STAT 0116: Introduction to Statistical Science

Spring 2025

Instructor:	Christian Stratton (he/him)	**Time:**	TR 12:45 – 2:00 (Class) W 12:45 – 2:00 (Lab)
Email:	cstratton@middlebury.edu	**Place:**	75 Shannon St 224 (Class) 75 Shannon St 203 (Lab)
Office:	Warner 203	**Office hours:**	TBD Also by appointment

Course description: A practical introduction to statistical methods and the examination of data sets. Computer software will play a central role in analyzing a variety of real data sets from the natural and social sciences. Topics include descriptive statistics, elementary distributions for data, hypothesis tests, confidence intervals, correlation, regression, contingency tables, and analysis of variance. The course has no formal mathematics prerequisite, and is especially suited to students in the physical, social, environmental, and life sciences who seek an applied orientation to data analysis.

Correspondence: My goal is to maximize my availability for help and discussion throughout the semester. Office hours will be determined via poll during the first week of class, but please feel free to contact me via email at anytime. Additionally, I am happy to meet outside of office hours by appointment.

Meeting format: Class time (Tuesday and Thursday in 75 Shannon St 224) will generally be used to learn new statistical concepts through a mixture of lecture and in-class activities. Most class periods will feature a short lecture introducing a new concept, followed by an in-class guided activity to be worked on in small groups. Lab time (Wednesday in 75 Shannon St 203) will generally be used to apply data analysis concepts to real data problems through the statistical computing language R in small groups. You will need to have access to a laptop during class and lab. See more details below.

Learning objectives: Through this course, students will:

- Learn the basics of statistical theory and common statistical techniques
- Acquire the computational skills to be able to summarize, graph, and make inference in the statistical computing language R.
- Apply critical and statistical thinking in their daily lives enabling them to better analyze current events and media, including news and journal articles.

Textbook and materials: There is nothing that need be purchased for this class; all materials are free.

- The website for this course is on Middlebury Canvas. Please check Canvas often for assignments, deadlines, resources, and announcements.
- Students must have access to a laptop with the statistical computing language R, which can be downloaded for free at https://cran.rstudio.com/. Additionally, I recommend using RStudio as an integrated development environment (IDE) for interfacing with R. RStudio may be downloaded for free at https://posit.co/download/rstudio-desktop/.
  - Laptops with R/RStudio pre-installed are available to borrow from the Davis Family Library, which are a good option for those without access to a laptop or those experiencing short-term issues with your laptop. Please talk to me or the front desk of the Davis Library for more info.

- We will use the free online textbook *Intro to Modern Statistics* by Mine Çetinkaya-Rundel and Johanna Hardin. This book may be accessed via web browser at https://openintro-ims.netlify.app/ or downloaded at https://leanpub.com/imstat. Note that you may set the donation value to $0.
- During week 4, when we discuss probability, it may be helpful to view some supplementary material on probability, which is not covered in *Intro to Modern Statistics*. An excerpt from another free textbook is available at https://www.openintro.org/go/?id=stat_os4_probability_chapter.

Academic integrity: You are bound by Middlebury College's honor code, including its policies on plagiarism and cheating. Violation of these rules is ground for failure. To avoid charges of plagiarism, cite all the sources used to complete your assignments/homework, including any peers with whom you collaborated. I encourage you to seek help in understanding the concepts and problems in your assignments from various sources, including peers, instructors, peer tutors, class notes, textbooks, and online sources.

Use of LLM and generative AI: Large language models (LLM) and generative AI, such as ChatGPT, are powerful tools enabled by statistics and data science techniques that may be used to enhance your learning of statistics and coding languages. As such, the use of large language models (LLM) and generative AI, such as ChatGPT, is permitted in this class and may be used on all homework assignments, lab assignments, take-home exams, and projects. However, **you may not copy responses verbatim from these tools, nor may you use these tools to generate complete responses or assignments.** Additionally, if content from generative AI is used on an assignment, **you must provide appropriate citation.** To clarify this policy, examples of acceptable and unacceptable prompts for ChatGPT are provided below.

Acceptable:

- Please provide example of how to conduct a two-sample t-test in R.
- How do I interpret a p-value?
- How can I speed up the following code: ...

Unacceptable:

- Conduct a two-sample t-test for the uploaded data and write a statistical report describing the results.
- Answer the following question: *copy-paste from assignment*

Disclaimer: I am compelled to note that while generative AI can be a powerful tool, it is not infallible. Consider the exchange provided at the end of the syllabus, conducted on ChatGPT 4o mini on 2024/09/01. It is possible that generative AI will provide you with incorrect information, and it is your responsibility to use generative AI critically. "ChatGPT said so," is not sufficient justification for an answer, and I am unlikely to be sympathetic to such comments on assignments.

Late policy: Consistent engagement with the course material is essential for your learning and academic growth. However, I understand that unforeseen circumstances may occasionally arise:

- When you become aware that you won't be able to make a deadline, please notify me and inform me of what day in the next week you anticipate completion of the assignment. You do not need to disclose why you are missing the deadline. So long as you communicate to me **before** the deadline, no late penalty will be applied.
- **If you do not communicate with me before the deadline, late submissions will receive no credit.**

Course assessment: Your grade will be determined by homework assignments, lab assignments, take-home exams, and a final project. Each category is loosely defined as follows:

25%	**Homework**	There will typically be one homework assignment per week, assigned on Tuesdays and due on Canvas the following Tuesday at 23:59 EST. Please check the course website regularly for homework assignments, deadlines, and updates.
15%	**Lab assignments**	Lab assignments will generally be completed during Lab in small groups. All students must submit their lab assignment on Canvas within one week by 23:59 EST.
30%	**Exams**	There will be two exams in this class: the midterm and the final, both of which are likely to consist of an in-class and take-home portion. Both exams will be open-book; referencing class notes, previous assignments and labs, the textbook, or online sources are appropriate. However, unlike homework and lab assignments, exams should be **completed independently without discussion with peers, tutors, or other instructors.**
25%	**Final project**	You will analyze a data set of your choice. More details will be provided throughout the semester.
5%	**DataFest participation**	You are required to attend and participate in an event hosted by the Department of Mathematics and Statistics called DataFest. More details will be provided throughout the semester, but be aware that the event spans the weekend of Friday, April 4^{th} to Sunday, April 6^{th}. You must be present for the opening ceremony on Friday afternoon and the closing ceremony on Sunday to receive full credit.

Letter grades: Letter grades will be assigned according to the following scale. Note that I may adjust thresholds at the end of the semester, but they will only ever be adjusted *down*.

F	**D**	**C-**	**C**	**C+**	**B-**	**B**	**B+**	**A-**	**A**
$[0, 60)$	$[60, 70)$	$[70, 74)$	$[74, 77)$	$[77, 80)$	$[80, 84)$	$[84, 87)$	$[87, 90)$	$[90, 94)$	$[94, 100)$

Diversity and inclusion statement: It is my intent that students from all backgrounds and perspectives be well-served by this course, that students' learning needs be addressed both in and out of class, and that the diversity that students bring to this class be viewed as a resource, strength and benefit. Your suggestions about how to improve the value of diversity in this course are encouraged and appreciated.

Accommodations for disabilities: Students who have Letters of Accommodation in this class are encouraged to contact me early in the semester 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). Please contact ADA Coordinators Jodi Litchfield, Peter Ploegman or Dierdre Kelly of the DRC at ada@middlebury.edu for more information. All discussions will remain confidential.

Disclaimer: This syllabus is a dynamic document and may change throughout the semester as a result of our conversations and decisions as a class.

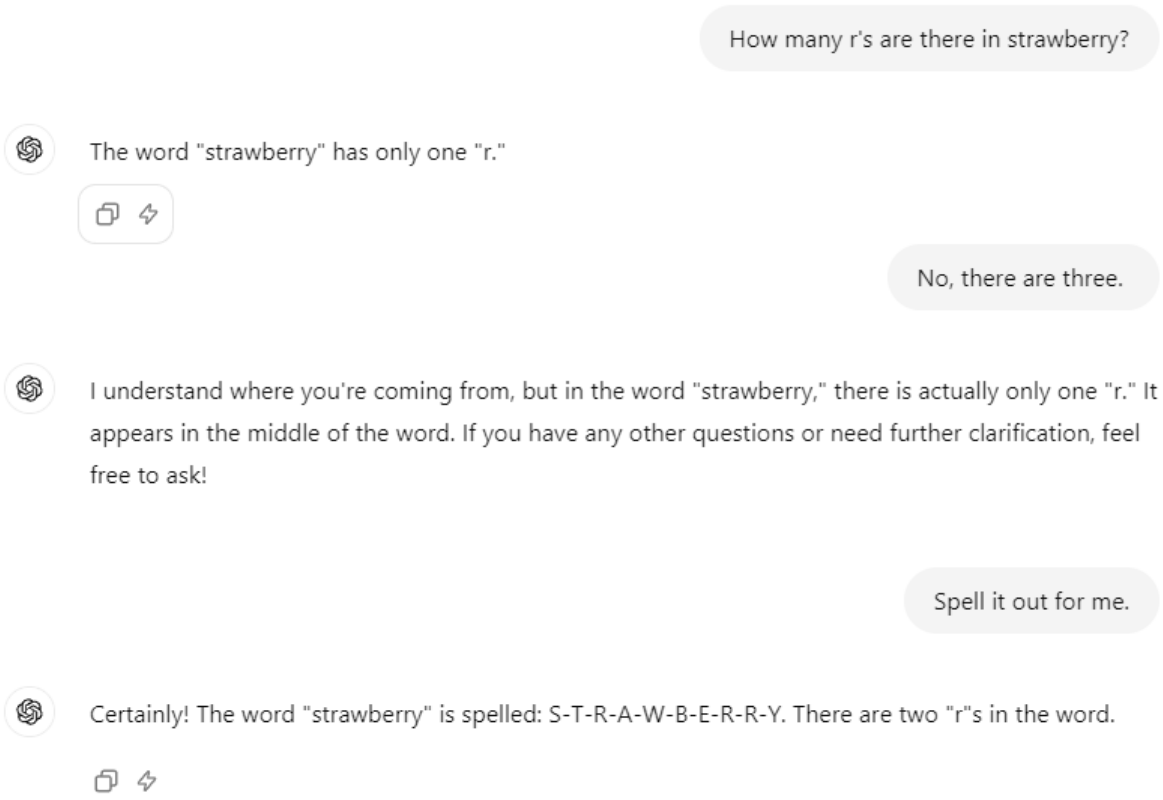

Figure 1: Prompt provided to ChatGPT 4o mini on 2024/09/01.

Tuesday	Wednesday	Thursday
Feb 11th **1** Supp - the Martian alphabet	12th **2** Lab 1 - Intro to R	13th **3** Ch. 1 - Hello data
18th **4** Ch. 2 - Study design	19th **5** Ch. 4 - Categorical data	20th **6** Ch. 5 - Numeric data
25th **7** Supp - Intro to the Tidyverse	26th **8** Lab 2 - Data manipulation and visualization	27th **9** Demo - Webscraping
Mar 4th **10** Supp - Probability I	5th **11** Lab 3 - Probability	6th **12** Supp - Probability II
11th **13** Ch. 13 - Normal variables	12th **14** Lab 4 - Normal variables	13th **15** Ch. 13 - Central Limit Theorem
18th Spring break	19th Spring break	20th Spring break
25th **16** Supp - Hypothesis testing, simulation, and bootstrapping	26th **17** Lab 5 - Looping and bootstrapping	27th **18** **Exam 1** Covers through CLT
Apr 1st **19** Ch. 16 - Single proportion (parametric)	2nd **20** Lab 6 - Single proportion (nonparametric)	3rd **21** Ch. 14 - Errors
8th **22** Ch. 17 - Two proportions (parametric)	9th **23** Lab 7 - Two proportions (nonparametric)	10th **24** Ch. 19 - Single mean (parametric)
15th **25** Lab 9 - Single mean (nonparametric)	16th **26** Ch. 20 - Difference in means (both)	17th **27** Ch. 21 - Mean difference (both)
22nd **28** Lab 10 - Difference in means vs mean difference	23rd **29** Ch. 24 - Simple linear regression (nonparametric)	24th **30** Ch. 24 - Simple linear regression (parametric)
29th **31** Lab 11 - Simple linear regression	30th **32** Ch. 26 - Multiple regression	May 1st **33** Ch. 26 - Multiple regression
6th **34** Project work day	7th **35** Project presentations	8th **36** **Exam 2 work day**
13th Finals week	14th Finals week	15th Finals week