IST 772 - (3 credits)
Catalog Title: Statistical Methods in Information Science and Technology
Location: Online
Instructor: Kevin Crowston; Phone: 315-443-1676 Email: crowston@syr.edu

Course Catalog Description:

Classical statistical procedures used in information transfer research. Emphasis on underlying rationale for each procedure and on criteria for selecting procedures in a given research situation.

Course summary and objectives:

Quantitative data analytics stands in 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 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 relevant operational 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 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 two known errata in the textbook. On page 224, the caption for Figure 10.5 is incorrect. It should read, "Boxplots of age and income variables, grouped by vote." On page 272, problem 2 refers to the Blackmore data set as being in the nlme package, whereas it is actually from the carData package.

Your instructor may 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 homework and exam responses. It is preferable to have access to a laptop computer so these open source packages can be installed (they work on Mac, Windowns, and Linux) on your own machine. Alternatively, the iSchool has an experimental R-Studio server that can be accessed through any browser with your SU NetID and password at https://rstudio.ischool.syr.edu. Have a laptop or desktop available during each class with R and R-Studio ready to run. Having two screens is a definite plus when doing work for the class and particularly for participating in the weekly synchronous sessions. You will also find it advantageous to have R and R-Studio available when you are reading the textbook and when you are reviewing the asynchronous material.

Synchronous Sessions

The synchronous sessions are conducted using a web-based video conferencing application. You will need a high speed Internet connection and a computer that is equipped with a video camera, speaker, and microphone. Many students prefer to use headphones or earbuds instead of speakers because they prevent feedback with the microphone. By far the highest quality experience comes from using a gaming headset that has headphones and a built in microphone. If you use your computer's speakers instead of headphones or earbuds, you may need to stay on mute much of the time to avoid putting room noise or feedback into the session.

To avoid distracting your fellow students and provide a good image for the video-conference, you should position your computer's camera 18-24 inches from your face with a wall, rather than a window or door, as the background. Although it is theoretically possible to participate successfully in the session when family members, friends, roommates, or pets are in the room with you, it is paramount that you avoid having any distracting activity in your room during the session. Likewise, although emergencies do occur, in general you should have your phone and other devices set to "do not disturb" for the duration of the session.

Even if you are on mute during some portions of the session, you are expected to be actively participating by answering the instructor's questions and interacting with your colleagues. Most synchronous sessions will include "breakout rooms" where you will work with one or more colleagues on a problem assigned by the instructor. Everyone learns best in an environment of mutual respect, so please remember that your colleagues may be at a different stage of knowledge or skill than you. Be supportive and assist others with their learning during the synchronous sessions.

Please join each session 5-10 minutes before the official start time so that you can test and adjust your camera, microphone, and speakers/headphones. Once a session has started, the instructor cannot pause if you have technical difficulties.

Student Assessment:

This course provides knowledge and practice in quantitative data analysis and particularly in communicating statistical results accurately and without bias. There are two major components of student assessment: your weekly homework and two examinations. Additionally, your active and constructive involvement in the live class sessions will benefit you and your classmates. Some discussions and activities will relate to your understanding of concepts, while others will pertain to the practice of quantitative analysis.

Your section instructor will assign work for you to complete in class, generally in small groups. Certain classes will contain "low stakes" practice tests (that is, you get credit just for trying the work). Generally, the instructor will assign exercises in class to expand and practice your skills and knowledge. In these cases, you will be evaluated based on the extent to which you engage with the problems presented, and the extent to which you improve your own understanding of quantitative methods. 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:

10 homeworks, 5 points each = 50 points
5 practice tests, 2 points each = 10 points (weeks 4, 5, 8, 9, and 10)
1 midterm based on the first 4-5 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

No rounding; no extra credit; no late submissions; no make up assignments; no shirt, no shoes: no service. Seriously, if you submit any of your work late, it will only be accepted and graded at the instructor's discretion and a late penalty may be applied.

Class structure:

Each week comprises a couple of hours of prerecorded (asynchronous) segments and a 90 minute live session. Here is a recommended workflow for you:

- **At least three days prior to the live session**, and before you begin reviewing the asynchronous material, 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 no prior statistical knowledge. Some may even find the book entertaining. 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 as you read.

- **One or two days prior to the live session**, review the asynchronous material. This material contains pre-recorded segments that account for about an hour, plus activities, questions, problems and exercises that may take about another hour. This material reinforces what you read in the book and provides an opportunity to test yourself on your knowledge.
- **The live (synchronous) sessions occur once per week**, require a high speed Internet connection and a device that will run the Zoom platform. Participation in the live sessions is obligatory: If you have a compelling reason to miss one class, make sure to inform the instructor in advance. If your obligations or circumstances will cause you to miss more than one class you should consider taking this course in a future quarter. One necessity for your success in the live session includes the capability for you to run R-Studio. Each live session has one or more breakout sessions where you will complete an exercise or activity with the help of R-studio and R code you create or modify.
- **Following the live session, you will have 72 hours to complete the homework** assignment for that week. 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 before!) you should be able to complete each homework in somewhere between one and three hours. You can ask questions to the section instructor 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; The homework for week one is exercises 1, 3, and 4 on page 20.
2	Chapter 2	Topic: **Basic Probability**; Explore descriptive statistics and distributions; View data sets in R; Initial skills assessment returned and discussed; Contigency table exercise; read data into R and diagnose. The homework for week two is exercises 1 and 2 on page 35, as well as problems 6, 7, and 8 on page 36.
3	Chapter 3	Topic: **Sampling Distributions**; Principles of sampling; sampling over the long run; sampling distributions of means: generating sampling distributions. The homework for week three is exercises 2 through 7 on pages 50 and 51.
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. The homework for Week 4 is exercises 7-10 on page 66.
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. The homework for week five is exercises 6 through 10 on pages 86 and 87.
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. The homework for week 6 is exercises 1-7 on pages 117 and 118.
MT	Mid-term	Complete mid-term test within 24 hours of its release at the end of live class 6. Typically, the exam will ask you to use R to produce some results and you will write up an interpretation of those results. The midterm may also contain knowledge and skill questions.
7	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. The homework for week seven is exercises 3,4, 8, 9, and 10 on pages 155 and 156.
8	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. The homework for week 8 is exercises 1-8 on pages 181-182.
9	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. The homework for week 9 is exercises 1, 5, 6 and 7 on page 234.
10	Chapter 11	Topic: **Time Series Analysis**; Non-independence of observations; repeated measures ANOVA; time-series analysis; change point analysis; final "quick paper" on the selection of statistical methods for different data situations. The homework for week 10 is exercises 2, 5, 6, 7, and 8 on pages 272 and 273.
11		Final examination: For the final exam, you will receive a custom dataset that is unique to you. You will have a set of analytical

		challenges and will be responsible for writing up the results. The exam will become available after Live Session 10 and will be due no later than seven days later. Example: If live session 10 ends at 7 PM EST on a Thursday, then the exam will be due no later than the following Thursday at 6:59 PM EST.

Preparing Your Homework for Submission

Prepare your responses to the assigned questions with a word processor, cutting and pasting output from R-studio as appropriate. If you are providing R output, you must include the snippet of code that created that output. When R provides output that is formatted as a table, you may find it helpful to switch to a monospace font such as Courier. As an alternative to preparing your homework in a word processor, you may use Markdown if you are familiar with it. This course does not provide instruction on how to use Markdown. If you wish to do this, make sure to provide the Rmd file as well as the output file in PDF.

Submit your homework as PDF file. Name your file HWX_Lastname.pdf, substituting the week number and your own last name. Submitting a PDF ensures that the way you submit the homework is the way it will appear when graded.

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 quoteation 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. **Each homework is due 72 hours following the completion of the live session for that week.** So, for example, if a live session meets on Mondays from 6:00-7:30 PM EST, the homework would be due on Thursday no later than 7:30 PM EST.

Academic Integrity
The Academic Integrity Policy holds students accountable for the integrity of the work they submit. Students should be familiar with the university's policy and know that it is their responsibility to learn about course-specific expectations, as well as about university policy. The university policy governs appropriate citation and use of sources, the integrity of work submitted in exams and assignments, and the veracity of signatures on attendance sheets and other verification of participation in class activities. The policy also prohibits students from submitting the same written work in more than one class without receiving written authorization in advance from both instructors.

As a supplemental matter closely related to academic integrity, all materials produced for this class are copyrighted, either by the course author, Jeffrey Stanton, by your section instructor, or by Syracuse University. Please respect the integrity of the course. Do not post materials from this course online. Do not post answers to homework questions or otherwise share your answers with other students. Similarly, do not post answers to exam questions or otherwise share your answers with other students.

Disability-Related Accommodations
Our community values diversity and seeks to promote meaningful access to educational opportunities for all students. I am committed to your success and to supporting Section 504 of the Rehabilitation Act of 1973 as amended and the Americans with Disabilities Act (1990). This means that in general no individual who is otherwise qualified shall be excluded from participation in, be denied benefits of, or be subjected to discrimination under any program or activity, solely by reason of having a disability. You are also welcome to contact me privately to discuss your academic needs. Syracuse University has policies, procedures, and a resource center to assist with disability related matters.

Educational Use of Student Work

I intend to use academic work that you complete this semester for educational purposes in this course during this semester. Your registration and continued enrollment constitute your permission. I also intend to use academic work that you complete this semester in subsequent semesters for educational purposes. Before using your work for that purpose, I will either get your written permission or render the work anonymous by removing all your personal identification.