We present a complete course in data handling and analysis to run internally in the department. The complete course, which encompasses a three-day pre-course on handling data through R and a week course on Biostatistics, can provide a total of 5 ECTS. The course in this first version is planned for a maximum of 15 students, due to the high demand for supervision and continuous feedback on the PhD projects that the teaching methodology requires.

Registration: Biostatistical modelling for Ag. Science 2027 - Eventværk

Registration deadline: 31 January 2027

Biostatistical modelling for Ag Science

Hands-on data: an applied statistics PhD course

Aim

Provide the participants with a place and the tools for methodological discussion around data analysis in the context of research on Agroecology. The participants will finish the course being able to recognize the elements of experimental or observational design involved in their research and model them in order to respond to the research question.

ECTS: 3.5 One week classes 90 hs. (40 hs. in class and 50 preparation).

Intended Learning Outcomes

- Manage R statistical software that allows reshaping and merging data of various types.
- Choose and apply exploratory data tools to discover patterns in data.
- Design an experiment.
- Conduct basic statistical algorithms (Generalized Linear Mixed Models, with main focus on general linear models) for data analysis and understand their strengths and weaknesses.
- Describe, interpret and discuss the results and shortcomings of an analysis based on statistical modelling.
- Build attractive and informative graphics and tables from applied statistical analysis.
- Develop a criterion on scientific ethics when reporting the results of an applied statistical analysis.
- Use AI tools critically and transparently across the statistical analysis workflow, formulating effective prompts, recognising their limitations, and maintaining responsibility for the validity and reproducibility of results.

Content

I. *Data handling and exploratory Analysis*
   i. Data types and random variables.
   ii. Descriptive qualitative data analysis.
   iii. Descriptive quantitative data analysis.

II. *Introduction to statistical modelling*
   iv. Notation for statistical modelling.
   v. Modelling experimental and observational data.

vi. Statistical modelling and uncertainty.
vii. Classification and Regression models.

III. *Experimental design in agricultural science*

viii. Randomization and Replication. Randomization restriction strategies.
ix. One-way classification experiments. Experiments with factorial treatment structure. Crossed and nested factors.
x. Experiments with plot structures. Completely randomized designs, blocked designs, split-plot designs. Combining factorial treatment structures with plot structures.
xi. Experiment with temporal and spatial structures.
xii. Type of sampling for observational studies.

IV. *General(ized) Linear Mixed Model*

xiii. Classification and Regression linear models. Simple and multiple linear regression. Estimation and confidence intervals. Hypothesis testing. Predicted values, confidence bands, and prediction intervals. Residual analysis. Model adequacy.
xiv. Random effect models. General concepts. Marginal models and subject-specific models. Models for residual covariance structure. Estimation of covariances in normal populations. Inference on random effects. Best Linear Unbiased Predictor (BLUP). Goodness-of-fit criteria and model selection. Models for longitudinal data.
xv. Generalized models. A practical guide to exponential probability families. Count data. Logistic regression.
xvi. Critical evaluation of AI-suggested model specifications and interpretations.

V. *Results and analysis communication*

xvii. Good practices in presenting results
xviii. Principles of reproducible research (Open Data, version control through Git and GitHub).
xix. Responsible use of AI in research: disclosure, data privacy and researcher accountability.
xx. Discussion on the limitation of biostatistical modelling

Instructors

Maarit Mäenpää & Franca Giannini-Kurina

Fundament

In the field of Agroecology, rigorous data-driven research has become crucial for addressing complex agricultural and environmental challenges. The ability to effectively manage and analyse data, including the responsible use of emerging AI tools, is paramount for producing high-quality research outcomes and making informed decisions.

The present course is designed after pedagogical research that investigated the expectations and perceptions of statistical literacy of both PhD students and supervisors to design a postgraduate course in applied biostatistics in the Department of Agroecology that constructively aligns the expectations of both parts. A survey of VIP staff was designed and carried out to assess perceptions on the level of proficiency, teaching focus and types of activities of the classical learning outcomes of a basic postgraduate course in applied statistics. The results underline the need for a curriculum that focuses on the development of practical independent skills in data handling and classical statistical analysis, prioritising individual work on data belonging to the PhD project and developing

competences on a critical quantitative scientific criterion. Consequently, the learning design of the course includes intended learning outcomes, activities, and assessment for and of learning inspired by the community needs.

Teaching context

This course attempts to differentiate itself from other postgraduate training programs in basic statistics by applying a student-centred pedagogical strategy where the teaching format includes lecturing and case-based teaching. The cases are based on PhD students' own research problems and data to make the cases more relatable, relevant and authentic, and the cases are to be presented and discussed in in-class activities.

Assessment

The assessment for and of learning has been designed to ensure that students not only gain theoretical knowledge but also develop the ability to effectively analyse, interpret and communicate their results in the context of their research questions, making their learning experience more meaningful and applicable to their future research endeavours. The course will culminate in a collective assessment in an open-door session where participants will present the report of their own analysis.

Program (1-5 of March, 2027)

Workshop 1

My research question into a statistical model.

Data types and random variables. Descriptive qualitative data analysis. Descriptive quantitative data analysis. Notation for statistical modelling. Modelling experimental and observational data. Statistical modelling and uncertainty.

Workshop 2

The beauty of a "simple" linear model: How simple and how linear is a LM?

Classification and Regression linear models. Simple and multiple linear regression. Estimation and confidence intervals. Hypothesis testing. Predicted values, confidence bands, and prediction intervals. Residual analysis. Model adequacy. Generalized linear (mixed) models: exponential family distributions, link functions, and models for counts and proportions.

Workshop 3

No replicates, no: Why we never should have a column named "replicate" and the definition of my experimental design.

Randomization and Replication. Randomization restriction strategies. One-way classification experiments. Experiments with factorial treatment structure. Crossed and nested factors. Experiments with plot structures. Completely randomized designs, blocked designs, split-plot designs. Combining factorial treatment structures with plot structures. Type of sampling for observational studies.

Workshop 4

What in hell is a random effect?... and in heaven?

Random effect models. General concepts. Marginal models and subject-specific models. Models for residual covariance structure. Estimation of covariances in normal populations. Inference on random effects. Best Linear Unbiased Predictor (BLUP). Goodness-of-fit criteria and model selection. Models for longitudinal data. Experiment with temporal and spatial structures.

Workshop 5

A life between ctrl + C and ctrl + V: Open to openness and how much work will take/save us.

Introduction to open science in the context of data analysis. Good practices in presenting results

Workshop 6

Open house: presenting quantitative results from our statistical analysis.

Discussion on the limitation of biostatistical modelling. Assessment

Bibliography

Casler, M. D. (2015). Fundamentals of experimental design: Guidelines for designing successful experiments. Agronomy Journal, 107(2), 692-705.

Efron, B., & Hastie, T. (2021). Computer age statistical inference, student edition: algorithms, evidence, and data science (Vol. 6). Cambridge University Press.

Glaz, B., & Yeater, K. M. (2020). Applied statistics in agricultural, biological, and environmental sciences (Vol. 172). John Wiley & Sons.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. New York: Springer.

Pedersen, E. J., Miller, D. L., Simpson, G. L., & Ross, N. (2019). Hierarchical generalized additive models in ecology: an introduction with mgcv. PeerJ, 7, e6876.

Piepho, H. P., Büchse, A., & Emrich, K. (2003). A hitchhiker's guide to mixed models for randomized experiments. Journal of Agronomy and Crop Science, 189(5), 310-322.

West, B. T., Welch, K. B., & Galecki, A. T. (2022). Linear mixed models: a practical guide using statistical software. CRC Press.

Pre-workshop (necessary to complete 5 ECTS)

Handling Ag Data through R

Aim

Reinforce the understanding and skill development on data handling and coding in R necessary to the course.

ECTS: 1.5 ECTS (38 hs., three sessions and 10 hs. homework)

Intended Learning Outcomes

- Recognize the R programming environment, including the installation and configuration of R and RStudio, and basic syntax and operations.
- Design a structure and wrangle data sets for a specific aim.
- Construct a variety of data visualizations using the ggplot2 package.
- Recognize the principles of reproducible research and open data, and document data sets with a data dictionary.
- Explain the basic functioning and limitations of generative AI tools, and use them responsibly to write, debug and document R code through effective prompting.

Content

I. Getting started with R.

a. Introduction to the R working environment.

i. What is R? What is RStudio? Download and installation of R and RStudio. Packages, documentation, and help in R.

ii. Starting a work session in R. Creating work projects in RStudio. First functions. R as a calculator. Language syntax. Statements. Assignments.

iii. Mathematical operations. Comments. Saving commands (scripts) and projects. Matrix Algebra. Nomenclature and use of functions.

b. Data files (data frame)

iv. Reading data, importing data from Excel and other formats, reading files from a working directory. Exporting data, saving objects, and R workspace.

v. Handling files and data, sorting, selecting rows, selecting columns, creating data subsets.

vi. Exploring data frames with specialized packages, identifying missing values. Lists.

c. Graphics

vii. The plot() function. Scatter plots. Graphical attributes. Histogram. Box plot. Bar chart.

viii. Introduction to Grammar of graphics and advanced plotting using ggplot2 package.

d. AI-assisted coding

ix. Foundations of generative AI: how large language models produce answers, typical failure modes (errors, fabricated functions or packages, outdated syntax), and why output must be verified.

x. Prompting for data analysis: giving context (research question, data structure, R packages in use), asking for explanations rather than just code, iterating and refining prompts, and testing generated code.

II. *Managing, handling, and storing data responsibly*

i. Principles of creating a data dictionary for reproducible research.

Instructors

Maarit Mäenpää & Franca Giannini-Kurina

Assessment

The pre-workshop will finish by building a collection of descriptive figures generated by the participant from their own data able to describe the experimental or observational design in their research, and a data dictionary detailing their data.

Program (24-26 of February, 2027)

Workshop 1

Building my first proud plot in R

What is R? What is RStudio? Download and installation of R and RStudio. Packages, documentation, and help in R. Starting a work session in R. Creating work projects in RStudio. First functions. R as a calculator. Language syntax. Statements. Assignments. Mathematical operations. Comments. Saving commands (scripts) and projects. Matrix Algebra.

Reading data, importing data from Excel and other formats, reading files from a working directory. Exporting data, saving objects, and R workspace. Handling files and data, sorting, selecting rows, selecting columns, creating data subsets. Exploring data frames with specialized packages, identifying missing values. Lists.

The plot() function. Scatter plots. Graphical attributes. Histogram. Box plot. Bar chart. Introduction to Grammar of graphics and advanced plotting using ggplot2 package.

Workshop 2

A first aid kit for our data journey

Defining functions in R. Function arguments. Writing, debugging and documenting functions with the help of AI tools: prompting well and checking the result.

Workshop 3

Dear future me, what on earth is “X17_final_v2”?

Principles of creating a data dictionary for reproducible research: variable names, descriptions, units, types, allowed values and missing-value codes. Naming conventions and tidy data.

Bibliography

Gandrud, C. (2018). Reproducible research with R and R studio. Chapman and Hall/CRC.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... & Yutani, H. (2019). Welcome to the Tidyverse. Journal of open-source software, 4(43), 1686.

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). R for data science. O'Reilly Media, Inc.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9.

Xie, Y., Dervieux, C., & Riederer, E. (2020). R markdown cookbook. CRC Press.