E&EB Guide to Building Data Science Skills @ Yale

(Compiled in May 2020 by Anusha Bishop with contributions from Matt Hack and Maya Juman)

1. Introduction

An understanding of statistics and data science is undoubtedly important for undergraduates interested in ecology and evolutionary biology research. However, currently the E&EB major at Yale does not have any explicit statistical requirements or built-in programs for developing data science skills. Instead, most resources for developing these skills must primarily be found externally, which can be challenging and confusing for many students. To try and address this issue, I have compiled a short guide to the resources that I have found during my time as an E&EB student. I have focused especially on learning to conduct data analysis in R, as this is becoming a fundamental skill to have in the field (however, other languages like Python are also very useful to know). This is by no means an all-encompassing guide, but hopefully this will provide a good jumping off point for E&EB majors.

There are many great formal resources at Yale (e.g. courses, the YCRC, etc.) that I will go over, but don't forget to take advantage of informal resources, especially other undergraduates, graduate students, post docs, advisors, etc. If you are interested in doing data analysis specific to some area (e.g. genomics), the best resource can often be talking to and learning from other people (e.g. through shared code, training recommendations, reading suggestions, etc.). Try not to be worried about asking “dumb” questions, everyone has to start somewhere and most people recognize that there can be a steep learning curve for data science (especially coding). That being said, while the start of learning these new skills can be challenging, many students find that it is very rewarding and helps them grow as independent researchers.

2. Selection of Practical Yale Courses

These are my personal course recommendations for E&EB majors looking to build up some statistics and data science skills *without having to take advanced math* (i.e. linear algebra). However, students can definitely benefit from taking more advanced math, especially if they are interested in fields like theoretical ecology. Additionally, I have left out some of the general statistics courses (i.e. Theory of Statistics, Probability Theory, Machine Learning) because they may not be as practically relevant to all E&EB majors. These are definitely still good courses to check out, and several are necessary if you are looking to complete the Certificate in Data Science.

Important Note: These are the courses I am aware of as of Spring 2020, but there is a lot of fluctuation. I would recommend keeping an eye out every semester for new courses. Relevant courses are not just limited to S&DS and can also be found in ENV (formerly F&ES), E&EB, CPSC, etc. Be sure to check out masters/graduate school courses as well, they are often totally accessible for undergraduates and can provide more practical learning (especially those taught in ENV (formerly F&ES) and E&EB). You can always talk to the professor if you are worried about the level of the class.

2.1. Introductory Statistics Courses

2.1.1. **S&DS 101-106 (Introduction to Statistics and Data Science):** My recommendation for most E&EB students with *no previous experience in statistics*, would be to start by taking S&DS 106. This is the intro stats course taught by Jonathan Reuning-Scherer who is a wonderful prof. I would advise taking 106 (the data analysis section) instead of 101 (the bio section), because currently 106 is the only section that has an integrated R component (however, the sections/professors fluctuate yearly so you should confirm this if you plan on enrolling). In this course, most people use minitab for at least the first half of the semester (before you split into sections), but I would recommend trying to use R the entire semester. Jonathan Reuning-Scherer provides instructions for both minitab and R in his lectures/notes and the ULA/TAs are more than happy to help people use R. In my opinion, if you feel comfortable doing some self-instruction in R, this is the best use of your time. However, if you don't feel comfortable doing this and want to learn R with some more guidance, you can follow up this course with S&DS 230 (see below).

2.1.2. **S&DS 220 (Introductory Statistics, Intensive):** This course is an intensive version of introductory statistics that uses R the entire semester (unlike the other introductory courses mentioned above). So far it has only been taught in the Spring by Joseph Chang.

2.2. R Courses (Beginner)

2.2.1. **S&DS 230 (Data Exploration and Analysis):** I recommend this course if you are looking to learn R and either (a) took one of the S&DS 101-106 courses OR (b) took some kind of statistics (i.e. AP Stats) in high school OR (c) have some basic statistical understanding (i.e. what is a standard deviation, normal distribution, etc.) and are willing to put in a little extra work to fill in any gaps if they come up. This course provides a comprehensive intro to R and some more intermediate statistics (PCA, ANOVA, etc.) for data analysis. It is very well taught and structured for practical learning.

2.2.2. **F&ES 720 (Introduction to R)**: This is an Introduction to R class taught in F&ES (now ENV) that could be a good thing to checkout. It does require some previous statistics and it is more applied than S&DS 230, since it's aimed at environmental science masters students. Normally, E&EB undergraduates can get spots in F&ES classes pretty easily, but this should be confirmed with the professor before enrolling. Note: F&ES switched to ENV in 2020, so this course may be listed in future terms as ENV 720.

2.3. R Courses (Intermediate)

2.3.1. **S&DS 363 (Multivariate Statistics for Social Sciences):** I strongly recommend this course if you are interested in developing your R and data analysis skills and you have (a) learned R in a previous course (e.g. S&DS 230) OR (b) learned R independently and feel comfortable with intro stats. This course is also taught by Jonathan Reuning-Scherer. Don't be thrown off by the course title, it is not just relevant for social sciences (many of the examples and datasets he uses are from environmental science) and the statistics material is surprisingly manageable (no linear algebra knowledge needed). My friends and I all agree that this is one of the lightest workload/least stressful classes we've ever taken. There are no exams, only a few psets, and everything (including homework) is done in groups (so take it with a friend if you can). In addition, for the homework and the final project you can analyze your own data. It is an incredibly useful course and you gain some really good data analysis skills.

2.3.2. **S&DS 674 (Applied Spatial Statistics):** This course provides a project based introduction to spatial statistics in R which could be very useful to E&EB students interested in applying spatial modelling in their research. It is taught by Timothy Gregoire, an F&ES professor with an extensive background in applying and developing statistical methods for studying natural resources and environmental phenomena. Gregoire is also slated to teach **ENV 753 (Regression Modeling of Ecological and**

Environmental Data) and **ENV 751 (Sampling Methodology and Practice)** in the coming year (2020-2021).

3. Other Yale Data Science Resources

3.1. Yale Center for Research and Computing (YCRC)

The YCRC tends to be underutilized by E&EB undergraduates and should not be overlooked. They offer extensive training opportunities which cover everything from general skills in R and Python to advanced skills in High Performance Computing and GIS. Their workshops are completely free and most are only a few hours. These training sessions can be especially valuable if you don't have time to commit to semester-long courses and you like to learn relatively independently.

3.2. Marx Science & Social Science Library

Like the YCRC, the Marx Science & Social Science Library (formerly known as the CSSSI) offers a lot of great (free) workshops, especially in introductory statistics and coding (R and Python). They also offer free research consultations which can be really helpful for figuring out how to properly apply statistics to your own personal projects.

4. Learning R Independently

In terms of independent learning material, I think to learn some of the basics of R a good place to start is R for Data Science. Some people also swear by swirl, which teaches you R interactively in the R console. To learn R syntax and how coding syntax works in general, I used the DataCamp app on my phone to practice when I had little pockets of time (on the bus, in between classes, etc). It is kind of like Duolingo for R and can be helpful if you have no previous coding experience in any language. Codecademy also has very similar resources to DataCamp.

Important Note: most people would probably agree that you learn R by using R. My biggest mistake while learning R was trying to do a ton of preemptive learning through reading and studying, when in reality nothing ever sticks until you actually have to apply it. I seriously think that the most important thing you can learn is how to effectively google your questions/problems and apply what you find.