. Post hoc procedures and violations of test assumptions 4.5.3. Summary of post hoc procedures 4.6. One-way ANOVA using R 4.6.1. Packages for one-way ANOVA in R 4.6.2. General procedure for one-way ANOVA 4.6.3. Entering data 4.6.4. Exploring the data 4.6.5. The main analysis 4.6.6. Planned contrasts using R 4.6.7. Post hoc tests using R 4.7. Calculating the effect size 4.8. Reporting results from one-way independent ANOVA Interesting real research
	Session 5	5 Analysis of covariance, ANCOVA (GLM 2) 5.1. What will this session tell me? 5.2. What is ANCOVA? 5.3. Assumptions and issues in ANCOVA 5.3.1. Independence of the covariate and treatment effect 5.3.2. Homogeneity of regression slopes 5.4. ANCOVA using R 5.4.1. Packages for ANCOVA in R 5.4.2. General procedure for ANCOVA 5.4.3. Entering data 5.4.4. Exploring the data 5.4.5. Are the predictor variables and covariates independent? 5.4.6. Fitting an ANCOVA model 5.4.7. Interpreting the main ANCOVA model 5.4.8. Planned contrasts in ANCOVA 5.4.9. Interpreting the covariate 5.4.10. Post hoc tests in ANCOVA 5.4.11. Plots in ANCOVA 5.4.12. Some final remarks 5.4.13. Testing for homogeneity of regression slopes 5.5. Robust ANCOVA 5.6. Calculating the effect size 5.7. Reporting results Interesting real research

	Session 6	6 Factorial ANOVA (GLM 3) 6.1. What will this session include? 6.2. Theory of factorial ANOVA (independent design) 6.2.1. Factorial designs 6.3. Factorial ANOVA as regression 6.3.1. An example with two independent variables 6.3.2. Extending the regression model 6.4. Two-way ANOVA: behind the scenes 6.4.1. Total sums of squares (SST) 6.4.2. The model sum of squares (SSM) 6.4.3. The residual sum of squares (SSR) 6.4.4. The F-ratios 6.5. Factorial ANOVA using R 6.5.1. Packages for factorial ANOVA in R 6.5.2. General procedure for factorial ANOVA 6.5.3. Entering the data 6.5.4. Exploring the data 6.5.5. Choosing contrasts 6.5.6. Fitting a factorial ANOVA model 6.5.7. Interpreting factorial ANOVA 6.5.8. Interpreting contrasts 6.5.9. Simple effects analysis 6.5.10. Post hoc analysis 6.5.12. Overall conclusions 6.5.13. Plots in factorial ANOVA 6.6. Interpreting interaction graphs 6.7. Robust factorial ANOVA 6.8. Calculating effect sizes 6.9. Reporting the results of two-way ANOVA
	Session 7	7. Sample size and power 7.1 Proportion 7.1.1 One sample 7.1.2 Two proportions (equal n) 7.1.3 Two proportions (unequal n) 7.2 Balanced one-way ANOVA 7.3 Chi-square test 7.4 General linear model 7.5 Correlation 7.6 T-tests (one sample, 2 sample, paired) 7.7 T-test (two samples with unequal n)
	Project	Practical research project (Data and description of project to be provided) - Presentation - Report

The Tutors:

About the Lead Tutor



Patrick Njage is a Statistician working as an Associate Professor in Genomic Epidemiology at the National Food Institute, Technical University of Denmark. His passion and ability to teach both common and advanced statistical topics with great simplicity in the open-access language R stems from his background in two worlds. He has background training in both natural sciences (BSc, MSc, PhD (Microbiology and Biotechnology) as well as Statistics (Master of Statistics, University of Hasselt, Belgium). He is experienced in the analysis of complex, unstructured quantitative data. He has taught statistical analysis in R environment to students in funded fellowships from 5 African countries and workshops as well as lectures to participants from different countries in Europe. He has also supervised and co-supervised numerous under, postgraduate and Postdoctoral projects.

Team of Dedicated Tutors:



Samuel Mwendwa
PhD Fellow, Aarhus University



Monica Masake
Lecturer and PhD Fellow, Machakos University



Barbara Wangari
BSc. Mathematics and Computer Science (Data Analyst)



Stephen Ndaiga (Research scientist, statistician and trainer- KALRO).



Leah Munyendo
PhD, University of Hohenheim, Germany



Tamana Clare Nelima
Food Safety and Quality (MSc.), Researcher, University of Nairobi



Joshua Owade
Ass. Professor, Oregon State University, USA



James Orwa
PhD, Stats, Ghent; Mstat, Hasselt University
Ass. Professor, Agha Khan University

What to expect:

R is powerful, versatile, and easy to use statistical software. It is the first choice for many data analysts working in both companies and academia. The course will help you master the modeling and data visualization of R in a short time, as a first step to become a skilled data analyst.

By the end of the course, you will have the most important R programming skills to handle data visualization and inferential analysis of various levels of complexity with enhanced confidence and reproducibility. You will also be able to identify appropriate analysis for each type of research problem you may encounter. Additional materials to help you understand the statistical software will be distributed during the course.

This course is designed to **strike a good balance between theory and practice** by using the computer as a tool for learning statistical concepts with the hope that you will gain a better understanding of both statistical theory and practice. This will allow you to quickly access the power of this great open-source endeavor.

Requirements

It is recommended that you have prior exposure to R or completed the course “***Essential Level: Data Analysis Bootcamp Using Open-Access Program R***”. Please prepare your personal data for research application of the concepts learnt. You are welcome to bring own data analysis questions to the resource persons.

Course Format

The course will use interactive teaching techniques and project work in groups. We hope that each participant will bring with them a case study on research they have been involved. Example datasets will also be provided for each concept. In addition to learning and skills development, the course will provide a valuable opportunity for those involved in quantitative data analysis to meet and later join an R programming support community consisting of the course alumni. **The final course evaluation will involve a group project presentation and report applying concepts learnt during the course.** Previous project presentations have been a very interesting and exciting pinnacle of the course where participants reveal the metamorphosis during the course to high mastery of R and data analysis.

The course week will be structured as follows:

- **30-60 minutes on Tuesdays:** Discussion on the statistical concepts and practical session of the week. Prior to this, participants will be sent an introductory video to aid in identifying any areas requiring clarification.
- **2 hours hands-on-session on Thursdays:** Guided hands-on practical session. Assistants answer questions and answers. A short tour of the codes for each session will be sent via a short video (10-15 minutes).

Registration

The course fee for this course ranges between USD450 (KES 58,315.50) - USD500 (KES 64,795.00). However, to achieve the vision of enhancing access to world-class data analysis skills by the “less reached groups in Africa”, the course does not include a tuition fee and only a registration fee of (**KES 5,000**) required to support maintenance of the online study portal, award of certificates and other expenses including honoraria for the volunteer instructors and moderators.

Registration fee payment

Registration will be via KEMRI Graduate School website available here: <https://kgs.ac.ke/short-course-programmes/> by clicking on the “**Data Analysis Bootcamp Using Open-Access Program R**”.

Registration deadline is 6th November 2026. **Payment deadline** is 8th November 2026. The course will run for 7 weeks from 10th November 2026; Tuesdays (7:30–9:30 PM E.A.T) and Thursdays (7:30–9:30 PM E.A.T).

Payment details

- Participants to log in with their existing eCitizen credentials to: https://kemri.ecitizen.go.ke/auth/login?redirect_to=%2Fservices%2F4%2Fapply
- If there is no existing eCitizen account, participant can create one.
- Fill in the online application form.
- In the list of courses, select “**Data Analysis Bootcamp Using Open-Access Program R**”
- Submit your application
- Proceed to make payment for your application.

Pre-course Zoom orientation (Q&A session)

Prior to the registration deadline, we have organized a pre-course Zoom orientation and Q&A session on Thursday 5th November 2026 at 19:30 EAT. You are welcome to attend this session via the Zoom link below. A Certificate of Completion will be provided upon successful completion of the course.

Zoom meeting link for pre-course Zoom orientation and Q&A session:

<https://sl1nk.com/25e8p2y>

Meeting ID: 696 4619 4764

Passcode: 925253

A Certificate of Completion will be provided upon successful completion of the course.