COMP 4441 Introduction to Probability and Statistics for Data Science

Course Overview

This course is a foundational course in the MS in Data Science program, introducing core methods and theories in statistical analysis. The course prepares students to carry out independent research and interpretation of data analyses to address research questions. The course provides an introduction to the use of R and RStudio for statistical programming.

Objectives

Students in this course will

- Apply a range of data visualization methods
- Develop models for data, and apply statistical techniques to assess the validity of the models
    - Use parametric and nonparametric methods for examining a single sample and two-sample means
    - Apply linear regression
- Derive sound theoretical footing for the methods, where practical, to guide them as they
    - Check data requirements
    - Apply method diagnostics
    - Interpret results
- Report results of analyses of real-world data using course methods
- Use basics of R, a widely used programming language for statistical analysis

Textbooks and Materials

Crawley, M. J. *Statistics: An introduction using R* (2nd ed.),
ISBN: 978-1-118-94109-6. (An e-book or hard copy will work.)
Text website

This text provides a practical introduction to statistical analyses implemented in the R programming language. It does assume a familiarity with data analysis needs. The lectures will provide orientation to these needs, as well as supplementation of the somewhat terse theoretical discussions.

Other References

Grinstead, C. M., & Snell, C. L. *A introduction to probability*.
https://www.math.dartmouth.edu/~prob/prob/prob.pdf

Grolemund, G., & Wickham, H. *R for data science*.
https://r4ds.had.co.nz/

Topics covered include

- Fundamentals of probability spaces
- Basic data visualization
- Summary statistics
- Concept of hypothesis testing
- One- and two-sample tests: Wilcoxon, *t*, and *z*
- Goodness of fit tests: Chi-squared and Fisher's exact
- Simple regression

Grading

Assignment/Assessment	**Points**	**Weight on Final Grade**
Problem Sets (x8)	400 (50 points each)	30%
Midterm Exam	100	15%
Final Project	215	25%
Final Exam	100	15%
Participation	50 (5 points each week)	15%

Grading Scale

- A is 93-100
- A- 90-92.99
- B+ 86-89.99
- B 83-85.99
- B- 80-82.99
- C+ 76-79.99
- C 73-75.99
- C- 70-72.99
- D+ 66-69.99
- D 63-65.99
- D- 60-62.99
- F < 60

Assignment and Assessment Information

Problem Sets

Problem sets will typically consist of data analyses. Students should turn in a .doc, .docx, or .pdf discussion of the results; the .R or .Rmd file of the code used to obtain the results; and the .RData workspace in which the results were calculated. Students are strongly encouraged to use R Markdown to generate the discussion of the results.

Each problem set will be disseminated at the beginning of the week prior to the week it is due. For example, the problem set due prior to the Week 5 Live Session will be disseminated to students at the beginning of Week 4 and will largely cover the content that students are learning in the Week 4 asynchronous coursework and live session.

One goal of the problem sets in this class is to encourage you to innovate with the techniques you have learned. This innovation can take the form of figuring out how a method applies to a new situation or how a principle can be generalized to create a new method.

This level of creativity can be a very enjoyable aspect of the practice of data science. It can also be very hard to pull off on a tight deadline. Please try to start the problem set early to give yourself time to step away and come back with new ideas. If you are stuck as the deadline approaches, consulting with a colleague or mentor (instructor) may be an option.

Midterm Exam
Students work within an RStudio template to create examples of basic probability computations, basic R functionality, and one-sample tests of center. The completed exam will provide a reference for the statistical concepts and the programming tools.

Final Project
In the final project, students use at least one method not covered in class as part of an analysis of data as well as a full explanation of the method. The final project consists of

- **A project plan** identifying the method and dataset to be used
- **A paper** explaining the process and R code for the analysis
- **A presentation** explaining detailed analysis of a dataset

More details about the final project, including a grading rubric, are available in the Toolbox in the Course LMS.

Final Exam
Students work within an RStudio template to create examples of the application of two-sample tests of center, multiple-sample tests for categorical data, and linear regression. The examples will include tests that assumptions of methods are met and interpretation of results. The completed exam will provide a reference for the methods and their R implementations.

Participation
Your participation grade each week is determined by

- Completion of the asynchronous formative assessments, for example, knowledge check questions and file uploads
- Submission of one news item relating to developments in data science or applications of data science prior to each week's live session, as applicable
- Attendance and active participation in each week's live session

Weekly Schedule

Please complete readings prior to beginning this week's asynchronous content for the indicated week.

Week 1—Introduction to R and Inference

- Readings :
    - Crawley, M. J., *Statistics: An introduction using R* (2nd ed.), Appendix.

Week 2—Introduction to Probability

- Problem Set 1 due prior to Week 2 Live Session.

Week 3—Parameter Estimation

- Readings:
    - Crawley, M. J., *Statistics: An introduction using R* (2nd ed.), Chapter 2, Selecting parts of a data frame: Subscripts.

    - Crawley, M. J., *Statistics: An introduction using R* (2nd ed.), Chapter 2, Sorting.
- Submit a recent article or news story relating to data science as part of asynchronous coursework 24 hours before Week 3 Live Session.
- Problem Set 2 due prior to Week 3 Live Session.

Week 4—Centrality and Variance

- Readings:
    - Crawley, M. J., *Statistics: An introduction using R* (2nd ed.), Chapter 3.
    - Crawley, M. J., *Statistics: An introduction using R* (2nd ed.), Chapter 4.
- Submit a recent article or news story relating to data science as part of asynchronous coursework 24 hours before Week 4 Live Session.
- Problem Set 3 due prior to beginning of Week 4 Live Session.

Week 5—Expectation

- Readings:
    - Crawley, M. J., *Statistics: An introduction using R* (2nd ed.), Chapter 5, Data summary in the one-sample case.
- Submit a recent article or news story relating to data science as part of asynchronous coursework 24 hours before Week 5 Live Session.
- Problem Set 4 due prior to beginning of Week 5 Live Session.

Week 6—Parametric Tests of Center: Single Sample

- Readings:
    - Crawley, M. J., *Statistics: An introduction using R* (2nd ed.), Chapter 5, Calculations using *z* of the normal distribution through student's *t* distribution.
- Submit a recent article or news story relating to data science as part of asynchronous coursework 24 hours before Week 6 Live Session.
- Midterm exam due 24 hours before Week 6 Live Session.
- Problem Set 5 due prior to beginning of Week 6 Live Session.
- Final project proposal due prior to beginning of Week 6 Live Session

Week 7—Tests of Center: Nonparametric and Two Sample

- Readings:
    - Crawley, M. J., *Statistics: An introduction using R* (2nd ed.), Chapter 6, Comparing two variances through tests on paired samples.
- Submit a recent article or news story relating to data science as part of asynchronous coursework 24 hours before Week 7 Live Session.
- Problem Set 6 due prior to beginning of Week 7 Live Session.

Week 8—Categorical Data Methods, Covariance and Correlation

- Readings:
    - Crawley, M. J., *Statistics: An introduction using R* (2nd ed.), Chapter 6, The binomial test through end of chapter.

- Submit a recent article or news story relating to data science as part of asynchronous coursework 24 hours before Week 8 Live Session.
- Problem Set 7 due prior to beginning of Week 8 Live Session.

Week 9—Linear Regression

- Readings:
    - Crawley, M. J., *Statistics: An introduction using R* (2nd ed.), Chapter 7, Linear regression through Measuring degree of fit.
- Submit a news news story showing the relevance of data science in current events as part of asynchronous coursework 24 hours before Week 9 Live Session.
- Problem Set 8 due prior to beginning of Week 9 Live Session.

Week 10—Topics in Regression

- Readings:
    - Crawley, M. J., *Statistics: An introduction using R* (2nd ed.), Chapter 7, Model checking through end of chapter.
- Final exam due prior to Week 10 Live Session.
- Deliver final project during Week 10 Live Session.

Collaboration and Academic Honesty

When you turn in work in this course, you are implicitly agreeing that you have followed the rules for collaboration set forth for that assignment. In general

- The midterm, final project, and final exam must be your own work.
- For problem sets, you may consult with the instructor. You may work individually or in a pair. You may consult with other students, but you should credit them. Do not do web searches for solutions.

Students will abide by the honor code.

Technology

You will need a good internet connection and a laptop that meets DU specifications. (See http://www.du.edu/uts/laptops/specs.html.)

The programming assignments will be completed using the R programming language, which may be downloaded from http://cran.r-project.org/.

The IDE RStudio works well with R and makes use of R Markdown particularly simple. RStudio may be downloaded from https://www.rstudio.com/products/rstudio/download/ after you have downloaded R. The free RStudio desktop version is suitable for this course.