**One Health Summer Institute – Addis Ababa, Ethiopia**

**Introduction to Data Analytics, Statistics & Programming in Epidemiologic Research**

**August 13 – 16, 2019**

**9:00 am – 5:00 pm**

**Course Instructors**:

Dr. Barbara Kowalcyk - kowalcyk.1@osu.edu

Dr. Ewy Mathé - ewy.mathe@osumc.edu

**Course Description**:

This course is an introduction to principles of data analysis, including data cleaning, quality checking and visualization, statistical analysis and handling large datasets. The programming language used will be R. The course objectives are to:

1. Provide guidelines for best practice in reproducible research;
2. Provide attendees with a core set of working code, with example data, so that they can readily apply it to their own data;
3. Provide attendees with the ability to analyze publicly available data.

Lectures, case studies and small group exercises will be used to achieve these goals. At the conclusion of this course, students will have an understanding of how to approach the analysis of datasets collected from epidemiological research and informing decision-making around foodborne and zoonotic diseases.

**Requirements**:

Each student will need to have access to a computer in class with the following programs and packages installed:

* R (<https://cran.rstudio.com/>): most of the hands-on portions of the course will be performed using the R statistical language.
* RStudio Desktop (<https://www.rstudio.com/products/rstudio/download/>): this is a very intuitive program that makes it easier to learn and use R.

**Recommended Tutorials**:

To get you started, and especially if you have no experience with the R statistical language, please look at the following tutorials before class:

* A quick introduction to R/RStudio can be found here:

http://ncss-tech.github.io/stats\_for\_soil\_survey/chapters/1\_introduction/1\_introduction.html

* And here is a much more thorough introduction to R:

<https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf>

* R/Rstudio are very widely used in many different fields of research (e.g. biomedical, epidemiology, biology, etc.). Most questions can thus be answered with google! (e.g. you can paste your error message and follow them by “R” as a good way to start searching).
* Gitlab: Git is a versioning software that automates the process of versioning and saving your files. Gitlab is an environment that uses Git, and is a nice way of sharing code and data. We will be discussing this more during class. But you can take a sneak peak here: https://www.tutorialspoint.com/gitlab/gitlab\_introduction.htm

*Here are other resources as well*:

* <https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf> (I learned with something like this.  I went through all the code, writing it into an R Console, rather than just reading, and it was really helpful)
* <http://www.r-tutor.com/r-introduction> (very basic)
* <http://tryr.codeschool.com/levels/1/challenges/1> (very basic but interactive)
* <https://www.coursera.org/learn/r-programming-environment> (you should be able to access all lectures, but not assignments, more advanced)
* And Lots more here: <http://heather.cs.ucdavis.edu/~matloff/r.html>

**Course Resources**:

A Gitlab repository has been created for this course, and it can be accessed here:

[https://github.com/Mathelab/OHSI\_DataAnalytics\_Aug2019](https://github.com/Mathelab/OHSI_DataAnalytics_Aug2019.git)

The lectures and data used for this class will also be placed in this repository.

**Schedule**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Day 1** | **Day 2** | **Day 3** | **Day 4** |
| **9:00 – 10:30** | Course Introduction (Kowalcyk) | Statistical Concepts (Kowalcyk) | Handling and assessing the quality of public molecular data (Mathe) | Research ethics and reproducibility case studies (Kowalcyk/Mathe) |
| **10:30 – 11:00** | Coffee Break | Coffee Break |  | Coffee Break |
| **11:00 – 12:30** | Designing epidemiologic studies (Kowalcyk) | Hypothesis Testing, Confidence Intervals and Sample Sizes (Kowalcyk) | Advanced Statistical Methods (Kowalcyk) | Developing a Study Protocol, Data Management and the FAIR principles (Kowalcyk/Mathe) |
| **12:30-1:30** | Lunch | Lunch | Lunch | Lunch |
| **1:30 – 3:00** | Producing .Rmd files, data input, data cleaning (Mathe) | Sources of publicly available data, data structures in R (Mathe) | Statistical Analyses in R (Mathe) | Group project work |
| **3:00 – 3:30** | Coffee Break | Coffee Break | Coffee Break | Coffee Break |
| **3:30 – 5:00** | Introduction of group project | Looking at your data (summary statistics and data visualization) (Kowalcyk/Mathe) | Group project work | Group presentations |
| **Homework Assignment** |  | Read:  Jinyu Xu, Jeffrey D. Galley, Michael T. Bailey, Jennifer M. Thomas-Ahner, Steven K. Clinton & Susan E. Olivo-Marston. *The Impact of Dietary Energy Intake Early in Life on the Colonic Microbiota of Adult Mice*. Scientific Reports volume 6, Article number: 19083 (2016)  Gang Su, Charles F Burant, Christopher W Beecher,Brian D Athey, and Fan Meng. *Integrated metabolome and transcriptome analysis of the NCI60 dataset*. BMC Bioinformatics. 2011; 12(Suppl 1): S36. | Read:  Boeckhout M, Zielhuis GA, Bredenoord AL. *The FAIR guiding principles for data stewardship: fair enough?*. *Eur J Hum Genet*. 2018;26(7):931–936. doi:10.1038/s41431-018-0160-0 |  |