STAT 0116: Introduction to Statistical Science

Fall 2025

Instructor:	Christian Stratton	**Time:**	TR 12:45 – 2:00 (Class) W 12:45 – 2:00 (Lab)
Email:	cstratton@middlebury.edu	**Place:**	75 Shannon St 203 (Class) 75 Shannon St 202 (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: Both class and lab 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. 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.
- Develop the ability to communicate statistical results orally and in writing.

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 two free online textbooks: *Intro to Modern Statistics* by Mine Çetinkaya-Rundel and Johanna Hardin and *Modern Dive* by Chester Ismay and Albert Kim. These books may be accessed via web browser at https://openintro-ims.netlify.app/ and https://moderndive.com/v2/index.html. The IMS book may also be 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 assignments, unless explicitly prohibited by the assignment. 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 in-class activities, readiness assessments, statistical reports, and oral exams. Each category is loosely defined as follows:

10%	**Activities**	Most class days will feature an in-class activity that demonstrates the day's concepts; the activity must be completed by the next class day. **Graded on completion**.
10%	**Readiness assessment**	Most class periods will begin with a short assessment based on the assigned reading for the day. These assessments are designed to be very easy if you completed the reading. **Graded on accuracy**.
40%	**Oral exams**	There will be two oral exams in this class: the midterm and the final. Oral exams consist of an approximately 15 minute conversation with me about topics learned in class. These conversations are meant to demonstrate high-level understanding of statistical concepts. For instance, you will not be asked to recite any formulas, but may be asked to explain the purpose of a p-value.
40%	**Statistical reports**	Throughout the semester, you will be asked to write a number of statistical reports. Full details will be provided as each report arises.

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**
$[0, 60)$	$[60, 70)$	$[70, 74)$	$[74, 77)$	$[77, 80)$	$[80, 84)$	$[84, 87)$	$[87, 90)$	$[90, 94)$	$[94, 100)$

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).

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

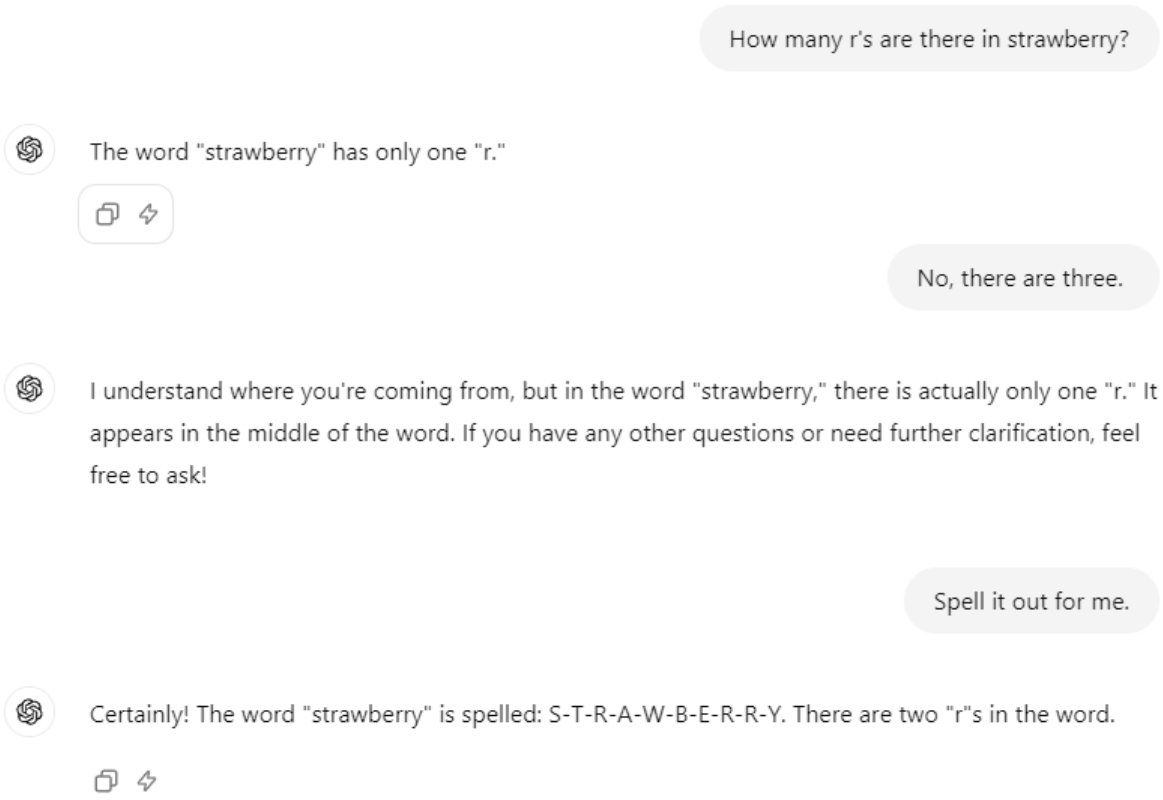

Figure 1: Prompt provided to ChatGPT 4o mini on 2024/09/01.

Tuesday	Wednesday	Thursday
Sep 9th **1** Supp - detecting AI	10th **2** Intro to R and RStudio	11th **3** Descriptive statistics
16th **4** Intro to the Tidyverse	17th **5** Data visualization with ggplot2	18th **6** dplyr/ggplot2 practice
23rd **7** Supp - Probability I	24th **8** Probability II	25th **9** Supp - Probability III
30th **10** Normal random variables	Oct 1st **11** The Central Limit Theorem	2nd **12** The CLT cont.
7th **13** Sampling distributions	8th **14** Intro to hypothesis testing	9th **15** Intro to confidence intervals
14th **16** Simulation based inference for one proportion	15th **17** Simulation based inference for one proportion with infer	16th **18** Theory based inference for one proportion
21st **19** Intro to Chi-squared tests	22nd **20** Chi-squared tests cont.	23rd **21** Simulation based inference for multiple proportions
28th **22** Simulation based inference for multiple proportions with infer	29th **23** Intro to linear models	30th **24** SLR - least squares, SSE, and R^2
Nov 4th **25** SLR - b_0, b_1, and r	5th **26** SLR - theory-based inference	6th **27** SLR - theory-based inference cont.
11th **28** SLR - simulation-based inference	12th **29** SLR - simulation-based inference cont.	13th **30** SLR - prediction and confidence bands
18th **31** MLR - indicator variables	19th **32** MLR - interaction models	20th **33** Flex day
25th *Thanksgiving break*	26th *Thanksgiving break*	27th *Thanksgiving break*
Dec 2nd **34** MLR - equivalence to CLT tests	3rd **35** MLR - equivalence to CLT tests cont.	4th **36** Errors and power
9th Reading days	10th Reading days	11th Finals week
16th Finals week	17th	18th