Scientific Data Management for Post-Graduate Students Using R Programming Language-Follow-up Training (4th-9th October 2021)

Strengthening Research skills in Eastern and Southern Africa

Concept Note

Resource Persons
Dr. Susan Balaba Tumwebaze (susantumwebaze@gmail.com)
Dr. Hellen Namawejje (hnamawejje@gmail.com)

CO-ORGANISERS:

Table of content

1.0 Background

The Follow-up Scientific data management (SDM) is drawn from the previous SDM that was conducted on 16th-21st August 2021. Based on the course objectives, content covered, evaluation and participant's expectations, of the just concluded SDM, it was concluded that the training was successful and met the participant's expectations. However, based on suggestions provided by participants on how to improve the course and facilitator's observation as well as past experience hence a follow up scientific data management training using R to be organized by RUFORUM was recommended thus the proposed follow-up training.

In this follow-up training, the basics for R programming language will not be repeated since there are videos that participants can download and review to enrich their knowledge and skills in R (see links).

Previous Training Links can be downloaded here

Day 1 Link: https://www.youtube.com/watch?v=6n6IsRhFb2A

Day 2 Link: https://www.youtube.com/watch?v=tnYqkxvf1P4

Day 3 Link: https://www.youtube.com/watch?v=Sl1HCJnYejU

Day 4 Link: https://www.youtube.com/watch?v=4KscSiZa8HQ

Day 5 Link: https://www.youtube.com/watch?v=hdRlBND_z4c&t=178s

Day 6 Link: https://www.youtube.com/watch?v=Sp7Ot6PpwjM

However, for every topic that was covered during the previous training, a recap of the R package required for analysis will be provided in the follow-up training.

Topics to be covered were suggested by participants from previous training, and what was not covered on the program and other relevant topics such as; Regression, experimental design and analysis using REML, Repeated measures, generalized linear models (logistic regression, log-linear models, ordinal and multinomial regression) and contingency tables will be covered. During this training, training hours will be 4 hours with breaks in between.

SDM training enhances the capacity of postgraduate students to meaningfully engage in conducting quality research by developing appropriate research proposals, design of studies, collection, and analysis of data for meaningful reporting. 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 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 postgraduate students with skills required in conducting their research efficiently and effectively. The content incorporated in this course is drawn from broader topics ranging from planning of experiments/surveys, designing, and implementing experiments, conducting data analysis for qualitative and quantitative data. The students will also be exposed to R programming language for data management, analysis, and reporting.

The training will be fully conducted using R programming language and participants are expected to have the latest version of R programming and RStudio downloaded and installed on their computers.

1.1 Target audience

The course is targeted for postgraduate students but not limited to in any of the following fields: Agricultural Economics, Plant Protection, Food Science, Natural Resource Management, Aquaculture, biological sciences, Fisheries Sciences among others.

1.2 Aim

The ultimate aim for the follow-up training is to provide detailed analysis on the use of different statistical methods to postgraduate students, as well as researcher, improve efficient flow of agricultural information and research specifically to achieve the following; understand the various statistical inference 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 postgraduate students with detailed skills and knowledge in use of R programming language for data 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 follow-up training will give them more hands-on skills they need to improve the quality and quantity of their research publications. The training will also be an opportunity for postgraduate students to prepare their draft thesis.

1.3 Specific Objectives

a) Students will be able to capture paired and unpaired dataset, use independent t-test, dependent t-test using surveys and experimental designs.
b) Students will learn how to compare three and more means using one-way ANOVA, two-way ANOVA and repeated measures ANOVA.
c) Students will learn how to manipulate their field data in R using regression analysis, for example, simple and multiple regression, logistic, log-linear, ordinal and multinomial regression.
d) 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) 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) Students will be able to use R programming language to analyze data using

categorical data analysis techniques and ranking of scores.

1.4 Course Outcomes (Expectations)

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

(i) Describe the fundamental concepts behind experimental/survey designs and statistical data analyses within the context of developing countries.

(ii) Apply key statistical concepts such as 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

Delivery will be mixed model, including interactive lectures and practicals 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. 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 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 practicals and discussion on the statistical results.

Participants will analyze their data using techniques already introduced on a daily basis. Discussions on interpretation and presentation of the results will be held every day during the plenary sections. The participants will evaluate the modules on a daily basis 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 first SDM training using R The modules provide 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 six working days each day starting at 10:00 am up to 2:00 pm, with breaks in between.

.

2.0 An overview training content

The following modules were extracted based on the previous training using R-programming language. These are follow-up modules to support detailed analysis in student's research. Since it is a follow-up training, the topics intends selected will be applied for both surveys and experimental designs.

2.1 Module 1- Comparing two means using R

Capturing paired and unpaired data in the excel sheets, use of dependent t-test in the context of cross-sectional research or surveys and experimental designs when using two related scale variables. Testing assumptions for a dependent t-test, statistical significance, effect size, data visualization in form of simple bar charts, independent t-test considering one scale and one dichotomous variable. Independent t-test for both surveys and experimental designs, test assumptions, statistical significance and graphs, Rmarkdown.

2.2 Module 2- Comparing three and more means using R

Use of one-way independent ANOVA (one scale variable and one categorical variable), application to both surveys and experimental designs, test assumptions, effect size, logic of F-ratio, Follow-up tests: Which groups differ from which? option 1: planned contrasts/ planned comparisons, option 2: post hoc procedures, post hoc comparisons, one-way ANOVA: effect sizes, two-way ANOVA and repeated measures ANOVA, residuals. REML;

2.3 Module 3-Regression analysis using R

Recap on simple linear regression, multiple regression, validation of assumptions under regression etc; model building: - modelling and analysis of experimental and survey data; Introduction to types of statistical models (Generalised linear models [log-linear models, logistic regression model, ordinal regression model, multinomial regression models) and their applications. Interpretation, presentation and discussion of results.

2.4 Module 4-Analysis of categorical data using R

Contingency tables, use of chi-square tests, analysis of score and rank data applications of GLM in survey data. Performing the test, causation/ use of control group. Hypothesis testing (type 1 error, type 2 error) how to construct a hypothesis test, how to prove a hypothesis test, p-values. Presentation, interpretation, and discussion of results with reference to the set objectives.

2.5 Module 5. **Cross cutting issues and wrap up**

This will cover individual student presentation of the analysis of their data and sharing key outputs at plenary.

3.0 An overview of course outcomes

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

3.1Outcomes for Module 1 on comparing two means using R

- Understanding when to apply dependent t-test on either survey and experimental data.
- Understanding when to apply independent t-test on either survey and experimental data.
- Capturing paired and unpaired dataset into excel sheets. Importing same data into R and exporting R datasets into excel.
- Testing assumptions and level of significance for independent and dependent t-test.
- Do data visualization applied to paired and unpaired datasets 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.

3.2 Outcomes for Module 2 on comparing three and more means using R

- Applying one-way independent ANOVA on both survey and experimental designs
- Applying two-way ANOVA on both survey and experimental designs
- Testing assumptions for both one way and two-way ANOVA.
- Understanding the effect size, logic of F-ratio.
- Describe Follow-up tests: planned comparisons and post hoc procedures/comparisons.
- Understanding effect sizes, residuals and repeated measures ANOVA

3.3 Outcomes of module 3 on regression analysis using R

- Distinguish simple and multiple linear regression analysis
- Validation of assumptions under regression and model building for both experimental and survey data
- Recognize when to apply different generalized linear models.
- Analyze data using log-linear models, logistic regression, ordinal regression, multinomial regression and interpret results correctly
- Build models using different scenarios of studies
- Validate final model of the analysis

3.4 Outcomes of module 4 on analysis of categorical data using R

- ❑ Understanding how to use cross-tabs and chi-square tests.
- ❑ Use of analysis of score and rank data applications.
- ❑ Constructing hypothesis test for a chi-square and p-values
- ❑ Prove different hypothesis test based on the research objectives.
- ❑ Presentation, interpretation and discussion of results

4.0 Annexes

4.1 **Annex 1**: **Training Programme (October 4th-9th, 2021)**

Time	**Monday (16/8/21)**	**Tuesday (17/8/21)**	**Wednesday (18/8/21)**	**Thursday (19/8/21)**	**Friday (20/8/21)**	**Saturday (21/8/21)**
10:00am-10:20am *10:25am-10:45am*	Registration Opening Ceremony Objectives & overview of course (**All**)	Comparing two means using R with experimental data Review of exercises done offline by students **HN/SB**	Review of exercises done offline by students **SB/HN** Performing analysis of variance with examples (dataset) Using general linear analysis REML approach using R **SB/HN**	Review of exercises done offline by students HN/SB	Review of exercises done offline by students • Multinomial regression model continued **SB/HN**	Review of exercises done offline by students Individual presentation **All**
10:45am-11:00a.m	Tea Break	Tea Break	Tea Break	Tea Break	Tea Break	Tea Break
11:00am-1:00pm	Recap on Installation of R Recap on Introduction to R programming language **HN/SB**	Comparing three or more means using R with provided datasets for both surveys and experimental data A brief theory on comparing three or more means -One-way ANOVA -Two-way ANOVA **SB/HN**	Regression analysis Recap on simple and linear regression, -Validation of assumptions under regression and model building **HN/SB**	Relationships & Association Generalized linear models using -log-linear models -Hypotheses testing. **SB/HN**	Performing cross-tabs and chi-square tests with examples (dataset) **HN/SB**	Individual presentation of research design and data analysis applied to their research **All**
1:00pm-1:15pm	***Lunch Break***	***Lunch Break***		***Lunch Break***	***Lunch Break***	***Lunch Break***
1:15pm-2:00pm	Comparing two means using R with survey data -Introduction to RMarkdown **HN/SB**	-Planned comparisons -Post hoc comparisons -Effect size, residuals, and repeated measures ANOVA **SB/HN** **SB/HN**	Relationships & Association Generalized linear models using • Logistic model **HN/SB**	Relationships & Association • Ordinal regression models • Multinomial regression model **SB/HN**	Relationships & Association • Use of analysis of score and rank data applications **HN/SB**	Individual presentation of research design and data analysis applied to their research and wrap-up **All**

2:05pm	*Offline activities (on RMarkdown)*	*Offline activities (Doing one-way and two-way ANOVA with experimental dataset)*	*Offline activities (Doing logistic regression model analysis for quantitative data exercise)*	*Offline activities (performing an ordinal and multinomial regression analysis)*	*Offline activities (analyzing data with chi-squares test, use of analysis of score and rank data)*	*End of training*

SB *– Susan Balaba Tumwebaze****; HN****- Hellen Namawejje*

Timelines

Day one:**(**4/10/21)

Participants will learn how to capture paired and unpaired data in excel and use R programming language to manage data, writing the code for the dependent t-test and independent t-tests. After day 1 of training, we expect participants to have understood how to use dependent and independent t-test using R.

Day two: (5/10/21)

Participants will be given field data sheets and analyze it using one-way and two-way ANOVA, performing post hoc procedures and planned comparisons using R

Participants will be given offline exercises on analyzing experimental data and survey data using and will present results the following day

Day three: (6/10/21)

Participants will be exposed to the theory of Regression and Generalized linear models (simple, multiple and logistic regression)

Provided with data sets, which will be analyzed offline and later be presented in the morning of 7/10/21

Day four: (7/10/21)

Participants will be exposed to the theory of Regression and Generalized linear models (ordinal and multiple regression)

Provided with data sets, which will be analyzed offline and later be presented in the morning of 8/10/21

Day five: (8/10/21)

Participants will be exposed to contingency tables and how to use a chi-square test as well as ranking scores. Offline exercises will be given to participants to present the morning of 9/10/21

Day six: (9/10/21)

Each student is given 10 minutes to present their research design and data analysis skills acquired and applied to their studies

Wrap-up

N.B: Evaluation of participant's expectation will be conducted every day in a week and an overall assessment of the training using google forms/surveys/monkey.

4.2 Annex 2: Training needs assessment tool (This questionnaire will be sent to participants in form of a google form or survey monkey)

Pre-Course Questionnaire for Follow Up Scientific Data Management Course Schedule for October -04, 2021

Section A : Personnel Information

1. Sex: □ Male □ Female
2. Name of degree being under taken (Options will be probided)
____
3.Organisation/University____
4. Title of your research____
5. Stage of research process (options will be provided)
____

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

Data Management

10. Knowledge in design of data collection tools: □ None □ Slight □ Moderate □ Good □ Excellent

11. Knowledge in data management techniques: □ None □ Slight □ Moderate □ Good □ Excellent

12. Knowledge in design of spreadsheet for data entry: □ None □ Slight □ Moderate □ Good □ Excellent

13. Knowledge in data checking: □ None □ Slight □ Moderate □ Good □ Excellent

14. Knowledge in importation to statistical Software: □ None □ Slight □ Moderate □ Good □ Excellent

Basic Statistics and Interpretation of Results

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

□ None □ Slight □ Moderate □ Good □ Excellent

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

□ None □ Slight □ Moderate □ Good □ Excellent

17. Knowledge in data interpretation and reporting: □ None □ Slight □ Moderate □ Good □ Excellent

18. Knowledge in design of data collection tools: □ None □ Slight □ Moderate □ Good □ Excellent

19. Knowledge in presentation of results: □ None □ Slight □ Moderate □ Good □ Excellent

Statistical Software and Computer Skills

20. What is your level of knowledge in R?

□ None □ Slight □ Moderate □ Good □ Excellent

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

□ None □ Slight □ Moderate □ Good □ Excellent

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

□ None □ Slight □ Moderate □ Good □ Excellent

SPECIFIC TOPICS

24. 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)										
	Interpretation of statistical results										

Course Facilitators: Dr. Susan Balaba Tumwebaze **(susantumwebaze@gmail.com);**
Dr**. Hellen Namawejje (hnamawejje@gmail.com)**

Thank you for taking off time to complete this questionnaire