
[STAT 201] ADVANCED INTRODUCTION TO STATISTICS AND DATA
SCIENCE
FALL 2024

| | | | |
|----------------------|----------------------|--------------|------------------------------|
| PROFESSOR: | Becky Tang (she/her) | TIME: | Section AZ: MWR 8:15-9:30AM |
| OFFICE: | Warner 214 | | Section BY: MWR 9:45-11:00AM |
| E-MAIL: | btang@middlebury.edu | ROOM: | MW (8:15 & 9:45): Warner 100 |
| OFFICE HOURS: | M 2-4pm, F 9-10am | | R (8:15): Warner 104 |
| PREREQS: | MATH 121 (or equiv.) | | R (9:45): Warner 101 |

COURSE DESCRIPTION

An introduction to statistical methods and the examination of data sets for students with a background in calculus. Topics include descriptive statistics, elementary distributions for data, hypothesis tests, confidence intervals, and regression. Students develop skills in data cleaning, wrangling, visualization, and model fitting using the Statistical Software R. Emphasis will be placed on reproducibility.

Note: this course is going to be a whirlwind! You will have to work hard, but you will learn so much!

KEY LEARNING OUTCOMES

- ☐ Learn the basics of descriptive statistics, statistical theory, and common statistical inferential methods.
- ☐ Develop intuition for and implement appropriate statistical method(s) to answer a particular research question.
- ☐ Learn to explore, visualize, and analyze data in a reproducible and shareable manner.
- ☐ Gain experience in data wrangling, exploratory data analysis, predictive modeling, and data visualization.
- ☐ Practice effectively communicating statistical results through written assignments and oral presentations.
- ☐ Develop working proficiency in coding in R and the tidyverse syntax to create plots, summarise data, and perform basic statistical inference.

TEXTBOOKS AND COURSE MATERIALS

Note: there is nothing that you need to purchase to participate in this class. All materials are used in the class are free and open-source.

■ **Course website:** Most of our course content and schedule will be housed on the course website: <https://midd-stat201-fall2024.github.io/>. Please bookmark this page for easy navigation.

📖 **TEXTBOOK:** We will mostly refer to *Intro to Modern Statistics* by Mine Cetinkaya-Rundel and Johanna Hardin and OpenIntro. There are several ways to access

this book:

- View online here: <https://openintro-ims.netlify.app/>
- Download a PDF here: <https://leanpub.com/imstat>. If you want to skip the optional contribution, set the price to \$0.
- Purchase a hard-copy here: <https://www.amazon.com/Introduction-Modern-Statistics-Mine-C3%87etinkaya-Rundel/dp/1943450145/>

 **TEXTBOOK:** We will occasionally refer to *Modern Dive* by Chester Ismay and Albert Y. Kim. There are several ways to access this book:

- View online here: <https://moderndive.com/index.html>
- Purchase a hard-copy here: <https://www.amazon.com/Statistical-Inference-via-Data-Sci-dp/0367409828/>







COURSE STRUCTURE

A typical class day involves the following:

1. Class session: our 75-minute meetings will typically look like 50 minutes of lecture and live code, with the remaining 25 minutes dedicated to working on problems. The problem-solving time will be further broken down into a bit of time to work individually before breaking out to work in groups. The lecture is intended to introduce new material and the group work session is intended for you to practice implementing methods or exploring the topics more deeply.
 - Some days may be dedicated fully to group work, depending on the topic.
 - Some problems may be designated as 'coding practice problem' (see **Types of Assignments** below for more details).
2. Homework problems: after every class, I will release a few problems that will be assigned to the weekly problem set. Some homework problems will be selected from the in-class practice problems.

A prepared student will attend the 75-minute class, and spend roughly one-two hours per day of class on work outside the classroom (reading, doing homework, attending office hours, studying, etc.). As this course meets three days a week, this represents a minimum 7-10 hour weekly commitment.

CLASS EXPECTATIONS

-  **You are expected to physically show up to class and actively participate,** conditional on classes being in-person. You are an integral part of the class community! Exceptions include previously-communicated illness or planned absence.
-  **Please arrive on time.** I expect everyone, myself included, to arrive on time and dedicate full attention during the class. In turn, I will do my best to always end class at the designated time.
-  **Laptops.** The use of laptops will often be necessary. Plan to bring a laptop everyday and ensure that your laptop has sufficient battery for the duration of the class.
-  **Cell phones should be turned to silent.** I don't mind cell phones in class, but please silence them so as to not disrupt the class.
-  **Please ask questions!**
-  **I encourage discussion amongst yourselves,** especially for clarification or help! However, please be mindful of volume so that the conversations will not be disruptive to the class.

- Ⓢ I expect all members of the class to make participation a harassment-free experience for everyone, regardless of race, creed, color, place of birth, ancestry, ethnicity, national origin, religion, sex, sexual orientation, gender identity or expression, age, marital status, service in the armed forces of the United States, positive HIV-related blood test results, genetic information, or against qualified individuals with disabilities on the basis of disability and/or any other status or characteristic as defined and to the extent protected by applicable law. We will not tolerate the use of violence against any individual.

RESOURCES

- ❑ **Office hours.** This time is meant for you! Please come by to ask questions, chat with me, or work on homework. You should never worry about disturbing me during this time.
- ❑ **TA hours.** TAs will be available on Sunday and Thursday evenings 7-9pm in the Quantitative Center in Armstrong Library to support you on your assignments or chat about their experiences in the course.
- ❑ **One-on-one meetings.** If you would like to meet with me one-on-one, please send me an e-mail or approach me after class so we can schedule a time.
- ❑ **Your peers.** Unless otherwise noted, I encourage students to work together and discuss material! However, unless the assignment explicitly states that it is to be completed as group work, the submitted material must be your own.

TIPS ON HOW TO SUCCEED

- Come to every class and attend office hours.
- Attempt some problems individually before working with others.
- If there is an assigned reading, please complete it in a timely fashion. Many concepts in this class will take time to sink in.
 - Read through the material at least twice. Once without taking notes for an overall overview, then a second time to take notes.
 - Do not skim past the Examples in the readings. Make sure you understand the solutions presented in the Example problems.
- Do not try to find answers on the internet or ChatGPT. Struggling through problems is how you learn!

COLLEGE POLICIES AND RESOURCES

ACADEMIC INTEGRITY

As an academic community devoted to the life of the mind, Middlebury requires every student to reflect complete intellectual honesty in the preparation and submission of all academic work. Details of our Academic Honesty, Honor Code, and Related Disciplinary Policies are available in Middlebury's handbook.

Using AI tools (e.g., ChatGPT, Bard) is highly discouraged in this class. You should not use them to assist in any part of your homework or other assignments. When you have questions, please attend TA hours or my office hours. *The use of AI tools for midterm examinations and to create any content for final project will be treated as a violation of Middlebury's Honor Code.*

HONOR CODE PLEDGE

The Honor Code pledge reads as follows: “I have neither given nor received unauthorized aid on this assignment.” It is the responsibility of the student to write out in full, adhere to, and sign the Honor Code pledge on all examinations, research papers, and laboratory reports. Faculty members reserve the right to require the signed Honor Code pledge on other kinds of academic work.

DISABILITY ACCESS AND ACCOMMODATION

Students who have Letters of Accommodation in this class are encouraged to contact me as early in the semester as possible to ensure that such accommodations are implemented in a timely fashion. For those without Letters of Accommodation, assistance is available to eligible students through the Disability Resource Center (DRC). <https://www.middlebury.edu/office/disability-resource-center>. The DRC provides support for students with disabilities and facilitates the accommodations process by helping students understand the resources and options available and by helping faculty understand how to increase access and full participation in courses. DRC services are free to all students. Please contact ADA Coordinators Jodi Litchfield and Peter Ploegman of the DRC at ada@middlebury.edu for more information. All discussions will remain confidential.

CENTER FOR TEACHING, LEARNING, AND RESEARCH (CTLR)

The CTLR provides academic support for students in many specific content areas and in writing across the curriculum through both professional and peer tutors. The Center is also the place where students can find assistance in time management and study skills. These services are free to all students. go.middlebury.edu/connect

TYPES OF ASSIGNMENTS

* **Coding practice problems.** Assigned roughly 1-2 times per week during the group work session and to be turned in individually on GitHub. These *brief* assignments are intended to give you an opportunity to practice coding in R and receive feedback in a timely manner. I will review your solutions prior to the following class period. These assignments are graded on good-faith effort so you can receive feedback quickly. **Coding practice problems are due to GitHub by midnight on the day they are assigned. For example, your solutions to coding practice problem assigned during class on Monday 9/16 would be due by 11:59pm on that same Monday 9/16.** While this seems like a quick turnaround, I will always leave time to complete the coding practice during class.

- No extensions on coding practice problems will be given, but one coding assignment may be missed without penalty.

* **Problem sets.** Assigned weekly and turned in individually unless otherwise specified (though you will be working with your peers on some problems). In the problem sets, you will apply what you have learned during lecture to dive deeper into the material and explore more interesting or nuanced problems. Most weeks, the problem sets will require R. **Problems will be assigned after every class, but each week's problem set is due to Canvas the following Monday at 11:59pm.**

- For some assignments, you will be asked to submit the R portion of the problem set separate from the non-coding portion of the problem set. This will be clearly denoted on the problem set. Please be sure to submit both files to their corresponding assignments on Canvas.

* **Midterm exams.** Two midterm exams are designed as an opportunity to assess the knowledge you've learned. The *tentative* dates for the exams are as follows. These dates will be finalized at least two weeks in advance of each assessment.

- **Midterm 1: Thursday 10/10 in class**
- **Midterm 2: Wednesday 11/20 and Thursday 11/21 in class**

Except in the cases of extreme illness or family emergency, students must take the midterms on the scheduled date and time.

* **Final project.** A final project presents the opportunity to demonstrate your learning from the entire semester as well as practice exploring a new topic on your own. You will be asked to 1) write a brief paper about a topic and 2) present your findings to the rest of the class. More details will be given after the first midterm. **The presentations will take place on the college-assigned final exam date for this course, which is to be determined. Please do not book travel until the final exam date has been confirmed.**

GRADING

- * Late work on problem sets/weekly assignments will always be considered within one week of the original due date. Unless otherwise stated, the late policy is as follows: for every 24-hour period the assignment is late, 10% from the maximum possible grade will be deducted. For example, if submit an assignments 18 hours after the original due date and it would have originally earned a 90%, it will now earn an 80%.
- * I will do my best to return assignments within one week of submission.
- * Regrade requests: I do allow regrade requests, which must be submitted to be in-person within one week of when the assignment is returned. Keep in mind that regrade requests do not guarantee points back.
- * **You must take each midterm and present the final project in order to pass the course.**

| Component | Percentage |
|-----------------------------|------------|
| Computing practice problems | 5% |
| Weekly assignments | 25% |
| Midterm I | 20% |
| Midterm II | 25% |
| Final Project | 25% |

- * Letter grades will be assigned based on the following course percentages, with the upper 3% and lower 3% of each category corresponding to + and –, respectively:
 - A: 90-100%
 - B: 80-89%
 - C: 70-79%
 - D: 60-69%
 - F: <60%

In the case of decimal points, grades will not be rounded up. For example, a 92.9% will correspond to an A-, not an A.

TENTATIVE COURSE CONTENT

(Last updated: 09/02/24)

NOTE: The following dates and content may be modified due to the requirements of the class may be moved backward or forward depending on class progress and my conference travel. **Midterm dates are tentative.**

| Week | Date | Topic |
|------|-------|---|
| 1 | 9/09 | M - Welcome! Course logistics and introduction W - Sampling schemes and study design R - Installation and intro to R |
| 2 | 9/16 | M - Numerical data - Problem Set 01 due W - Numerical data (cont.) R - Visualizations with ggplot |
| 3 | 9/23 | M - Categorical data - Problem Set 02 due W - Data wrangling R - Data wrangling (cont.) |
| 4 | 9/30 | M - Probability basics - Problem Set 03 due W - Conditional probability R - Simpson's paradox |
| 5 | 10/07 | M - Variability of statistic + Introduction to bootstrap - Problem Set 04 due W - Confidence intervals (CIs) via bootstrap R - Midterm I |
| 6 | 10/14 | M - Recap CIs + Introduction to hypothesis tests (HTs) W - HTs with randomization R - HTs with randomization (cont.) |
| 7 | 10/21 | M - Normal distribution - Problem Set 05 due W - Central Limit Theorem + CIs via mathematical models R - CIs via mathematical models (cont.) |
| 8 | 10/28 | M - HTs via mathematical models - Problem Set 06 due W - HTs via mathematical models (cont.) R - HTs via mathematical models (cont.) |

| | | |
|-----------|-------|--|
| | | - Last day to drop classes |
| 9 | 11/04 | M - Introduce final project + Merge conflicts - Problem Set 07 due W - Introduction to simple linear regression (SLR) R - SLR (cont.) |
| 10 | 11/11 | M -Inference in regression - Problem set 08 due W - Multiple linear regression (MLR) R - MLR (cont.) |
| 11 | 11/18 | M - Project work day W - Midterm II (R) R - Midterm II (written) |
| | 11/25 | Thanksgiving Break |
| 12 | 12/02 | M - ANOVA + Project work day W - Project work day/Peer review R - Project work day |
| | 12/11 | Section AZ final project presentations, 7pm-10pm |
| | 12/12 | Section BY final project presentations, 9am-12pm |