Class logistics

Syllabus: Spring 2022

Course details

- **Time:** Tuesdays, 6:00pm to 8:50pm
- **Location:** Online (initially), then FMH 214
- **Instructors**
  - Dr. Gareth Russell
  - Dr. Simon Garnier
- **Office hours:** GR: by appointment (russell@njit.edu). SG: by appointment (garnier@njit.edu).
- **Pre-requisites:** Technically none. Most students students will have had at least one basic, undergraduate-level course in statistics. If you have not, please see one of us before the beginning of class.
- **Textbook:** Because of the variety of methods covered, there is no appropriate single textbook, and so one is not required. Instead, students will be working with a set of interactive 'notebooks' which can be downloaded from the course website. Students will, however, be advised on useful texts for various disciplines and applications. A short list is provided below. Most of these are available in the library, and some are available for consultation from the instructor's private library. Students are encouraged to select and purchase one or more suitable reference texts when it becomes clear what sort of research they are likely to be doing.
  - Ray Hilborn and Marc Mangel: The Ecological Detective. Princeton UP. *A good overview of statistical inference — not just for ecologists.*
  - Manley: Randomization, Bootstrap and Monte-Carlo methods in Biology. Chapman and Hall.
  - Derek A. Roff: An Introduction to Computer Intensive Methods of Data Analysis in Biology. Cambridge UP.
  - Emanuel Paradis: Analysis of phylogenetics and evolution with R (2006). Springer.
  - Robert Sokal and James Rohlf: Biometry. W. H. Freeman.
  - Nicholas Gotelli: A Primer of Ecological Statistics. Sinauer.
  - Bolker, Benjamin: Ecological Models and Data in R. Princeton University Press. *A 'modern' approach.*
- **Computer:** *Students are expected to bring a laptop to class* with the following free software packages installed (instructions will be given in the first class):
- **Software:** You will use three free software packages in this class. Links are also provided on Canvas.

- Wolfram Mathematica. This will allow you to open and manipulate the lecture notebooks. This is the software that you will use to access the interactive theory notebooks. You can get it by following the instructions here.
  - Note for Rutgers students: this access is via NJIT's license, which you can use while registered for NJIT courses. If you want, you can also get it through Rutgers. I think you go here, search for Mathematica, and follow the instructions. But I can't help you if you do it this way, because it requires a Rutgers NetID login. If you are not sure, get it through NJIT. More info at the Wolfram homepage.
- R. This is the statistical 'engine' that you will be using for your own statistical analyses. http://cran.wustl.edu
- RStudio Desktop. This is an alternative front-end interface for R that many prefer to the one that comes with R. It also facilitates the creation of reports in R (which mix code, results and formatted text), and you will need this ability for your class projects. https://www.rstudio.com/products/rstudio/

Learning objectives

Statistics

Students will:

1. Understand why and when statistics are necessary (and when they are not).
2. Have a sense of the broader 'landscape' of statistics, and where specific methods (like ANOVA, or regression) fit in.
3. Understand some of the philosophical differences underlying the main types of statistical inference (frequentist, information-based, Bayesian).
4. Understand why certain techniques, such as linear model fitting with Normal errors, are so common.
5. Understand how to pick an appropriate statistical approach to a problem (choosing the right distribution, model, etc.).
6. Know how to read the common kinds of outputs made by almost every statistics package (e.g., ANOVA tables, Q-Q and residual plots, posterior probability distributions).
7. Understand how statistics can (and should) help you *design* an experiment or data collection protocol.
8. Know, and know how to avoid, common mistakes!
9. Know when and why to manipulate data before analysis (e.g., transformation, removal of correlation/reduction), and when not to.

R

Students will be able to:

1. Install the R and RStudio environments.
2. Create and manipulate the basic R structures such as variables, vectors, matrices, and data frames.

3. Import and manipulate data in R using the readr and dplyr packages.
4. Run basic statistical tests in R (e.g., t-test, linear models, etc.), and understand the outputs.
5. Create visualizations using the ggplot2 package.
6. Summarize the process of data analysis in R notebooks.

Assessment and grading

Material questions: 10%

- For each class you must submit *at least one* question on the material, based on reading the notebook(s), by the end of the Monday night *before* the class.

Mid-term (written exam): 40%

- Short answer questions, definitions, etc. Tests *understanding of concepts*, not calculation ability.

R project: 40%

- Project deliverable is an *R script*, plus *data file*, that anyone can run. The script should generate an *HTML notebook*.
- The project should demonstrate all the R learning objectives by including *at least one example* of each of the following
  - **0.1.** Load some data and manipulate it in some way (sort, subset, transform, reduce…)
  - **0.2.** Visualize the data
  - **0.3.** Do some statistical analyses (linear model, GLM, logistic regression, anything appropriate).
  - **0.4.** Visualize the analysis (fitted line(s), residual plot, Q-Q…)
  - **0.5.** Do at least some of the above with an external package: more advanced graphics, data processing, special type of analysis, etc.
  - **0.6.** Put it all together into a notebook. In this notebook you need to write to show *statistical thinking*: explain your data; describe a hypothesis (and a null hypothesis), pick a statistical approach and justify it, test your model's assumptions, interpret the results.

R project presentation: 10%

- Details TBD! This is usually a front-of-the-class short presentation by each student, but there are more students in this class than ever before, so we will likely have to adjust our plans. We will talk about this in class.

Class summary and philosophy

This is an overview class

Most introductory statistics classes start with a brief overview of some concepts and then teach a few specific methods, such as linear regression, analysis of variance or contingency tables (e.g., chi-square tests). At the end, you *might* be told about more advanced techniques, such as general linear models, or 'AIC,' or Bayesian approaches. But you won't learn much about them, and in practice, when confronted with some data, you will be likely to fall back on what you do know, cramming almost any kind of data into, say, an ANOVA. (There is a famous saying that when you get your first hammer, everything looks like a nail.)

The reality of modern science in the era of big data is that most of you are likely to need much more sophisticated techniques right away (e.g., for your MS or PhD research). And you will probably apply these techniques using packages built by others. To give just one example, almost all phylogenetic studies employ tools like *BEAST* or *MrBayes* that, behind the scenes, are doing 'Bayesian' inference. If you do research in this field, you won't be expected to write a package yourself to analyze your data. But you *will* be expected to understand in general what a Bayesian analysis is, what the alternatives are, why it is appropriate in that context, what the strengths and weaknesses are, and so on. This course attempts to provide that perspective.

You will still learn how to do some specific things

Mainly what you will learn is how to work in R, the most widely used statistical environment in academia. This will include some actual statistical analyses, although which ones you do will be partly up to you. But it will also include the equally valuable skills of data import and manipulation, visualization, and so on.

This is a flipped class

Most classes will consist of no more than one half lecture/discussion, with the remainder being activities, generally in R. The lecture component can be reduced because the material is available as a set of interactive notebooks. **These must be read, and more importantly *thought about*, prior to each class!** Otherwise you will be lost, and do poorly. The lecture component will then be a *review* of the notebook material, answering questions, clearing up misunderstandings, etc. To 'encourage' this reading/thinking, and to discover the main sticking points in understanding, you are asked to submit at least one question about the material in advance of the class. (This will be done online: see the Canvas site.)

This year you will start out using of R in a sort of 'canned' way to accompany the theoretical material; you will run some simple, provided scripts that perform analyses so that we can examine and learn to understand the output. Roughly half-way through the semester, as you start to develop your class project, Dr. Garnier will spend more time instructing you in the R language itself, especially techniques

for data manipulation and visualization which, alongside the statistical analyses themselves comprise the toolbox any scientists needs to know.

Topic outline

This outline may be modified in any given semester depending on a number of factors: how fast we get through the material, whether there are any special topics people want to talk about, etc. Check Canvas for the latest schedule and notebook updates.

Reminder: Other than Week 1, notebooks should be studied thoroughly **before** that class under which they are listed!

Week 1 (Jan 18)

Notebook 1: *Models, Randomness, Probability and Mind.*
Class activity: *Installing running and navigating R and RStudio.*

Week 2 (Jan 25)

Notebook 2: *Probability Distributions: Characteristics and Fitting to Data*
Coding in R

Week 3 (Feb 1)

Notebook 3: *Fitting Models By Maximum Likelihood and Least Squares*
Coding in R

Week 4 (Feb 8)

Notebook 4: *Fitting Models By Maximum Likelihood and Least Squares*
Coding in R

Week 5 (Feb 15)

Notebook 5: *Statistical Inference Three Different Ways*
Coding in R

Week 6 (Feb 22)

Notebook 6: *Linear Models Are Not Just Straight Lines!*
Coding in R

Week 7 (Mar 1)

Notebook 7: *So Many Kinds of Errors!*

Coding in R

Week 8 (Mar 8)

MID-TERM EXAM!

SPRING BREAK!

Week 9 (Mar 22)

Notebook 8: *Designing for Power*
Coding in R

Week 10 (Mar 29)

Notebook 9: *Non-Parametric Statistics: Jackknifes, Bootstraps and a Feud*
Coding in R

Week 11 (Apr 5)

Notebook 9: *Non-Parametric Statistics: Jackknifes, Bootstraps and a Feud*
Coding in R

Week 12 (Apr 12)

Notebook: *Autocorrelation: Remove It or Model It, But Don't Ignore It*
Coding in R

Week 13 (Apr 19)

Notebook: *Bayesian Models in the Real World/additional topic by request*
Coding in R

Week 14 (Apr 26)

Project presentations

[Week 15 (May 3) — no classes. This is a Friday schedule.]