18th RUFORUM ANNUAL GENERAL MEETING 2022

Theme: Strengthening Africa's Agri-food Systems in the Post COVID-19 Era – Opportunities and Challenges

Scientific Data Management for Post-Graduate Students Using R Programming Language for Researchers and Post-Graduate Students

Strengthening Research skills through Capacity Building Training

Date: 10th- 16th December, 2022 CAT

Venue: Lecture Room 1

Registration Link: https://bit.ly/3BeMMIB

Contact Persons:
Dr. Susan Balaba Tumwebaze; susantumwebaze@gmail.com
Dr. Thomas Lapaka Odong; Thomas.l.odong@gmail.com
Dr. Hellen Namawejje; (hnamawejje@gmail.com)

CO-ORGANISERS:

Table of content

1.0 Background

Understanding of scientific data management and analysis enhances the capacity of postgraduate students and researchers to meaningfully engage in conducting quality research by developing appropriate research proposals, design of studies, collection, and analysis of data for quality reporting. Researchers, PhD and MSc students are heavily involved in large scale experiments or surveys that sometimes lead to complex designs and to subsequent messy data. Figuring out how to handle data resulting from such experiments/surveys takes time, and getting appropriate assistance is difficult. The Researcher and students are also constrained on how to effectively analyze data using appropriate statistical software, interpret the results and communicate well to the target audience. In recognition of these shortcomings, this course is structured to encompass broad biometrical needs that will equip the Researcher and students with skills required in conducting their research efficiently and effectively as well as act as a refresher course to lecturers. The content incorporated in this course is drawn from broader topics ranging from planning of experiments/surveys, designing, and implementing experiments, conducting data analysis. The Researchers and students will also be exposed to R programming language for data management, analysis, and reporting.

Given that most commercial statistical software are expensive, R a free statistical programming language has become a very powerful statistical tool among researchers worldwide. Most universities worldwide are moving away from commercial software such as SAS, STATA, SPSS to free open-source software especially R and Python. According to TIOBE index, R is popularly ranked as 8th among scholarly users throughout the world (Nayeemuddin, 2019). R has numerals number of advantages that support anyone who is interested in data analysis and any user can quickly learn R whether a data scientist or not.

R-programming has an effective, coherent and integrated collection of tools for data analysis, provides graphical facilities for data analysis and display, widely used for statistical computing and design especially in big data and data analytics, has field-specific advantages such as great data visualization features among other benefits (Imarticus, 2019). An R user can modify the

different functions in R and make their own packages. In addition, R is issued under the General Public License (GNU) and there are no restrictions on its usage.

R plays a paramount role in various fields, for example in data science- a user can run a code without any compiler, do many calculations done with vectors, applied in biology, genetics, statistics, among others. R is good for business, can be used in developing amazing web-apps, enjoys a vast community support from boot camps and R meetups and it is well maintained, and R updates are always available from CRAN. It is being used in almost every industry for example banking, manufacturing, health, insurance, agriculture to mention but a few (Data Flair, 2019).

RStudio is an integrated development environment (IDE) for R, also supports statistical computing and graphics. Here, a user can manipulate data and can be able to store used R commands for future references. The environment also provides an R markdown which supports conversion of work to different formats like, word, pdf, power point, HTML etc. This is a great deal to academicians since scholarly articles can be written directly in R markdown environment and later be published into a manuscript.

1.1 Target audience

The course is targeted for Researchers and postgraduate students but not limited to in any of the following fields: Plant Breeding, Crop and Horticultural Sciences, Animal Sciences, Agricultural Economics, Plant Protection, Food Science, Natural Resource Management, Aquaculture and Fisheries Sciences.

1.2 Aim

The ultimate aim of this training is to build research capacity of next generation African scientists specifically to achieve the following: understand the various biometrical components pertaining to design and analysis of experiments/surveys; apply various statistical techniques correctly at all stages of research and report the results effectively. The training will equip lecturers and

postgraduate students with the skills and knowledge in use of R programming language for data management, analysis, and presentation of results in a format that would ensure their wide dissemination as peer reviewed publications and policy formulation. It is expected that this training will give them the hands-on skills they need to improve the quality of their research publications. The training will also be an opportunity for postgraduate students to prepare their draft thesis.

1.3 Specific Objectives

a) Researchers and students will be able to learn how to download R, RStudio, R packages and install them on their computers
b) Researchers and students will learn how to import their data in different formats (csv, text, excel, SPSS etc. into R and export R data into other statistical programs
c) Researchers and students will learn how to manipulate their data before doing any data analysis using interactive commands since R supports matrix arithmetic and data structures such as vectors, arrays, data frames and lists.
d) Researchers and students will learn how use R Markdown where they can store the different R commands, write R scripts, which they can after convert into pdf, HTML, word document, and power point.
e) Researchers and students will learn different efficient techniques used in data visualization applied in R programming language that they can relate to their datasets as well as being able to produce publication-quality graphs they can use in writing up their manuscripts.
f) Researchers and students will be able to use R programming language to analyze data using all inferential statistics tools such as correlation & regression analysis; categorical data analysis techniques and generalized linear models.

1.4 Course Outcomes (Expectations)

At the end of the training participants would be able to:

(i) Explain the fundamental concepts behind experimental/survey designs and statistical data analyses.

(ii) Apply key statistical concepts such as correlation & regression analysis; categorical data analysis techniques and generalized linear models using R, etc.

(iii) Use R programming language to describe, analyze and model the state of a biological or agricultural system in both a quantitative and qualitative manner as well as other fields like economics, and health.

1.5 Delivery Method and Requirements

The module will have a balanced approach, including a brief theoretical underpinning, analytical tools, and practical application of the learning to solve real-world problems. Delivery will be blended learning, face to face learning, including interactive online lectures and practical designed to complement the lecture material. The approach will be participatory, with students expected to be active learners, and to commit themselves to intensive and critical self-study. Assignments will be designed to train and test critical thinking skills and application of what is learnt. Real life data sets brought by facilitators or drawn from students prior to the start of the course will be used throughout in examples and exercises. The mode of instruction is divided into two parts namely, limited theory/examples and computer exercises, delivery using online and face to face training. Each participant will be expected to have a laptop and a set of data. The daily programme will be divided into sections that will allow for an overview of the topics followed by computer-based practical and discussion on the statistical results. Basic principles followed by computer examples will be introduced first. Participants will analyse their data using techniques already introduced daily. Discussions on interpretation and presentation of the results will be held every day during the plenary sections. The participants will evaluate the modules daily and shortcomings addressed immediately. An overall course evaluation will be done at the end of the module.

1. 6 Course Pre-requisite

This course builds on the knowledge acquired by participants during their postgraduate and undergraduate studies. It assumes understanding of basic Biometrics applied to quantitative and qualitative data, and in addition, numeracy skills acquired overtime. The module provides a solid understanding of statistical techniques that relate to quantitative/qualitative aspects from application, and analytical perspective, thus balancing between theory and applied concepts.

1.7 Duration

The course will take seven(7) working days each day starting at 8:30am up to 5:00 pm, with breaks in between.

.

2.0 An overview training content

The following modules, which cover the whole range of applied biometrics, say, from basic concepts to computing and results presentation will provide a framework for the training material. The topics will concentrate on introducing statistical concepts in a non-theoretical way, with computer-based practical being used to illustrate the different concepts. These are highly practical topics intended to increase the participants' awareness of biometrical techniques for data management and analysis in their own specialist areas. The lecturers and students will be grouped according to their area of interest and given assignments with data related to their area followed by presentations at the end of the day which will be conducted online.

2.1 Module 1- Introduction to R programming language

Introduction to basics-theory, what is R, downloading R, RStudio and different R packages, arithmetic with R, variable assignment, basic data types in R, importing and exporting of datasets, setting a directory, saving datasets into R and excel sheets, creating data frames, sorting a data frame, what's a factor and why would they be used?, factor levels, summarizing a factor using categorical data, ordered factors, creating and naming lists, data manipulation in R, Data visualization, RMarkdown. This section will be a revision for those lecturers and students who have already used R before (these students will be requested to share their experiences with others and help their colleagues).

2.2 Module 2- Research design, data collection and Management

Introduction to planning and design of experiments and surveys, an overview of types of experiments (CRD, RCBD, Latin Square, Incomplete Block Design) and their applications; types and use of surveys; sampling techniques; sample size determination; and tools for study designs. Software and techniques for data entry and effective retrieval in R language.

2.3 Module 3- Exploratory data analysis

Introduction to Exploratory data analysis (What is exploratory data analysis? Importance of exploratory data analysis, Overview methods for exploratory data analysis); Summarizing quantitative variables-univariate (Five number summary, Description of population distribution characteristics (measures of center, Spread, Modality, Shapes) and Identification of outliers); Summarizing qualitative variables -univariate (Frequency table (Frequency vs Relative frequency), Pie chart, Bar graph); Relationship between two variables - qualitative and quantitative (Five number summary and Side-side boxplots for different categories); Relationships between two qualitative variables (Contingency table/Cross-tabulation/two-way table; Joint, marginal, and conditional distribution; Side-by-side bar plot or stacked bar plot); Relationships between two quantitative variables (Scatter plot/multiple scatter plots, Correlation coefficient, Variance-covariance matrix)

2.4 Module 4-Linear models

The introduction to the concept of statistical models; Analysis of variance (ANOVA) for the different experimental designs, Assumptions of Analysis of Variance; Post ANOVA analysis, Correlation and Regression analysis (simple linear regression, multiple regression), Model building steps, validation of assumptions linear models (ANOVA, Regression); Introduction to Mixed model (REML)

2.5 Module 5-Generalized Linear models

Transition from linear to generalized linear model; Contingency table and chi-square test for association; Logistic regression and loglinear models.

3.0 An overview of course outcomes

Each of the modules to be covered during the training will result to the following outcomes:

3.1 Outcomes for Module 1 on introduction to R programming language

- ❑ Download and Installation of R, R studio, and other R packages
- ❑ How to use R programming language to manage data, writing the code for the different statistical methods
- ❑ Import different datasets for example from excel, txt, into R and export R datasets into excel.
- ❑ Manipulate data in R using interactive commands since R supports matrix arithmetic and data structures such as vectors, arrays, data frames and lists.
- ❑ Do data visualization applied in R programming language as well as produce publication-quality graphs.
- ❑ Use R Markdown to store the different R commands, write R scripts and convert R Markdown into pdf, HTML, word document, and power point.
- ❑ Implement the different statistical methods with different R packages to analyze data

3.2 Outcomes for Module 2 on research design, data collection and management

- ❑ Identify primary and secondary source of data
- ❑ Apply the most appropriate data collection process
- ❑ Classify variables and scale of measurements appropriately
- ❑ Evaluate different techniques of data management
- ❑ Describe the underlying principles of experimental and survey design
- ❑ Distinguish between the different types of experimental designs (CRD, RCBD, Latin Square, Incomplete Block Design) and different treatment structures (Factorial, non-factorial)

3.3 Outcomes of module 3 on exploratory data analysis

- ❑ Choose appropriate exploratory data analysis for your study
- ❑ Perform exploratory data analysis
- ❑ Interpret the output from exploratory data analysis (including identify trends and outliers in your data)
- ❑ Select appropriate inferential data analysis tools based on exploratory data analysis

3.4 Outcomes of module 4 on linear models

- ❑ Distinguish between linear and non-linear models and know when to use each one
- ❑ Distinguish between random, fixed, and mixed effects models
- ❑ Choose appropriate statistical model/tool (ANOVA, Correlation, Regression, Mixed model) for data analysis
- ❑ Perform data analysis using the different statistical models/tools
- ❑ Build models using different scenarios of studies
- ❑ Validate final model of the analysis

3.5 Outcomes of module 5 on generalized linear models

- ❑ Distinguish between linear and generalized linear models and know when to use each one
- ❑ Choose appropriate method for analysis of categorical data (Chi-square, Logistic, Loglinear model etc.)
- ❑ Perform categorical data analysis using the different statistical models/tools (Chi-square, Logistic, Loglinear model etc.)
- ❑ Interpret the results of categorical data analysis

4.0 Annexes

4.1 Annex 1: Training Programme Week 1 (10-16/12/2022)

Time	Saturday (10/12/22)	Sunday (11/12/22)	Monday (12/12/22)	Tuesday (13/12/22)	Wednesday (14/12/22)	Thursday (15/12/22)	Friday (16/12/22)
8:30a.m -	Registration Opening Ceremony Objectives & overview of course	Overview of Exploratory data analysis (What is EDA? Importance of EDA, Types of EDA) Summarizing quantitative variables-univariate using R Summarizing qualitative variables -univariate using R **SB/HN**	An overview research process Principles of study design (Experiments and surveys) **TLO/SB** Overview of experimental designs (CRD, RCBD, Latin Square, Incomplete Block Design) and their applications. Use of R software for randomization in Experiments **TLO/SB**	Correlation and Regression analysis **SB/HN**	Analysis of variance (ANOVA) for the different experimental designs Assumptions of Analysis of Variance; Post ANOVA analysis, **TLO/HN**	Analysis of variance (ANOVA) for the different experimental designs Assumptions of Analysis of Variance; Post ANOVA analysis, **TLO/SB**	Contingency table and chi-square test of association Generalized linear models • Logistic model • Loglinear, etc. **HN/TLO**
1000-	*Health Break*	*Health Break*	*Health Break*	*Health Break*	*Health Break*	*Health Break*	*Health Break*

1030-	Installation of R Introduction to R programming language. RMarkdown **HN/SB/TLO**	Relationship between two variables - qualitative and quantitative Relationships between two qualitative variables Relationships between two quantitative variables **SB/HN**	Overview of sampling surveys (sampling techniques, sample size determination). Use of R for sampling **HN/SB**	Correlation and Regression analysis **SB/HN**	Analysis of variance (ANOVA) for the different experimental designs Assumptions of Analysis of Variance; Post ANOVA analysis, **TLO/SB**	Contingency table and chi-square test of association Generalized linear models • Logistic model • Loglinear, etc. **HN/TLO**	Overview of Analysis of Categorical data Transition from linear to generalized linear model **HN/TLO**
1200-	*Lunch Break*	*Lunch Break*	*Lunch Break*	*Lunch Break*	*Lunch Break*	*Lunch Break*	*Lunch Break*
2:00-	Introduction to R programming language. RMarkdown **HN/SB/TLO**	Relationship between two variables - qualitative and quantitative Relationships between two qualitative variables Relationships between two quantitative variables **SB/HN**	Overview of sampling surveys (sampling techniques, sample size determination). Use of R for sampling **HN/SB**	Correlation and Regression analysis **SB/HN**	Participants to participate in Vice Chancellors Forum	Participants to participate in the Official Opening of the AGM	participants to participate in the Closing ceremony of the AGM in the afternoon

CO-ORGANISERS:

4:00-	*Health Break*	*Health Break*	*Health Break*	*Health Break*	*Health Break*		
04:30-	Introduction to R programming language RMarkdown HN/SB/TLO	Computer practicals with R	Computer practicals with R	Computer practicals with R	Participants to participate in Vice Chancellors Forum	Participants to participate in the Official Opening of the AGM	participants to participate in the closing ceremony of the AGM in the afternoon

SB – Susan Balaba Tumwebaze; HN- Hellen Namawejje, TLO-Thomas L. Odong

4.2 Annex 2: Training needs assessment tool

Pre-Course Questionnaire for Statistical Data analysis Course Schedule for December, 2022

Section A : Personnel Information

Sex: ☐ Male ☐ Female

2. Name of degree being under taken (for students only) _____

SECTION B: Competence in Data Management, Study designs and Analysis

Data Management

3. Knowledge in design of data collection tools: ☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

4. Knowledge in data management techniques: ☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

5. Knowledge in design of spreadsheet for data entry: ☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

6. Knowledge in data checking: ☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

7. Knowledge in importation to statistical Software: ☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

Basic Statistics and Interpretation of Results

8. Knowledge in some basic descriptive statistics (*measures of central tendency and measures of dispersion*):

☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

9. Knowledge in some basic inferential statistics (*hypotheses testing, t-test, ANOVA, confidence intervals*):

☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

10. Knowledge in data interpretation and reporting: ☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

11. Knowledge in design of data collection tools: ☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

12. Knowledge in presentation of results: ☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

Statistical Software and Computer Skills

13. Which of the following statistical software you have access to and use/used?

Statistical software aware	Aware of	Have access to	Use or used
R			
SPSS			
SAS			
GENSTAT			
MINITAB			
Any other specify:			

14. What is your level of knowledge in R?

☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

15. What is your level of knowledge in other statistics software (e.g.Genstat, *SPSS, SAS, etc.*)?

☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

16 What is your level of knowledge in Spreadsheet/Microsoft Excel?

☐ None ☐ Slight ☐ Moderate ☐ Good ☐ Excellent

SPECIFIC TOPICS

17. Please tick in the appropriate cell in the grid below, your assessment of your needs and capabilities in the topics listed (*VD= very deficient, ND= Not deficient, MR= Major enhancement required, LR= Little enhancement required)*

Component	**Specific Topics**	**Level of competence on this topic**					**Level of enhancement required**				
		VD				**ND**	**MR**				**LR**
		1	2	3	4	5	1	2	3	4	5
Data entry and management	General survey procedures										
	Agricultural & market survey procedures										
	Large scale experimental designs										
	Techniques of data										

	checking										
	Data storage and retrieval										
	Data management strategy										
Data analysis and Interpretations	Exploratory data analysis										
	Analysis of variance (ANOVA)										
	Chi-square and t-test										
	Non-parametric methods for survey data										
	Use mixed models										
	Use of generalised linear models (e.g. Logistic regression, Log-Linear)										
	Regression analysis (simple and multiple)										
	Quality management – principles and applications in agricultural research and data										
	Interpretation of statistical results										

Please email the form to the Course Facilitators:

Dr. Thomas Lapaka Odong; thomas.l.odong@gmail.com
Dr. Susan Balaba Tumwebaze; susantumwebaze@gmail.com
Dr. Hellen Namwejje; hnamawejje@gmail.com