Data Analysis and Visualization

CS555 B1

Tom Goulding MS, PhD
tlg@bu.edu
Room: CAS 315
Tuesday 6:00 – 8:45 PM

Office hours: By appointment via ZOOM
By appointment via ZOOM
Office : 808 Commonwealth Avenue, Room 254, Boston, MA02215
Phone: 978-772-5648

Course Description

This course provides an overview of the statistical tools most commonly used to process, analyze, and visualize data. Basic topics in statistics will be reviewed and discussed. They include measures of central tendency, probability, Test of hypothesis, statistical inference, 1 and 2 sample tests of means and proportions, simple linear regression, multiple regression, logistic regression, analysis of variance, and regression diagnostics.

The purpose of this course will be to explore these topics as our primary mission. Our secondary purpose is to focus on understanding how to use and interpret output from stat software as well as how to visualize results. We will primarily utilize R, an open source and free app with capabilities similar to Matlab.

Teams will be created the first day of class. Teams will work collaboratively throughout the semester in class and out of class. Homework problems, projects and classroom discussions will require extensive collaboration among team mates. Examinations are solo endeavors.

In each topic area, the methodology, including underlying assumptions and the mechanics of how it all works along with appropriate interpretation of the results, are discussed by teams in class. Teams will be heavily engaged in developing and presenting statistical concepts and solving problems in the context of real world examples.

About every two weeks a major topic discussion is led by the faculty member. During the first half of each 2 week period, the teams will engage in extensive collaboration resulting in an understanding and

application of various visualization concepts and tools. Problems are then assigned and the teams present the following week their solutions using the analytic tools of there choosing, but primarily R.

This collaborative problem solving approach to teaching & learning has an auspicious history.

The Pedagogy:
The professor is a disciple and academic descendant of R.L. Moore the famous University of Texas Mathematician who was a practitioner of the Socratic method. The professor thus guides student and engineering development project through questions which activate inquiry, relentless experimentation and ultimately success.

For more on R. L. Moore see: . http://legacyrlmoore.org/reference/burton_jones.html

Learning Objectives

By successfully completing this course you will be able to:

- Appreciate the science of statistics and the scope of its potential applications
- Summarize and present data in meaningful ways
- Select the appropriate statistical analysis depending on the research question at hand
- Form testable hypotheses that can be evaluated using common statistical analyses
- Understand and verify the underlying assumptions of a particular analysis
- Effectively and clearly communicate results from analyses performed to others
- Conduct, present, and interpret common statistical analyses using R, and other tools.

Prerequisites

Some background in statistics or CS546 (Quantitative Methods for Information Systems) and CS544 (Foundations of Analytics).

Books

You may find the web based tutorials and youtube videos sufficient to meet your needs.

However, the books below could also be used as reference material to help support you in your assignments. These supplemental texts may be of use to you as reference text if you continue to use R in the future. Both of these books can be purchased from Barnes and Noble at Boston University. The professor can recommend others. Feel free to ask.

Chang, W. (2013). *R graphics cookbook.* Sebastopol, CA: O'Reilly. ISBN 9781449316952.

Teetor, P. (2011). *R cookbook.* Sebastopol, CA: O'Reilly. ISBN 9780596809157.

Class Policies

1) **Attendance & Absences** – Full attendance and participation is expected. If there is a reason to miss a session, advanced notice through email should be sent to the lecturer.

If professional commitments, military deployments or an important personal matter (eg baby arrival) should develop do not hesitate to inform the instructor who will work with you to find ways of accommodation.

2) **Assignment Completion & Late Work** – All assignments should be submitted on time. If there is a delay, the student must be in touch with the instructor.
3) **Academic Conduct Code** –Cheating and plagiarism cannot be tolerated in any Metropolitan College course. Please take the time to review the Student Academic Conduct Code: http://www.bu.edu/met/metropolitan_college_people/student/resources/conduct/cod e.html.

NOTE: [Collaboration is essential and in fact required in this course.]

Grading Criteria

- Presentations and demonstrations.
  There will be approximately 5 presentations and demonstrations which are focused on applying theory learned in the week's modules to a set of data and analyzing that data in R. 5 bi weekly team submissions should be a single Microsoft Word or Powerpoint. The code, if any, used to generate your results should be appended to the end of your bi-weekly team presentations.

  In lieu of a final examination a term project in the health care field will be offered to each team.

The final grade for this course will be based on the following:

Deliverable	Weight
Bi Weekly presentation	50.00%
Mid Term Exam	25.00%
Term Project	25.00%

NOTE: Acquire a simple (non statistical calculator) calculator for use on examinations. This must be approved for use by the professor. *Smart phone calculators will not be allowed.*

Study Guide: A proposed schedule is a preliminary guideline: It will be adapted and changed as the background and progress of the students becomes evident.

Lecture 1 Introduction to the science of statistics part 1

- Fundamental Elements of Statistics
- Qualitative and Quantitative Data Summaries
- Measures of Central Tendency

Lecture 2 Introduction to the science of statistics part 2

- Normal distribution
- Sampling
- The Central Limit Theorem
-

Lecture 3 Confidence intervals and hypothesis tests part 1

- Statistical Inference
- Stating Hypotheses
- Test Statistics and p-Values
- Evaluating Hypotheses

Lecture 4 Confidence intervals and hypothesis tests part 2

- “Significance Test "Recipe"
- Significance Tests and Confidence Intervals
- Inference about a Population Mean
- Two-Sample Problems

Lecture 5 Understanding the association between two continuous or quantitative factors part 1

- Scatterplots
- Correlation

Lecture 6 Understanding the association between two continuous or quantitative factors part 2

- Simple Linear Regression
- F-test for Simple Linear Regression
- t-test for Simple Linear Regression

Lecture 7 Regression diagnostics

- Residual Plots
- Outliers and Influence Points
- Assumptions of least-square regression

Lecture 8 Multiple linear regression

- Equation of multiple linear regression
- Interpretation of multiple linear regression
- F-test for Multiple Linear Regression
- t-tests in Multiple Linear Regression
- Cautions about Regression

Lecture 9 Analysis of Variance (ANOVA) part 1

- One-Way Analysis of Variance
- F-test for ANOVA
- Evaluating Group Differences
- Type I and Type II Errors

Lecture 10 Analysis of Vriance (ANOVA) part 2

- Issues with Multiple Comparisons
- Assumptions of Analysis of Variance
- Relationship between One-Way Analysis of Variance and Regression
- One-Way Analysis of Covariance
- Two-Way Analysis of Variance
- Two-Way Analysis of Covariance

Lecture 11 Analysis for proportions part 1

- One-Sample Tests for Proportions
- Significance Tests for a Proportion
- Confidence Intervals for a Proportion

Lecture 12 Analysis for proportions part 1

- Two-Sample Tests for Proportions
- Confidence Intervals for Differences in Proportions
- Significance Tests for Differences in Proportions
- Effect Measures
- Logistic Regression
- Multiple Logistic Regression
- Area under the ROC

- Lecture 13 Review

Instructor Bio:

INDUSTRY: Principle Engineer to FORTUNE 100 Senior-VP/GM:

Guided $80/year systems business to $300M/year while achieving industry leading profitability.

Professor led ground breaking new technologies, including the first digital circuit switched networks for the US public telco network. The first fault tolerant, fully redundant tech control systems for the Cheyenne Mountain USA Air Defense Network, the VISA financial network and the Canadian power grid. His teams also developed the first fiber optic multiplexers, routers, hubs and switches. His individual contributions include developing the protype target acquisition and tracking algorithms and software for the 1st guided cannon launched munition. (Copperhead - TS clearance). (Motorola, Siemens, Sanmina, Martin-Lockheed)

Professor also has years of experience in medical device development for eldercare communities as well as community medicine research focused on parasitic and opportunistic disease characteristics in rural indigent Appalachian populations. Extensive experience building collaborative teams of physicians, engineers and academics

ACADEMIC LEADERSHIP His academic leadership roles included 14 years as Professor and Chairman of Math, CS. & Networking. He led rather modest computer science programs into a period of high enrollment growth to become a leading project-based computer science programs. He specializes in guiding virtual novices to success developing complex software.

GRANTS & RESEARCH: Dr. Goulding has received research grants from Coleman Foundation, Microsoft Research, Electronic Arts, Microsoft XNA group and NSF. He is an Electronic Arts Scholar with over 40 peer reviewed research papers on a wide range of topics.

TEACHING: 30+ years mentoring students often part-time and frequently online while serving in industry.

EDUCATION

University of Florida	PhD, MS	Theoretical Mathematics

Boston University Library Link

As Boston University students you have full access to the BU Library—even if you do not live in Boston. From any computer, you can gain access to anything at the library that is electronically formatted. To connect to the library use the link http://www.bu.edu/library. You may use the library's content whether you are connected through your online course or not, by confirming your status as a BU community member using your Kerberos password.

Once in the library system, you can use the links under "Resources" and "Collections" to find databases, eJournals, and eBooks, as well as search the library by subject. Go to http://www.bu.edu/library/research/collections to access eBooks and eJournals directly. If you have questions about library resources, go to http://www.bu.edu/library/help/ask-a-librarian to email the library or use the live chat feature.

To locate course eReserves, go to http://www.bu.edu/library/services/reserves.

Please note that you are not to post attachments of the required or other readings in the water cooler or other areas of the course, as it is an infringement on copyright laws and department policy. All students have access to the library system and will need to develop research skills that include how to find articles through library systems and databases.

Academic Conduct Policy

For the full text of the academic conduct code, please go to http://www.bu.edu/met/for-students/met-policies-procedures-resources/academic-conduct-code/.

A Definition of Plagiarism

"The academic counterpart of the bank embezzler and of the manufacturer who mislabels products is the plagiarist: the student or scholar who leads readers to believe that what they are reading is the original work of the writer when it is not. If it could be assumed that the distinction between plagiarism and honest use of sources is perfectly clear in everyone's mind, there would be no need for the explanation that follows; merely the warning with which this definition concludes would be enough. But it is apparent that sometimes people of goodwill draw the suspicion of guilt upon themselves (and, indeed, are guilty) simply because they are not aware of the illegitimacy of certain

kinds of "borrowing" and of the procedures for correct identification of materials other than those gained through independent research and reflection."

"The spectrum is a wide one. At one end there is a word-for-word copying of another's writing without enclosing the copied passage in quotation marks and identifying it in a footnote, both of which are necessary. (This includes, of course, the copying of all or any part of another student's paper.) It hardly seems possible that anyone of college age or more could do that without clear intent to deceive. At the other end there is the almost casual slipping in of a particularly apt term which one has come across in reading and which so aptly expresses one's opinion that one is tempted to make it personal property."

"Between these poles there are degrees and degrees, but they may be roughly placed in two groups. Close to outright and blatant deceit-but more the result, perhaps, of laziness than of bad intent-is the patching together of random jottings made in the course of reading, generally without careful identification of their source, and then woven into the text, so that the result is a mosaic of other people's ideas and words, the writer's sole contribution being the cement to hold the pieces together. Indicative of more effort and, for that reason, somewhat closer to honest, though still dishonest, is the paraphrase, and abbreviated (and often skillfully prepared) restatement of someone else's analysis or conclusion, without acknowledgment that another person's text has been the basis for the recapitulation."

The paragraphs above are from H. Martin and R. Ohmann, *The Logic and Rhetoric of Exposition, Revised Edition.* Copyright 1963, Holt, Rinehart and Winston.