Matéria de Análise de Dados

Doenças Infecciosas – Pós-Graduação

Segundo Semestre 2020

In your careers, you will encounter vast amounts of data. Journal articles, theses and dissertations, even newspaper articles focus on data. We have access to rivers of data. As health and medical professionals, we need to know how to analyze data, to separate studies that communicate convincing medical and biological conclusions from those that are simply garbage. We need to build our own datasets and conduct our own analyses.

This course will help you with the fundamentals of data analysis: biostatistics and the computational skills necessary to organize your data, analyze it and present your conclusions. It is not a traditional theoretical statistics course. It will show you how to read the statistical analyses in articles, papers and presentations. It will show you how to use effectively programming tools that allow you to construct reproducible, robust scientific results.

From the first week, you will be building datasets and analyzing them. Each week your skills will grow in scope (more techniques) and deepen. We will start with basic statistical analyses that help us describe data and move on to understand how we can use biostatistics and bioinformatics to draw conclusions from sample data about the populations the samples may represent. From there, we will finish the course by examining several “machine learning” modeling algorithms that can process very large datasets and assist us to predict outcomes not immediately visible in the dataset we are using.

Students are encouraged to bring their own data to the course and to work on those data throughout the course. Each laboratory or research group has their own problems and students really want to learn this material to advance their own research programs. Working together in teams will also help students learn from each other and share out tasks that will become larger as the semester advances.

Why is this *ementa* written in English? Not just because the professor is a native English speaker. We want students in the discipline to have more opportunity to work

in English during their post-graduate study as papers, conferences and general communications with scholars and physicians beyond Brazil will generally be in English. The course lectures will be given in English. Most of the written support materials for the course are in English. Journal articles we will study are also in English. However, this is not an English language course. All students are encouraged to participate to the maximum degree possible with questions, observations and presentations. If students feel more comfortable doing this in Portuguese, that is not a problem. Any doubts about something I might say in English that you do not understand clearly, I will gladly resolve in Portuguese.

MAD will not focus on theoretical mathematics. Students do not need a strong background in mathematics. You will need to know some basic mathematical tools such as the concept of a sum (∑), logarithms and exponents, and the equation for a straight line. In the early sessions of the course, we will review some of these concepts that you may have forgotten.

We will be conducting our data manipulation and analysis using the statistical programming language "R". R is an open-source program that is freely available. Students will be expected to start the semester with R and its Interactive Development Environment, RStudio, installed on a laptop. Using a language rather than a "canned" software program like SPSS or GraphPad Prism allows us to build recipes (programs) that anyone can reproduce later. This will help us build the kind of reproducible analyses that science needs but has not had in the past. We will learn as part of our work with R basic concepts of computer programming.

Four Pillars of MAD

MAD will address four principal themes:

1. The basic concepts of biostatistics
2. The organization, cleaning and practical analysis of data
3. The computation and programming tools that support the manipulation and analysis of data
4. The workflow that needs to be followed to execute successfully a data project

That is the *what* of the course. How will we accomplish this? We will have presentations of the material by the professor and by the work groups that the students will form. We will also have traditional problem sets that students will execute in their work groups. The third key component is a project that students will develop in their work groups throughout the semester. This can be based on data coming from the student's own research program or from other sources.

The distance learning model that we need to follow because of the Covid-19 pandemic works with the educational model outlined in the previous paragraph. While my preference would be to get to know all the students personally and have informal exchanges with you, we need to follow the protocols that EPM has established for this

academic year. However, we will use technology to the maximum possible to get to know each other.

The Four Pillars in Greater Depth

In order to build the structure of the four pillars, we will look at the following set of topics to lead us to being able to construct and an entire data analysis study. The course will focus on classic, or "frequentist" statistics. It will only peripherally present modern Bayesian methods that use a different understanding of probability than the classic methods.

1. Introduction to the Course:
    a. What is statistics and why we use it
    b. Basic introduction to R and how to use it
2. Data analysis workflows
3. How to collect data and build a dataset
4. Basic descriptive statistics and visualization of data
5. Data Cleaning – refining the dataset
6. Data distributions – theoretical and sample
7. Basic statistical inference
    a. Central Limit Theorem
    b. Inferences concerning 1 and 2 variables
    c. Parametric vs. Non-Parametric
8. Simple Linear Regression (the most basic model)
    a. Correlation and Causation
    b. Extensions to SLR
9. ANOVA – Analysis of Variance
10. Machine Learning Models
    a. Supervised methods: Regression/Classification
    b. Unsupervised methods: clustering, Principal Components, Naïve Bayes
11. Postscript: Bayesian Analyses

Additional Support for Students

In addition to the formal classes, problem sets and group projects, there will be an extensive set of open-source readings made available to students, including the current draft of a data analysis textbook that Dr. Hunter is preparing (see the Bibliography below). In addition, there will be a two-hour period weekly, known in the US as "office hours", when Dr. Hunter will be available for consultations and help, both electronically and in his EPM office. Students in previous sessions of similar courses have greatly benefited from these sessions.

James R. Hunter, Ph.D. Biography

James Hunter earned his Ph.D. in Infectious Diseases at UNIFESP in 2019. He has previously earned a B.A. and a M.C.P. (Master of City Planning), both from Yale University in the United States. Since 1970, he has taught courses on quantitative methods, statistics and operations research in the United States, England, Canada and Brazil.

Dr. Hunter has recently completed a Bioinformatics Post-Doctoral Fellow in the CAPES PRINT program at Escola Paulista de Medicina. He continues his research into HIV-1 and related viral diseases in the EPM's Retrovirology Lab. He has been named a Professor Afiliado at EPM. He has lived in Brazil since 1999 and came to UNIFESP in 2014.

Bibliography

All the materials used in the classes (slides, Dr. Hunter's text, etc.) will be placed in a freely available repository on GitHub. Below is a list of books that will be useful to students in the course. Most of them are free.

Bibliography — Statistics

- Diez, Barr & Cetinkaya-Rundel, OpenIntro Statistics 4, (http://openintro.org)
- Navarro, D. Learning statistics with R: A tutorial for psychology students and other beginners, (http://learningstatisticswithr.com)

Bibliography — R, Programming and Analysis of Data

- Irizzary, Introduction to Data Science (https://rafalab.github.io/dsbook)
- Irizarry & Love, Data Analysis for the Life Sciences (Leanpub)
- Kabacoff, R in Action: Data analysis and graphics with R, 2e (Manning)
- Peng, R Programming for Data Science (Leanpub & Bookdown)
- Peng, Kross & Anderson, Mastering Software Development in R (Leanpub & Bookdown)
- Wickham & Grolemund, R for Data Science, (http://r4ds.had.co.nz ou O'Reilly)

Livros de Leanpub: https://leanpub.com; Livros de Bookdown: https://bookdown.org/

5 Books You Should Read about Statistics and Data Analysis (Because They are Good)

- Leonard Mlodinow, O Andar do Bêbado
- David Salsburg, Uma Senhora Toma Chá
- Ian Stewart, 17 Equações que Mudaram o Mundo
- Peter L. Bernstein, Desafiando os Deuses: A História do Risco
- Randall Munroe, E Se?: Respostas Científicas para Perguntas Absurdas