STAT E-80: Basic Probability Using R

Course Syllabus: Preliminary Version 0.1

Spring 2020

Instructor: Theodore Hatch Whitfield, Sc.D. (call me Hatch)

Time: Wednesday, 5:40 PM – 7:40 PM

Location: 1 Story Street, Room 302

Start Date: Wednesday, January 29, 2020

E-mail: hatch.extension@gmail.com

Introduction

STAT E-80: Basic Probability Using R is an introductory course in probability theory that covers the fundamental concepts of the subject, along with many classic problems. The course has a number of unique features that distinguish it from other introductory probability courses:

- It's designed to teach the basic concepts of probability to students with a limited background in higher mathematics.
- The course is taught with an emphasis on computation, and students implement all probability concepts by writing executable programs in the R statistical programming language.
- Exact calculation and approximation using simulation are integrated throughout the course.
- The course places strong emphasis on the properties of the classical parametric probability distributions (e.g. binomial, geometric, normal, etc.).
- Students learn to estimate model parameters using data.
- Data visualization methods are used extensively.
- The course is available online.

Primary Learning Outcomes

At the end of the course, students will be able to:

- Use the probability theory to perform exact calculations to solve complex concrete problems.
- Develop probabilistic simulation models to verify their exact calculations.
- Visualize their work.

As a side benefit, students will also develop their skills in R, and programming in general, although the course is not primarily intended as a programming course.

Intended Audience

STAT S-80 is designed for beginning students in applied data science who want to develop their ability to use probability theory to solve concrete problems. The course provides a strong foundation for further work in statistics, machine learning, or data science.

What STAT S-80 is NOT

STAT S-80 does not cover certain topics:

- The course is not a rigorous mathematical development of the formal theory of probability, and there is little emphasis on proving theorems.
- The course is not a comprehensive introduction to the R programming language, and R concepts are introduced only when necessary to study probability.
- The course is not concerned with statistical analysis of data.

Prerequisites

Only high-school algebra and geometry is required for the course. Prior experience with calculus is helpful but not required. Similarly, no prior experience with programming is required.

Course Format: On-Campus and Online

The course is listed as "On-campus with an online option". What exactly does this mean?

- First, every Wednesday night, from 5:40 PM to 7:40 PM, from January 29, 2020, through May 6, 2020, I will be at 1 Story Street, Room 302, delivering a lecture. There will be one exception to this schedule: on March 18, we won't have any class at all, because that will be Spring Break week.
- During the lectures, there will be a live internet feed, and you can log in remotely and watch the class. We'll be using the Zoom in the Room webconferencing platform, so if you want to participate in the class you can.
- The lectures will be recorded, and posted on the course website within 24 hours. Thus, even if you can't be in class or watch remotely, you can still view everything as on-demand video.

If you're in the Boston Metro area, I would strongly encourage you to attend lectures if possible. Face-to-face interaction is still the best form of communication, and if you come to lecture then it's easier for me to stay in touch with you and for you to talk to me about any questions or concerns you might have. However, this is not strictly necessary, and even if you live right in Harvard Square, if your preferred course format is to sit in your living room in your pajamas and watch the lectures online while eating a bowl of cereal, that's fine too. There is no requirement

that you physically attend lectures, and there is no penalty for missing them. In addition to the recorded lectures, we will also be using the Piazza social media platform, which will allow for public discussion of course topics. So there will be many ways for you to interact with both the teaching staff and your fellow students, and I hope that you will use these extensively.

Assessments

The course assessments will consist of three components:

- 12 problem sets
- 1 midterm exam
- 1 comprehensive course project

The final grade is calculated as a weighted average of the three components: the overall problem set score is 20\%, the midterm exam is 30%, and the final exam is 50%. Grades are calculated on an absolute scale, with 93 and above receiving an A, 90 to 93 an A-, etc. No scaling or curving is used.

About Exams

Since many students in the class are distance learners, and will be unable to come to Harvard to sit for exams, both the midterm exams and the comprehensive final project will be administered remotely for all students. These exams will be released on Canvas, and you will download the exam, work out the problems, and then upload a PDF of the solutions in 1 file. The times and dates for the midterm exam and the final project are:

- The midterm will be released on Thursday, March 12, at 5:00 PM, and submission will close on Wednesday, March 18, at 5:00 PM.
- The final project will be released on Thursday, May 7, at 5:00 PM, and submission will close on Wednesday, May 13, at 5:00 PM.

Some courses have online exams that feature a "clock" i.e. once the exam is downloaded, you only have a limited time to complete it, typically a few hours. The exams for this course do not feature such a clock, and you can take as much time as you like to work on the exam, as long as you submit it by the final deadline. Also, the exams will not require a proctor, and you are expected to comply with all rules on academic integrity a.k.a. cheating.

Course Materials

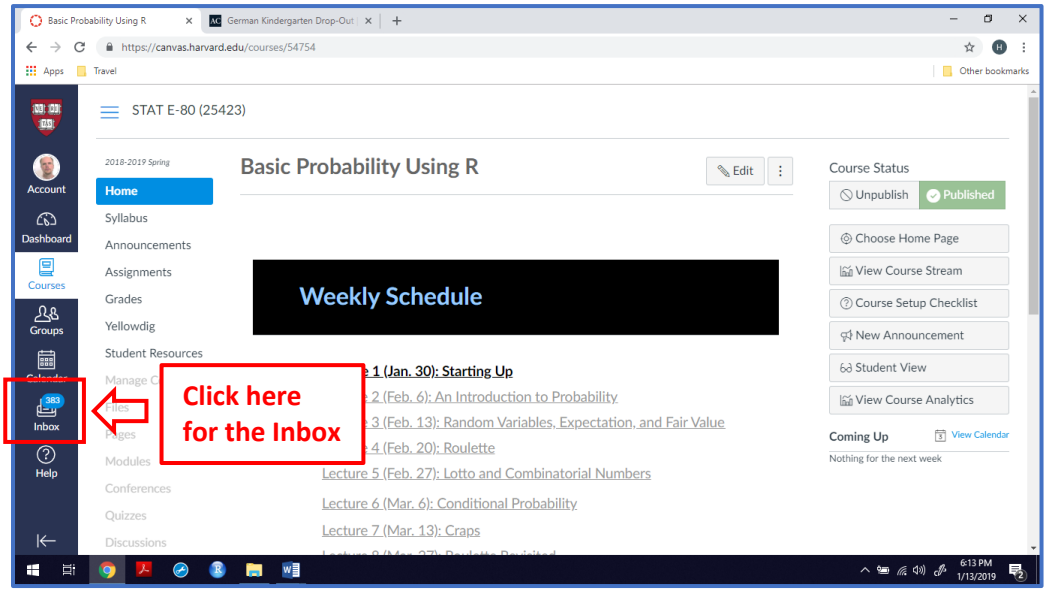
The readings, slides, and R notebooks for the course form the full set of course materials, and no additional text or reference is required. However, for students who would like an additional reference for programming in R, a recommended text is:

- *Learn R in 24 Hours*

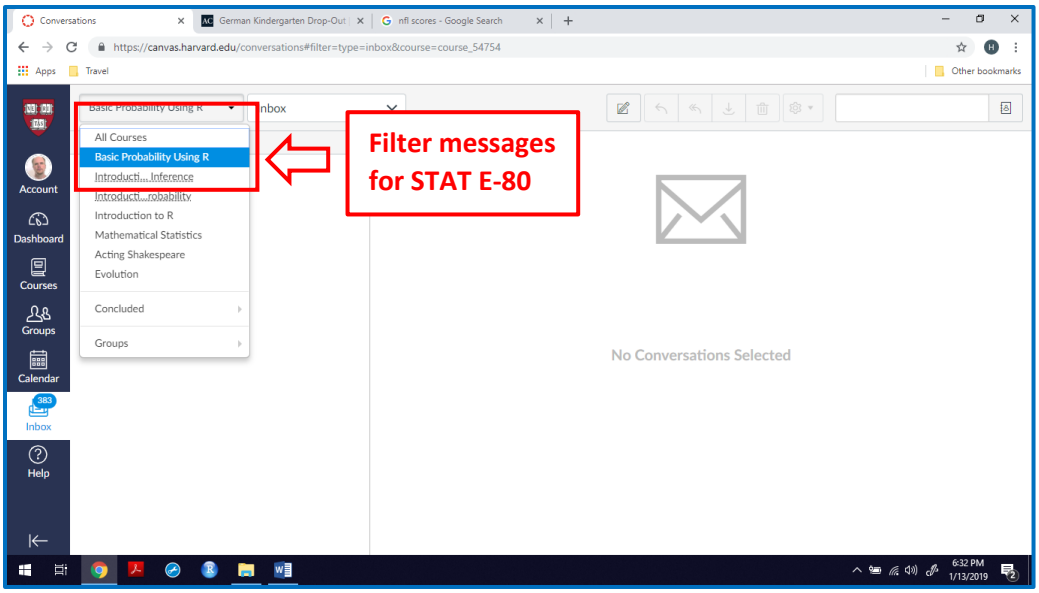
Communicating with the Teaching Staff

We always want to hear from you! But we also want to communicate with you in an orderly way. The Canvas system provides us with a dedicated email system for STAT E-80 called "Inbox", and this is the best method to get in touch with us and ask us questions. Unfortunately, many people are unaware of the Inbox, so let's take a moment to get familiar with it and learn how to use it.

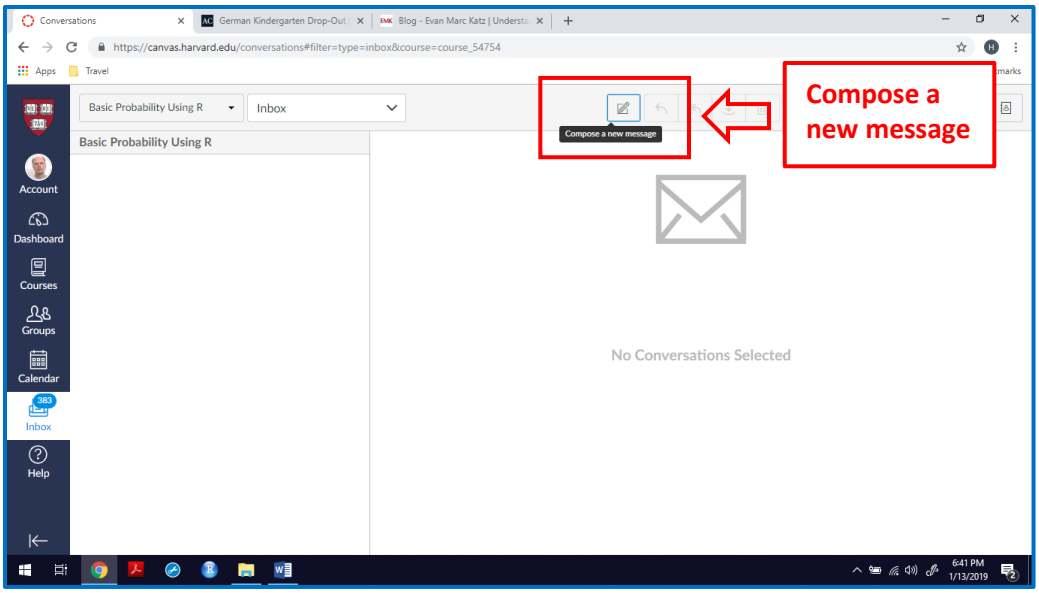
The first thing we need to do is to get to the Inbox page. You can do this from almost anywhere within Canvas, but for the sake of concreteness let's start from the STAT E-80 course home page. On the left-hand side you can see a vertical menu of navigation icons, and near the bottom there is an icon titled "Inbox":



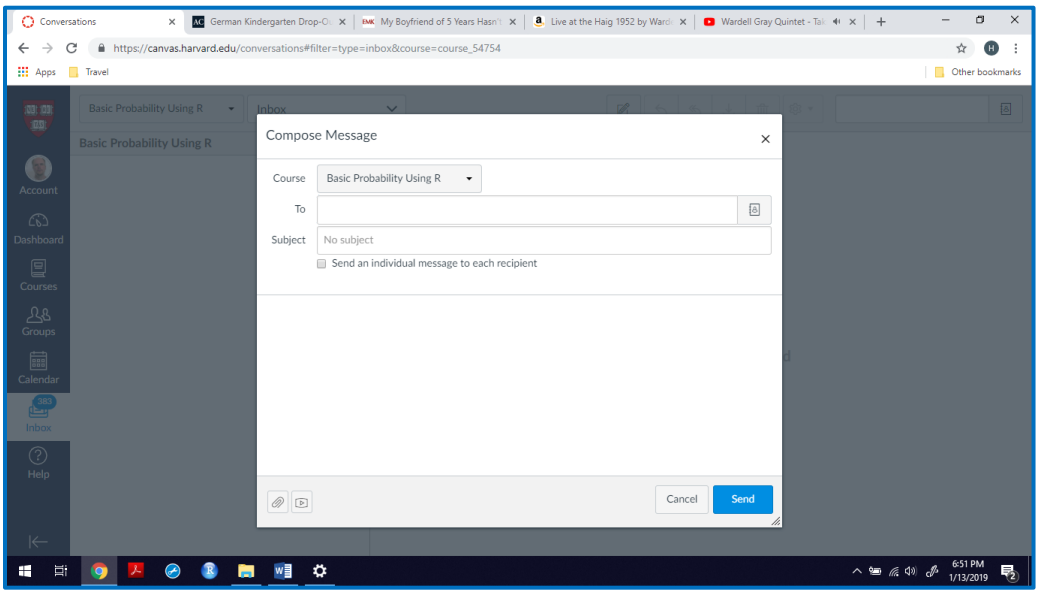
Once you get to the Inbox page, you'll see all your messages from all your Canvas courses. You can filter these messages so that you just see the mail for STAT E-80 by going to the upper-left hand corner where there is a drop-down menu that allows you to select an individual course for display:



To compose a new message, there is an icon at the top of the page with a graphic of a little pencil writing inside a square and the text “Compose a new message”:



When you click on this icon, you'll get a pop-up window for writing your message:



Using the Canvas Inbox is the best way to communicate with us, because it provides us with a great way of tracking all your messages. If you don't use this, your mail can easily get buried or lost.

Video Policy

This course is available online, and so we record video for the lectures and the TA sections. That means that you could potentially end up on camera, and thus there might be video of you. By signing up for the course, you are implicitly allowing us to make this available to other members of the class. The video will only be available to other members of the class, and it will be disabled after the semester ends, but during that time it will be accessible. If you don't want to be on camera, there will be a designated location in the class that will be a No Filming zone, and you won't end up on camera. However, other than this you can't opt out of video filming, and by enrolling in the course you are granting us permission to film you.

Disability and Accessibility Policy

Here's the official Harvard policy for students with disabilities:

> *The Extension School is committed to providing an accessible academic community. The Accessibility Office offers a variety of accommodations and services to students with documented disabilities. Please visit www.extension.harvard.edu/resources-policies/resources/disability-services-accessibility for more information.*

If this is an issue for you, talk or e-mail me personally, and we'll work it out.

Academic Integrity aka Cheating Policy

Here's the official Harvard policy on cheating and plagiarism:

> *You are responsible for understanding Harvard Extension School policies on academic integrity (www.extension.harvard.edu/resources-policies/student-conduct/academic-integrity) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting the wrong draft, or being overwhelmed with multiple demands are not acceptable excuses. There are no excuses for failure to uphold academic integrity. To support your learning about academic citation rules, please visit the Harvard Extension School Tips to Avoid Plagiarism (www.extension.harvard.edu/resources-policies/resources/tips-avoid-plagiarism), where you'll find links to the Harvard Guide to Using Sources and two free online 15-minute tutorials to test your knowledge of academic citation policy. The tutorials are anonymous open-learning tools.*

The rule for this course is very simple: all the work that you submit must be your own. If you have a conversation with someone or read something and that helps you to understand an issue better, that's great, but the actual thing that you give to us must be entirely your own work. Here are some things to consider:

- The Harvard rules are quite explicit, and you have to follow the rules.
- Cheating is spiritually and morally degrading. Doing the right thing is spiritually and morally uplifting.
- You deny yourself the sense of satisfaction from successfully meeting a difficult challenge.
- The exams are based on the homeworks, so if you cheat on the homeworks and don't really understand them, you'll have difficulty doing the exams.

Bottom line: don't cheat.

Schedule

Lecture 1: Getting Started (Wednesday, January 29)

We'll start with a quick review of official policies and course administration. Then we'll learn how to work with R and R notebooks, and practice with some simple graphics primitives. We'll finish with the celebrated Monty Hall game.

- Problem Set 1 assigned, due on Wednesday, February 5 at 5 PM.

Lecture 2: Basic Theory (Wednesday, February 5)

Basic concepts of elementary probability: the sample space, events, and the Kolmogorov axioms. Definition of a probability distribution. Independence of events. Sampling with and without replacement. Beginning counting. The Chevalier de Mere dice games. The complement trick and the "birthday" problem.

- Problem Set 1 due at 5 PM
- Problem Set 2 assigned, due on Wednesday, February 12 at 5 PM

Lecture 3: Random Variables (Wednesday, February 12)

Random variables, expectation, variance, cumulative and survival probabilities, transformations and sums of random variables. Discrete vs. continuous. Independence of random variables. Payout and profit for a game. Dice games. Elementary roulette.

- Problem Set 2 due at 5 PM
- Problem Set 3 assigned, due on Wednesday, February 19 at 5 PM

Lecture 4: Conditional Probability (Wednesday, February 19)

Conditional probability. The Law of Total Probability. The Law of Total Expectation. The Monty Hall problem revisited.

- Problem Set 3 due at 5 PM
- Problem Set 4 assigned, due on Wednesday, February 26 at 5 PM

Lecture 5: Bayes' Theorem (Wednesday, February 26)

Bayes' theorem. Epidemiological disease screening. Authorship problems.

- Problem Set 4 due at 5 PM
- Problem Set 5 assigned, due on Wednesday, March 4 at 5 PM

Lecture 6: Multivariate distributions (Wednesday, March 4)

Multivariable probability distributions. Joint, marginal, and conditional probability. Covariance. Two-player zero-sum games. Penalty kicks in soccer.

- Problem Set 5 Due
- Problem Set 6 assigned, due on Wednesday, March 11 at 5 PM

Lecture 7: Casino Games (Wednesday, March 11)

Roulette. The casino dice game "Craps".

- Problem Set 6 Due
- Midterm Exam assigned, due on Wednesday, March 18 at 5 PM

Lecture 8: The Uniform Distribution (Wednesday, March 25)

Properties of discrete and continuous uniform distributions. Introduction to parameter estimation. The German tank problem.

- Problem Set 7 assigned, due on Wednesday, April 1 at 5 PM

Lecture 9: The Bernoulli, Binomial, and Poisson Distributions (Wednesday, April 1)

Bernoulli random variables. The binomial distribution as a sum of independent Bernoulli random variables. The mathematics of the binomial coefficient. The Poisson distribution as the limiting distribution of a rare event across a large population.

- Problem Set 7 Due
- Problem Set 8 assigned, due on Wednesday, April 8 at 5 PM

Lecture 10: The Geometric Distribution and Survival Curves (Wednesday, April 8)

The geometric distribution. Times to events. Social Security Administration life tables. Analysis of the dice game Craps.

- Problem Set 8 Due
- Problem Set 9 assigned, due on Wednesday, April 15 at 5 PM

Lecture 11: The Hypergeometric Distribution (Wednesday, April 15)

Sampling without replacement. Theory of the hypergeometric distribution. Analysis of the casino game Keno.

- Problem Set 9 Due
- Problem Set 10 assigned, due on Wednesday, April 22 at 5 PM

Lecture 12: The Exponential and Gamma Distributions (Wednesday, April 22)

Continuous distributions. Exponential distribution. The gamma distribution as a sum of independent exponential random variables. Quantiles. Parameter estimation.

- Problem Set 10 Due
- Problem Set 11 assigned, due on Wednesday, April 29 at 5 PM

Lecture 13: The Normal Distribution and the Central Limit Theorem (Wednesday, April 29)

The normal distribution. Motivation as the sum of large numbers of independent random variables. The Central Limit Theorem.

- Problem Set 11 Due
- Problem Set 12 assigned, due on Wednesday, May 6 at 5 PM

Lecture 14: The Bivariate Normal Distribution (Wednesday, May 6)

The bivariate normal distribution. Comprehensive review of course.

- Problem Set 12 Due
- Comprehensive Course Project assigned, due on Wednesday, May 13 at 5 PM