**BA 386T: Probability and Statistics**

Instructor of record: John C. Butler

Email: butlerjc@mccombs.utexas.edu

Teaching Assistant: Ming Lu

Email: mingmichelle@163.com

**Background and objectives.** This course presents a non-technical introduction to probability and statistics. The main goals are: 1) to give you some practical tools for data analysis in R; and 2) to prepare you for further education in a field where data and statistics are used routinely. The emphasis will be on conveying the basic principles of probability, statistics, and data science, without digging much into formal statistical theory. The philosophy here isn’t exactly “no math.” Rather, it’s more like “minimum viable math”: the least amount of formal mathematics necessary to convey the ideas you need to be an informed practitioner of basic data science.

**Course materials.** There’s nothing to buy. Your main points of reference will be:

* “[Data Science in R: A Gentle Introduction](https://bookdown.org/jgscott/DSGI)” referred to as “DSGI” in the syllabus and assigned readings.
* “Introduction to Probability Lecture Notes,” by Dimitri P. Bertsekas and John N. Tsitsiklis. This is posted as a PDF file on the course Canvas page and linked where appropriate in the course readings.
* [Statistical Inference for Data Science](https://leanpub.com/LittleInferenceBook/read), by Brian Caffo. This is free to read online, or you can buy a physical copy on a “name your own price” model (where $0 is a valid price).

**Software.** We will use R and RStudio for data analysis. They are free, and you will need both. For details on how to install them and get them up and running for the first time, see the first chapter of “[Data Science in R: A Gentle Introduction](https://bookdown.org/jgscott/DSGI).”

**Course format.** This is an asynchronous online course with no formal in-person component. You are free to complete the course requirements at your own pace. You will interact with all course materials through our course’s Canvas site. The course is structure around a set of Modules, which must be completed sequentially. (That is, you have to finish one Module before Canvas will unlock the next Module in the sequence.) A single module covers a single unit in the Course Topics outline below and has three components.

* **Read**: reading materials for that unit. Most, albeit not all, of these are drawn from “[Data Science in R: A Gentle Introduction](https://bookdown.org/jgscott/DSGI).” Some reading material is conceptual in nature, while other material takes the form of software walkthroughs. These walkthroughs are inherently “hands-on,” where you are expected to follow along with some data analyses and reproduce the results in your own RStudio session. See the sections in the reading on [“How you’ll get feedback”](https://bookdown.org/jgscott/DSGI/getting-started-in-r.html#how-youll-get-feedback) and [“My advice”](https://bookdown.org/jgscott/DSGI/getting-started-in-r.html#my-advice) for my suggestions on how best to use the readings to learn. The basic idea is simple, though: if you treat the readings as a passive experience, rather than something that requires your simultaneous active participation in R, you are unlikely to build the skills necessary to answer the quiz questions.
* **Watch**: video lectures for that Module, typically around 10-15 minutes in length each (with some shorter and others a bit longer). You should watch these in the order listed, because they build upon each other sequentially.
* **Quiz**: the quiz (or quizzes) for that Module. These quizzes form the basis for your course grade. (See the section on **Grading**, below.) **You must submit all the quizzes in a Module in order to unlock the next Module.**

**Course Administrative Notes**

* This course is a custom program offered for admitted Master of Science students in the McCombs School of Business in partnership with University Extension. Please direct all administrative and registration questions to MSPORegistrar@mccombs.utexas.edu.
* Please do not contact University Extension with any course related questions or reference their calendars for course registration information.
* This course is self-paced and follows a custom calendar. The first day students may begin the course is **January 9, 2023**. The last possible day to complete the course is **May 26, 2023**.
* If you complete the course prior to the last class day, please note that once a final grade is assigned, it is not possible to change the grade or drop the course.
* All content related questions should be addressed with your TA; Professor Butler is responsible for administrative issues.
* This course never meets in person, has no scheduled lectures.

**Grading.** Each module in Canvas will require that you take one or more quizzes before you move on to the next module. Your course grade is based entirely on these quizzes. Each quiz has roughly 5 to 10 questions covering a related set of topics from that Module. Some questions are conceptual in nature and are drawn directly from the readings and video lectures. Other questions require that you download a data set and conduct an analysis in RStudio, which provides the basis for your answer.

You are allowed two attempts at each quiz. Once you have submitted your first attempt, Canvas will show you which questions you’ve answered correctly and which you haven’t. This will give you a chance to revisit those questions, refer back to the course materials, learn from your mistakes, and make a second submission. Canvas will record your higher of the two attempts, so there is no possible penalty for trying again.

Your final course grade is calculated as the number of quiz questions you answer correctly, divided by the total number of quiz questions in the course. This percentage is calculated *across the entire set of quiz questions for the course*. The implication is that each *quiz question* (and not each *quiz*) carries the same weight in determining your final grade, so that a quiz with 10 question contributes more to your grade than a quiz with 5 questions. Said another way, we do *not* compute a percentage grade for each quiz and then average those percentages. So, for example, if there are 197 quiz questions over the entire course, and you answer 185 of them correctly, your grade would be 186/197 = 0.944, or 94.4%. (If it helps, you can think of the quizzes as analogous to one long final exam, distributed in little chunks over the entire length of the course.)

The grade cutoff values below provide the basis for translating your overall course average into a letter grade. **Course grades will not be rounded numerically; there will be no exceptions to this policy.** While we reserve the right to lower these cutoff values (i.e., make them more generous) at our sole discretion, we will not raise them (i.e., make them less generous).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Grade** | **A** | **A-** | **B+** | **B** | **B-** | **C+** | **C** | **C-** | **D** | **F** |
| **cutoff** | **94%** | **90%** | **87%** | **84%** | **80%** | **77%** | **70%** | **65%** | **60%** | **<60%** |

**Course topics outline**

**Unit 1: Data exploration**

*Getting started with data* ([DSGI](https://bookdown.org/jgscott/DSGI/) Lessons 1-2)

* Data frames, cases, and variables.
* Samples vs. populations.
* The unit of analysis.
* Case study: getting data into R

*Introduction to probability* ([DSGI](https://bookdown.org/jgscott/DSGI/) Lesson 3, Bertsekas Ch 1))

* What is probability?
* The basic rules of probability
* Probabilities estimated from data.
* Case study: music streaming data
* Independence and conditional independence.
* Simpson’s paradox
* The rule of total probability
* Bayes’ rule
* *Data visualization* ([DSGI](https://bookdown.org/jgscott/DSGI/) Lesson 4)
* The data vis Hall of Shame
* The grammar of graphics
* Five key plots
* Case study: Olympic medalists over time
* *Data summaries* ([DSGI](https://bookdown.org/jgscott/DSGI/) Lesson 5)
* Describing the center and shape of a data distribution
* Describing variation
* Extremes, quantiles, outliers, and z-scores
* *Data wrangling* ([DSGI](https://bookdown.org/jgscott/DSGI/) Lesson 6)
* Complex summaries: two concrete examples
* Six key data verbs
* Case study: EPA gas mileage
* Case study: NYC flights data
* *Fitting equations to data* ([DSGI](https://bookdown.org/jgscott/DSGI/) Lesson 7)
* What is a model?
* What are models good for?
* Case study: Austin restaurants
* Exponential models
* Power laws
* Case study: price elasticity for milk

**Unit 2: Statistical inference**

*Foundations for inference*. ([DSGI](https://bookdown.org/jgscott/DSGI/) Lesson 17.1-17.2, Bertsekas 2.1-2.2, 2.4)

* Introduction to random variables
* Expected value and variance
* Making sense of expected value
* The Poisson distribution
* The binomial distribution
* The normal distribution

*Statistical uncertainty.* ([DSGI](https://bookdown.org/jgscott/DSGI/) Lesson 8)

* Where uncertainty comes from
* Measuring uncertainty: variance versus bias
* Quantifying sampling variance by Monte Carlo simulation

*The bootstrap* ([DSGI](https://bookdown.org/jgscott/DSGI/) Lesson 9)

* The bootstrap
* Coding sessions: bootstrapping simple summaries (1 and 2)
* Bootstrapping a difference of means
* Bootstrapping a difference of proportions
* Bootstrapping a regression model
* Does the bootstrap work?

*p-values.* ([DSGI](https://bookdown.org/jgscott/DSGI/) Lesson 10)

* Case study 1: the New England Patriots. (Sidebar: a more involved simulation)
* The four steps of hypothesis testing.
* Interpreting p-values.
* Case study 2: a cancer cluster?

*Large-sample inference.* ([DSGI](https://bookdown.org/jgscott/DSGI/) Lesson 11, Bertsekas Ch 7, [Little Inference](https://leanpub.com/LittleInferenceBook/read) Ch 7)

* Intro to large-sample inference
* De Moivre’s Equation.
* The Central Limit Theorem.
* CLT case study
* Large-sample confidence intervals for a population mean.
* Beyond de Moivre’s equation
* Large-sample inference case study
* The t distribution: a correction for small samples.

**Unit 3: Regression models**

*Regression with categorical predictors* ([DSGI](https://bookdown.org/jgscott/DSGI/) Lesson 14)

* Baseline/offset form
* Models with one dummy variable
* Models with multiple dummy variables
* Interactions among dummy variables
* Diagnosing interactions
* Introduction to ANOVA: the analysis of variance

*Regression with multiple numerical predictors* ([DSGI](https://bookdown.org/jgscott/DSGI/) Lesson 15)

* Numerical and grouping variables together
* Correlated predictors and causal confusion: partial vs. overall relationships
* Same slope vs. different slopes: interactions revisited
* Interaction vs. correlation
* Partial slopes: regression with multiple numerical predictors
* Case study: spending more to get less?
* Summarizing regression models
* ANOVA revisited
* Building regression models for identification
* Case study: houses in Saratoga, NY
* Statistical vs. practical significance
* Confounders vs. colliders