Syllabus (Fall 2021)
FC721 Statistical Reasoning for the Basic Biomedical Sciences
Room: R103
Tuesdays and Thursdays 4:00-6:00pm.

Thomas B. Kepler, Department of Microbiology, Department of Mathematics & Statistics
tbkepler@bu.edu, 617 358 5989, W508D, Medical Campus

By definition, scientific research operates at the boundary between what is known and what is unknown. The practitioner must operate effectively in the face of uncertainty though intuition is easily misled under such circumstances. Statistics provides a basis for identifying and correcting errors of judgement as well as the tools for drawing reliable inferences from data. In this class, we eschew the dry, detached approach to statistics and instead treat the subject as an integral part of scientific research—a natural extension of basic scientific methods. On the one hand, we will develop *judgment* for the design of experiments and the inference of knowledge from them. On the other, *technique*, to carry out data analyses using specialized computational tools effectively.

We will develop judgment by examining the key concepts of statistics and their theoretical underpinnings, while developing sound intuitions through extensive in-lecture short exercises in R. We will develop technical ability through extended R exercises using R Studio and best reproducible data-analytical practices using R Notebooks, and Git. All software used in class is available free of charge; each has extensive online communities providing help at all levels. No previous experience with any of these tools is assumed.

The class discussions are essential and require vigorous participation from all students. I am not, for example, going to define *hypothesis testing* prior to our discussion of it. Instead, we will call on our own experience with the way data are used in biomedical investigations and figure it out ourselves. I will exert just enough guidance to ensure that we end up at the consensus views, but we will get there on our own. Once we understand what we want to do, we will learn how to get it accomplished in R. The same goes for a great many other topics.

Zoom

To facilitate sharing of completed assignments, we will use Zoom for screen sharing. We will not use Zoom as an alternative to in-person class attendance. Do not share the link and login information outside of the class.

GMS FC721
https://bostonu.zoom.us/j/99646567362?pwd=ekVwQ01zdDNoMjAvQkQzVlc3Y3pXUT09
Meeting ID: 996 4656 7362
Passcode: 310433

Computers and Software

Please bring your own personal computer to class. The software we will be using can be run on all the major platforms, Windows, Mac, and Linux. We will be using the following freely available software packages, to be installed on each student machine.

R https://www.r-project.org

R Studio https://www.rstudio.com/products/rstudio/download

Git https://git-scm.com/downloads

Readings

The course will be taught from the instructor's notes and the freely available online text, *R for Data Science* by Hadley Wickham and Garrett Grolemund https://r4ds.had.co.nz/index.html. The source code containing the entirety of the book is hosted on GitHub at https://github.com/hadley/r4ds. We will be referring directly to material in this book and the exercises it contains. All other books are for reference only, and are optional.

Further material on statistical graphics using ggplot is provided by the textbook by Hadley Wickham, Danielle Navarro, and Thomas Lin Pedersen entitled *ggplot2: Elegant Graphics for Data Analysis* available online at https://ggplot2-book.org/.

There are quite a few other reliable textbooks online that can be used to supplement the material in this course, including

- *Modern Statistics for Modern Biology* by Susan Holmes and Wolfgang Huber, at http://web.stanford.edu/class/bios221/book/.
- Peter J. Diggle and Amanda G. Chetwynd (2011) *Statistics and Scientific Method: An Introduction for Students and Researchers.* The table of contents is here http://www.gbv.de/dms/tib-ub-hannover/65406072x.pdf.

From time to time, we will read additional material from the primary literature.

Grading

The course will be graded pass/fail and the grade based on performance on the in-class discussions, homework assignments, and a final project. The assignments will typically ask students to choose the most appropriate statistical methods for a specific experiment, to justify their choices, and carry out the analyses. They will be asked to design experiments under relevant constraints and carry out simulated experiments to check their assumptions. A solution page will be used to grade each assignment. The final examination will entail the complete analysis of a large, complex dataset of the student's choosing and their presentation of that analysis to the class.

Criteria: A Pass will require regular participation in class discussions, including discussions of the assignments, as well as satisfactory completion of the final project. I will assume that students come to class prepared for discussion and presentation of their assignments. Students in danger of failure will be notified and invited to review their situation with the instructor.

Teaching Assistance

Jamie Strampe (jstrampe@bu.edu) will be our teaching assistant this semester. She is a Bioinformatics PhD student working with John Connor of the Department of Microbiology. Please feel free to contact her or me with questions or concerns.

Calendar of Topics

This is the planned order of investigation. We may take diversions and add or subtract material as needs or opportunities arise.

Month	Date	Day	Notes
September	2	Thursday	Syllabus, Software, Introduction
September	7	Tuesday	R Studio and Git
September	9	Thursday	NO CLASS
September	14	Tuesday	descriptive statistics, data wrangling
September	16	Thursday	data wrangling
September	21	Tuesday	exploratory data analysis
September	23	Thursday	exploratory data analysis
September	28	Tuesday	probability
September	30	Thursday	probability
October	5	Tuesday	probability
October	7	Thursday	NO CLASS
October	12	Tuesday	estimation and hypothesis testing
October	14	Thursday	estimation and hypothesis testing
October	19	Tuesday	linear models
October	21	Thursday	linear models
October	26	Tuesday	generalized linear models
October	28	Thursday	generalized linear models
November	2	Tuesday	nonlinear regression
November	4	Thursday	NO CLASS
November	9	Tuesday	model graphics
November	11	Thursday	model selection
November	16	Tuesday	experimental design/ANOVA
November	18	Thursday	experimental design/ANOVA
November	23	Tuesday	experimental design/ANOVA
November	25	Thursday	Thanksgiving Holiday
November	30	Tuesday	multiple comparisons
December	2	Thursday	Multivariate Statistics/RNA-seq
December	7	Tuesday	Multivariate statistics/RNA-seq
December	9	Thursday	Machine Learning
December	14	Tuesday	Exams
December	15	Wednesday	Exams
December	16	Thursday	Exams
December	17	Friday	Exams