May Institute on computation and statistics for mass spectrometry and proteomics

April 30, 1:30pm – May 2, 5:00pm **Beginner R and beginner statistics**

Venue:
Northeastern University main campus, West Village H, first and second floor rooms: 110, 108, 210A/B, 212. Please see the annotated map.

Lead instructor:
Ryan Benz, rwbenz@gmail.com

Speakers
Olga Vitek

Teaching assistants
Sarah Szvetecz, Ritwik Anand, Ethan Rogers, Vartika Tewari

Description
This course has two objectives. First, it will introduce the practical steps of statistical analysis using the open-source environment R. In addition to discussing basic data management tasks in R, such as reading in data and performing basic analysis, it also contains an introduction to reproducible research using R markdown. It will also introduce the concept of *tidy data* and the tidyverse ecosystem of R packages. In particular, we will focus on the dplyr R package for performing basic data manipulations on data tables (data frames), providing a foundation for performing data analysis on your own data.

Second, it will introduce the participants to the basics of statistical experimental design, data analysis, and statistical inference. It will provide basic introduction to topics such as randomization, error bars, hypothesis testing, and simple linear regression. The course will contain both lectures and practical hands-on exercises.

We will be using the RStudio Integrated Development Environment (IDE) application (free, open source) to work with R code and the hands-on exercises. The course will alternate between short, lecture style sessions to introduce and explain the material, followed by hands-on exercises for practice. Time will also be allocated throughout the course for questions.

References
The course will combine lectures and practical hands-on exercises. The discussion of programming with R is based on the following textbooks:

- Grolemund & Wickham. R for Data Science, O'Reilly, 2017
- Susan Holmes and Wolfgang Huber. Modern Statistics for Modern Biology, Cambridge University Press, 2019
- Points of Significance Collection, Nature Methods.

Target Audience

Target audience are experimental scientists with no prior knowledge of statistics or R. Participants are expected to have an understanding of basic data manipulation/analysis (e.g. experience working with data in Excel).

Set-up before the course

Beginners Python and beginners statistics

Tentative schedule

Wednesday April 30, 2025
12:30 p.m. Registration
1:30 p.m. **Introduction to Statistics**, Olga Vitek
3:00 p.m. Refreshments
3:30 p.m. **Introduction to R and RStudio**, Ryan Benz
5:00 p.m. Q&A and adjourn

Thursday, May 1, 2025
9:00 a.m. **R Coding Essentials: variables, vectors, data frames, functions**, Ryan Benz
10:30 a.m. Refreshments
11:00 a.m. **Principles of statistical inference**, Olga Vitek
12:30 p.m. Lunch on your own
1:30 p.m. **Data exploration with dplyr**, Ryan Benz
3:00 p.m. Refreshments
3:30 p.m. **Data visualization with ggplot2**, Ryan Benz
5:00 p.m. Q&A and adjourn

Friday, May 2, 2025
9:00 a.m. **Experimental design. Class discovery and class prediction**, Olga Vitek
10:30 a.m. Refreshments
11:00 a.m. **Reproducible data analysis with Quarto**, Ryan Benz
12:30 p.m. Lunch on your own
1:30 p.m. **Code version control with GitHub**, Ryan Benz
3:00 p.m. Refreshments
3:30 p.m. **Practice & exercises**, Ryan, Olga, TAs
5:00 p.m. Wrap-up and adjourn

Link to participant lunch order pickup

If you would like us to pick up your prepaid lunch order from Tatte or Anna's Taqueria, fill this form: https://forms.gle/5E8apFjvjW3qZhgQA

Course evaluation

Please help us improve the program in the future by filling in this form: https://forms.gle/s6bbgAUGhdDy6GjG9