IST 772 - (3 credits)

Catalog Title: Quantitative Reasoning for Data Science
Time/Day: Monday, 2:15–5:00 PM
Location: Hinds Hall 018

Instructor: Kevin Crowston
Phone: +1 315–443–1676   Email: crowston@syr.edu

Note that this syllabus is subject to change.

Course Catalog Description:

Multiple strategies for inferential reasoning about quantitative data. Methods for connecting data provenance to substantive analytical conclusions.

Course summary and objectives:

Quantitative data analysis stands at the center of the field of data science. Yet most people who make use of analytical results – and a surprising number who produce those results – have little idea of how the nature of the data and the analysis impacts their interpretation. Uncertainty is an intrinsic and inescapable characteristic of data. If we fail to understand uncertainty when working with our data, at best we will make poorer decisions than necessary and at worst we will make catastrophic errors based on mistaken assumptions. **Statistical inference is the judgment process by which we make sense of uncertainty in data, and this course focuses on establishing a thoughtful and thorough understanding of statistical inference.**

This course will help you understand contemporary methods of statistical inference regardless of the specific type of analysis you are undertaking. The primary goal of this course is for you to learn the techniques and concepts that facilitate drawing sensible conclusions from samples of quantitative data. The course has three major goals focusing on knowledge, skills, and practice. By the end of this course you will be able to:

Demonstrate knowledge of contemporary inferential statistical concepts (from the perspective of two contemporary philosophies) and data analysis strategies by making sensible choices about: • How data collection, the data themselves, and the analysis processes relate to the kinds of inferences that can be drawn • What kinds of analysis will be feasible and developing the skill of planning data collection and measurement to facilitate appropriate analysis
Practice effective data science analytics: • Preparing data for analysis, including screening data, dealing with missing data, doing data transformations • Testing assumptions that data must meet for analyses and inferences to be reasonable • Interpreting data analysis results and outputs and communicating them to others using language that accurately describes uncertainty

• Leaving a documentation/provenance trail for other analysts to follow and reproduce your work
Demonstrate competence and/or mastery of the skills needed for use of a popular statistics and data management platform to conduct sound and reproducible analyses including: • Installing R and R-studio, and creating readable code to conduct analyses • Exploring the limitations of existing data sets and how their provenance influences what analyses to perform and what inferences to draw • Choosing appropriate R procedures and configuring the relevantoperational parameters

Note for Students Who Have Taken SCM651

This course covers several of the same statistical procedures as another course in the data science program, SCM651. It does so from a different perspective, however, and with distinctive goals. Both courses cover the basics of R, bivariate Pearson correlation, Analysis of Variance, Least Squares Multiple Regression, and Binomial Logistic Regression. The focus of this course is on the foundations of statistical inference with a special focus on the conceptual and procedural connections among traditional frequentist inference methods and Bayesian inference. **This course focuses on the principles of correct interpretation of statistical evidence**. Knowledge from this course can be applied to any new inferential technique that you encounter. Students who (will) have professional responsibilities that involve communicating about statistical results in writing, verbally, and graphically will sharpen those skills in this course.

Textbook and Readings:

One textbook is required for this course: Stanton (2017), *Reasoning with Data: An Introduction to Traditional and Bayesian Statistics Using R* (abbreviated below as RwD; ISBN-13: 978-1462530267; ISBN-10: 1462530265). The paperback version of this textbook is available from the bookstore. Paperback, hard back, and electronic versions are all available on Amazon and other online book sellers:

https://www.amazon.com/Reasoning-Data-Introduction-Traditional-Statistics/dp/1462530265

There are four known errata in the textbook.

1. On page 18, the book states that the Poisson distribution is good for modeling arrival delays. In fact, the exponential distribution is better for modeling arrival *delays*, while the Poisson is better at modeling the number of arrival *events* within fixed time periods (note that a Poisson variable is always a natural number, i.e., a count).
2. On page 36 in Exercise 8, the reference to a previous exercise should be to Exercise 7 rather than Exercise 6.
3. On page 224, the caption for Figure 10.5 is incorrect. It should read, "Boxplots of age and income variables, grouped by vote."
4. On page 272, problem 2 refers to the Blackmore data set as being in the nlme package, while it is actually from the carData package.

In addition to the textbook, I will also provide supplemental readings for you based on your interests and background.

Note that we will use R and R-Studio extensively throughout this class. R-Studio is the preferred platform for developing code in this class. You must have access to a laptop computer on which these open source packages can be installed (they work on Mac, Windows, and Linux). Bring a laptop to each class with R and R-Studio ready to run. You will also find it advantageous to have R and R-Studio available when you are reading the textbook.

Class Sessions (Note: Flipped Format)

This class is a "flipped" class. The idea behind a flipped class is that, given the right content, you can prepare yourself quite extensively before we meet in person allowing the in-person time to be spent on activities that benefit from interaction rather than lecture (defined as "a process by which the contents of professor's notes are transferred to the student's notebook without passing through the mind of either").

To prepare yourself for the in-person session each week, you must read the designated chapter in *Reasoning with Data* and then review the videos that are linked in Blackboard for the given week. Take notes while you read and watch and make note of points where you have questions. Each weekly set of videos ends with a brief, two question assessment meant to reinforce and solidify your understanding of the material for that week.

The in-person session will meet once a week for 165 minutes. Rather than lecture, the class will generally consist of discussion of the materials you already reviewed and in-class exercises to reinforce your ability to apply what you've learned. Certain classes will include "low stakes" practice tests (that is, you get credit just for trying the work). Once your questions are cleared up and we've gone through the exercises, you will have the remaining class time to work on that week's problem set or practice exam, while being able to ask questions or reopen discussion of unclear points, with the goal of submitting the assignment by the end of class or shortly after.

During the class, you are expected actively participate by answering the instructor's questions and interacting with your colleagues. Most sessions will include exercises where you will work with one or more colleagues on a problem assigned by the instructor. To be able to take part in these exercises, you **must** bring to class a laptop with Zoom and RStudio installed. All of the exercise materials needed for participation in the live session are stored in Blackboard. People learn best in an environment of mutual respect, so please remember that your colleagues in these exercises may be at a different stage of knowledge or skill than you. Be supportive and assist others with their learning during the class.

In summary, in a traditional class, in-person time is spent on an introduction to content, leaving students to work out how to apply it on their own. In a flipped class, students introduce themselves to the material on their own time, and in-person time is spent practicing applying the knowledge, with support from the professor and other students. Done correctly, a flipped class is much more effective for student learning. However, it requires discipline on your part to prepare. If you have not done the required review before hand, you will get little out of the in-person session and will not be prepared to do the weekly assignment or the exams.

Weekly Workflow

The textbook, *Reasoning with Data*, serves as the main knowledge base for the class. Each week will be based upon one chapter of the book. We will go through the whole book in sequence from Chapter 1 to Chapter 12. Preparing for class by reading the book, running the R code examples, and taking notes is the major key to success in this course. Here is a recommended weekly workflow to help you stay on pace:

- **At least three days prior to class**, read the assigned chapter in *Reasoning with Data.* Each chapter runs about 15 pages and the whole book was designed to be highly readable and accessible, even to those with a limited background in math and/or limited prior statistical knowledge. Depending upon how quickly you read, each chapter should take you no more than about 90 minutes for a thorough consideration. You will find it advantageous to run the code examples shown in the chapters while you read. I have provided Jupyter notebooks to make this easy. Take written notes while you read!
- **One or two days before class:** Review all of the video lectures provided for that week. The lectures are provided in bite sized segments so that you can take notes and run code in between segments. After reviewing the last video in each week's list, there is a two-question, short answer test on Blackboard for you to complete. Each of these is timed, limited to 60 minutes, and must be completed in one sitting. You may use your notes and the book as reference material.
- **The class sessions occur once per week**. Participation in class is obligatory: If you have a compelling reason to miss one class, make sure to inform me in advance. If your obligations or circumstances will cause you to miss more than one class you should take this course in a future semester. One necessity for your success in class includes bringing a laptop with the capability to run R-Studio and Zoom. Each in-class session will contain coding and interpretation activities based on the exercises at the end of each chapter. At the end of the class session you will be submitting the code file you developed as you worked on the exercises.
- **Following the in-person session**, you will have 72 hours to complete the homework assignment for that week (though as noted above, there will usually be time to work on these during the in-person session), i.e., the homework is due 5pm Friday. The homework assignment generally comprises problems drawn from the exercises at the conclusion of each chapter of Reasoning with Data. Most of the homework problems require the use of R-studio. If you tackle these problems immediately after the live session (or even during!) you should be able to complete each homework in somewhere between one and three hours. After the in-person session, you can ask me questions by email, but you must allow 24 hours for a response, so getting started early is paramount.

Course Calendar:

Week	**Reading**	**Goals**
1	Introduction & Chapter 1	Topic: **Getting started and Statistical Vocabulary** Personal introductions; Get R and R-Studio Installed; Try R and R-Studio; Initial learning and skills assessment; Read Appendix A and Appendix B if you are not yet an R user;
2	Chapter 2	Topic: **Basic Probability**; Explore descriptive statistics and distributions; View data sets in R; Contingency table exercise; read data into R and diagnose.
3	Chapter 3	Topic: **Sampling Distributions**; Principles of sampling; sampling over the long run; sampling distributions of means: generating sampling distributions.
4	Chapter 4	Topic: **Statistical Inference Part I**; Practice exam; Inductive reasoning, the logic of inference; comparing means of independent samples; point estimates and interval estimates; confidence intervals.
5	Chapter 5	Topic: **Statistical Inference Part II**; Practice exam; Bayesian thinking; Bayes' rule; Markov chain, Monte Carlo; posterior distribution of mean differences; null hypothesis significance test.
6	Chapter 6	Topic: **ANOVA & Experimental Groups**; Analysis of variance; between and within groups variance; the F-distribution; Bayes factors; experimental data collection and analysis.
7	No Reading	**Mid-term exam**: The mid-term will occur in the classroom during class seven. The exam may ask you to use R to produce some results and/or to write up an interpretation of some provided results. The mid-term may also contain knowledge and skill questions.
8	Chapter 7	Topic: **Measures of association**; Association, covariance, and correlation; cross products and Pearson product moment, inferential reasoning about the correlation coefficient; categorical associations; correlation data collection and analysis; chi-square data collection and analysis.
9	Chapter 8	Topic: **Multiple Regression/Linear Prediction**; Criteria and predictors; point clouds; least-squares criterion; measures of model quality; multicollinearity; Bayesian and frequentist hypothesis testing.
10	Chapter 9	Topic: **Interactions in ANOVA and Regression**; Interactions in ANOVA; two-way experimental designs and more complex designs; degrees of freedom for ANOVA interactions; interpreting Bayesian output; interactions in multiple regression; centering; graphic interactions.

11	Chapter 10	Topic: **Categorical Analysis**; The logistic curve; generalized linear model and link functions; log odds and odds; measures of model quality; Bayesian estimation of logistic regression; in class categorical prediction exercise.
12	Chapter 11	Topic: **Time Series Analysis**; Non-independence of observations; repeated measures ANOVA; time-series analysis; change point analysis.
13	Chapter 12	Topic: **Dealing with Too Many Variables;** Principal component analysis; scale reliability.
14	No reading	**Review session**. Come prepared with questions or topics you'd like to discuss as you prepare for the final exam.

Student Assessment:

This course provides knowledge and practice in quantitative data analysis and in communicating statistical results accurately and without bias. Note that the primary goal of these assessments is to enhance your learning. If you work hard, jump in with both feet, and do all of the assigned work, it will be a success, and I can assure you that you will obtain a fair result at the end. Here is the breakdown of points for the course:

11 homework assignments, 5 points each, with lowest score dropped = 50 points
5 practice tests, 2 points each = 10 points (weeks 4, 5, 8, 9, and 10)
1 midterm based on the first 5 weeks of the course = 15 points
1 final exam comprising an analysis of real data sets = 25 points

And the grading table:

A = 95-100; A- = 90-94.99; B+ = 85-89.99; B = 80-84.99; B- = 75-79.99, C+ = 70-74.99; C = 65-69.99; C- = 60-64.99; >60 = F

Keep in mind these important grading rules: I will not round-up your final points count to the next highest integer; there will not be any extra credit assignments; if you submit late, it will only be accepted and graded at my discretion and a late penalty will be applied. I will not offer any make up exams – so you must be present on exam days. My evaluations of your work are based on nearly three decades of grading statistics homework and exams: I will always be happy to fix any bookkeeping errors that may occur, but my judgments on the correctness of your work are final.

Preparing Your Homework for Submission

Prepare your responses to the assigned questions in the provided RStudio notebook, then knit the notebook to create a PDF to submit. Double check that the PDF correctly captures the work you did. Submit your homework as a PDF file. Name your file HWX_Lastname.pdf, substituting the week number and your own last name.

The homework intentionally models the kinds of actions and language you may use as a data scientist, so it is critical to format and present the homework in a professional manner. Likewise, a key goal of this course pertains to accurate and unbiased communication of statistical results: Please write your interpretations in complete, grammatical sentences.

The main purpose of the homework is to practice the skills you have learned that week and crystallize the knowledge that you have gained. As such, homework should be a solo activity, so that you can prove to yourself and the instructor your capacity to accomplish the work independently. To the extent that you do collaborate with someone else – including seeking coaching, feedback, suggestions, or code examples – you must acknowledge your sources at the top of the homework file. This is the "give credit where credit is due" principle and it is paramount for data scientists. The same idea holds with respect to consulting outside resources, such as the R-Bloggers website. This point is important enough to repeat: Do not cut and paste anything without proper citation and quotation marks! Based on these principles, your homework should begin with a statement like one of these:

- Homework 1 by Fred Flintstone: I produced the material below with no assistance.

or

- Homework 1 by Fred Flintstone: I consulted with Barney Rubble about how to tackle these problems, but we each wrote our code and text independently.

or

- Homework 1 by Fred Flintstone: I consulted StackOverflow.com for information about how to write this code. Line 43-45 of this code file were copied from https://stackoverflow.com/questions/bayesian-inference-in-R

Any variation on these statements is reasonable as long as it is forthright. If you submit a homework that reports collaboration with another student, your instructor may have new advice, guidance, or suggestions for you to enhance your learning.

Solutions for the problem sets will not be posted. However, each student will once in the semester prepare and post a video explaining how they answered one or two of the problems, which can be consulted as a reference.

Examinations

A midterm exam will be given in class in class 7.

For the final exam, you will receive a custom dataset for the final exam that is unique to you. You will undertake a set of analyses described in the exam document, saving (and submitting to Blackboard) all of your output for later use and then create and submit to Blackboard a written report of your results prior to a deadline during the final exam period.