

QR 260: Applied Data Analysis and Statistical Inference

M/Th 1:30-2:40 pm, W1 2:15-3:25, PNE 339

Fall 2015 Syllabus

Instructor: Cassandra Pattanayak, cpattanayak@wellesley.edu

Office Hours: Tue 2-3:30 every week and Wed 1-2 if no lab, unless announced otherwise; or by appointment. My office, Clapp 238, is in the back of the main floor of Clapp Library, off of Brackett Reading Room, behind the Sanger Room.

My experience is that the concepts in this course are best explained in person. I strongly encourage you to take advantage of office hours, and I look forward to discussing your questions and reactions to the material.

Teaching Assistant: Anne Corbett '16, acorbett@wellesley.edu. Office hours TBA.

Description: This is an intermediate statistics course focused on fundamentals of statistical inference and applied data analysis tools. Emphasis on thinking statistically, evaluating assumptions, and developing practical skills for real-life applications to fields such as medicine, politics, education, and beyond. Topics include t-tests and non-parametric alternatives, multiple comparisons, analysis of variance, linear regression, model refinement, missing data, and causal inference. Students can expect to gain a working knowledge of the statistical software R, which will be used for data analysis and for simulations designed to strengthen conceptual understanding. This course is offered through Wellesley's Quantitative Analysis Institute.

Goals: After this course, you should be able to:

- Evaluate the strengths, weaknesses, and appropriateness of a variety of statistical techniques
- Given a data set: state hypotheses, explore the data using statistical software, identify and apply appropriate analysis methods, and assess assumptions
- Communicate statistical results graphically and in writing
- Handle common practical challenges of data analysis, including missing data, multiple comparisons, and data cleaning
- Use the statistical software R

Prerequisites: Any Quantitative Reasoning Overlay course. To earn credit toward economics major, must have taken Econ 103. To earn credit toward psychology major, must have taken Psyc 205.

Distribution: Mathematical modeling.

Note: This course can be counted as a 200-level course toward the major or minor in economics or psychology. Students who earned a QAI Certificate are not eligible for this course. Students who participated in the 2015 QAI Wintersession Pilot are eligible.

Google Group and Directories: I will use the google group for announcements, and you should feel free to use it to communicate with each other. You can also email me personally. I will make every attempt to answer your emails within 24 hours. It is not always possible for me to answer last-minute questions just before a deadline. You will be invited to access a shared google drive directory, where I will post course material.

Computing: This course involves learning the statistical software R. R is popular among statisticians and other researchers because it is free, downloadable, open source, field-neutral, and powerful. No matter where you are after college, R will always be available. Instructions for downloading R and other R resources will be posted.

No previous experience with R is necessary. However, you should expect that it will take time to familiarize yourself with R, and I expect to answer lots of R questions! Computing questions (like conceptual questions) are usually easier to answer in person than by email, so plan ahead and attend office hours.

Blended Learning: As part of Wellesley's Blended Learning Initiative, we will make use of online materials throughout this course. You will need to create an account at edge.edx.org – instructions will be sent out.

Textbook: The textbook for this course is *The Statistical Sleuth: A Course in Methods of Data Analysis*, Ramsey and Schafer. A few notes:

- Do not overpay! E-books, used hard copies, and rentals are available for \$40-\$100. Check out amazon and the publisher's website (cengagebrain.com). We'll cover approximately 12 chapters, and the second half of the book may be helpful as a future resource.
- You do not need the CD that comes with the book. All the data sets are free online.
- It doesn't matter whether you have the 2nd edition or the 3rd edition. The chapters are almost identical. Some problems were changed, but when that's an issue, I will reproduce the whole problem on the assignment. The 2nd edition costs less.
- A copy of the 2nd edition will be on reserve at the Science Library.

Other references:

OpenIntro Statistics, Diez, Barr, and Cetinkaya-Rundel

<http://www.openintro.org/stat/textbook.php>

This is a free, downloadable introductory statistics textbook that may be helpful if we refer to statistical concepts that you'd like to brush up on or have never seen.

Practical Regression and Anova using R, Faraway

<http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf>

This is a higher-level free, downloadable textbook that includes R code.

Causal Inference in Statistics and Social Science, by Imbens and Rubin. This is the textbook I use for another course (Econ 309/QR 309/Soc 319). For this course, I will post a few relevant chapters.

Course Requirements and Grading:

35% - Problem sets (approx weekly, typically due 11:59 pm on Thursdays)

15% - Midterm (self-timed take-home, due Mon, Oct 19)

15% - Project (in small groups, due Thursday, Dec 3 – everyone must be present that day)

25% - Final Exam (self-scheduled)

5% - Mini-assignments

5% - Class participation (I expect you to be present and engaged during class meetings)

Policies:

- Problem sets should be submitted electronically by 11:59 pm on the due date, unless specified otherwise. Submission instructions will be posted. Typically, you will be required to submit your R code in addition to your write-up.
- I will deduct 25% of the possible points from graded problem sets for each day of lateness: if the deadline is 11:59 pm on Thu, then assignments received by 11:59 pm Fri can score no higher than 75%; assignments received by 11:59 pm Sat can score no higher than 50%; assignments received by 11:59 pm Sun can score no higher than 25%; and assignments received after 11:59 pm Mon will receive no credit.
- The lowest problem set score will be dropped when your grade is calculated.
- Extensions will be granted only in exceptional circumstances, such as serious illness or a family emergency, or to accommodate special circumstances as described below. **If you think you may miss a deadline because of travel, interviews, or other scheduling conflicts, submit the assignment early.**
- You may bring two two-sided pages of notes (on 8.5" x 11" paper) and a calculator to the midterm and four two-sided pages of notes (on 8.5" x 11" paper) and a calculator to the final. No other resources are allowed.
- Mini-Assignments: Twice a week, I will circulate a google form with a short question, to be answered electronically before the next class meeting. The purpose is for me to be up-to-date on your thoughts and vice versa. I may be checking your understanding of the previous class, collecting examples for discussion, asking for updates on your projects, etc. The mini-assignments will be graded pass/fail. You will not receive credit for late mini-assignments, but two mini-assignments will be dropped.

Labs: Please bring a laptop – you can borrow one from the library if needed.

Honor Code and Collaboration Policy: The Honor Code will be strictly enforced. Sources must be cited in written assignments. Unless specified otherwise, all assignments and the project should be submitted individually. However, you are encouraged to (orally) discuss the assignments with your classmates. Each student must write up solutions separately. Be sure that you have worked through each problem yourself and that the answers you submit

are the results of your own efforts. For assignments and the project, you also may not share another student's computer code, submit output from another student's computer session, or allow another student to share your code or output. A good rule of thumb: if a fellow student asks you if you would like to discuss a problem on an assignment, you are encouraged to say "yes"; if a fellow student asks to see your answer to a problem or code, the answer is "no." **You are expected to explicitly acknowledge collaborators by writing their names at the top of your assignments.**

Laptop/Phone Policy: I expect that you will actively engage in class meetings. This means that laptops are used only for note-taking or class exercises and phones are away. My experience is that it is easier to take notes by hand in a quantitative class, and screens can be distracting to others. Please let me know if you plan to take notes electronically.

Accommodations: If you have documentation from Disability Services or anticipate conflicts with the course due to religious observance or a Wellesley-sponsored activity (such as an athletic team), please let me know early in the semester so that arrangements can be made.

Course Outline (subject to change) and Deadlines

		Topic, deadlines
Week 1	Mon, Aug 31 Thu, Sep 3	Bias, sampling, randomization/permutation tests
Week 2	<i>No class Sep 7</i> Wed, Sep 9 Thu, Sep 10	Non-parametric tests Asst 1 due Thu
Week 3	<i>No class Sep 14</i> Open lab time Th, Sep 17	Central Limit Theorem Z-tests, t-tests <i>Open lab time:</i> <i>Tue, Sep 15 12:30-3:30</i> <i>Wed, Sep 16 12:30-3:30</i> Asst 2 due Thu
Week 4	Mon, Sep 21 Open lab time <i>No class Sep 23</i> Thu, Sep 24	Robustness to assumptions <i>Open lab time:</i> <i>Mon, Sep 21, 2:40-4:30</i> <i>Tue, Sep 22, 9:30-12:30, 2-3:30</i>
Week 5	Mon, Sep 28 Thu, Oct 1	Causal inference Asst 3 due Thu
Week 6	Mon, Oct 5 Wed, Oct 7 Thu, Oct 8	Tests for multiple groups, ANOVA Asst 4 due Thu
Week 7	<i>No class Oct 12</i> Thu, Oct 15	Multiple comparisons Fall break
Week 8	Mon, Oct 19 Wed, Oct 21 Thu, Oct 22	Non-parametric methods for multiple variables, regression Midterm due Mon
Week 9	Mon, Oct 26 Thu, Oct 29	Regression Asst 5 due Thu
Week 10	Mon, Nov 2 Wed, Nov 4 Thu, Nov 5	Prediction, robustness of regression Asst 6 due Thu
Week 11	Mon, Nov 9 Thu, Nov 12	Multiple regression Asst 7 due Thu
Week 12	Mon, Nov 16 Wed, Nov 18 Thu, Nov 19	Model selection, missing data Asst 8 due Thu
Week 13	Mon, Nov 23 <i>No class Nov 26</i>	ANOVA v. regression output Thanksgiving
Week 14	Mon, Nov 30 Thu, Dec 3	Generalized linear models (logistic, etc.) Thursday, Dec 3: Poster session during class time
Week 15	Mon, Dec 7	Concluding topics
Finals Period		Self-scheduled exam