Land Acknowledgement

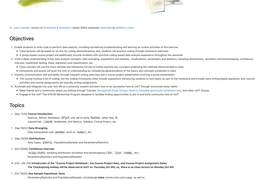
The land on which the University of Toronto operates has for thousands of years been the traditional land of the Huron-Wendat, the Seneca, and most recently, the Mississaugas of the Credit. Today, this meeting place is still the home to many Indigenous people from across Turtle Island and we are grateful to be able to pay respect to the Ancestors, Elders and Land Stewards.

STA130: Introduction to Statistical Reasoning and Data Science

Dashboard Overview Workspace

Statistical reasoning uses data to understand uncertainty and make decisions. Data science is the science of learning from data, and of measuring, visualizing, and communicating information. These two fields are closely related, and they are both essential for making sense of the world in which we live. In this course, we will learn how to think like statisticians and data scientists. We will see how to use computers to analyze real-world datasets, how to critically evaluate statistical claims we encounter in our daily lives, and how to communicate our conclusions to others in a way that is clear, honest, and useful.

This course will focus on the reasoning, computations, and communication skills which are central to the interpretation of statistics and data science. We'll use the Jupyterlab programming environment, the Python programming language, and the Pandas and Seaborn software libraries to explore real-world datasets, to simulate random processes, to perform statistical computations, and to create and interpret informative visualizations and scientific and general-purpose charts and graphs. We'll also discuss the ethical implications of statistical and data science reasoning, including issues related to fairness, accountability, transparency, and the responsible use of data and algorithms.



Dr. Jaclyn Lovelace's version of Greenbaum & Wickham's classic SOAS scheduler; Illustration by Wickham, Aoki

Objectives

- Enable students to write code to perform data analysis, including translating broad/vague/ambiguous/messy/incomplete data into a precise statement of the exercise to be performed, and interpreting the results of the performed analysis in the context of the original broad/vague/ambiguous/messy/incomplete data
- A group-based project (over an extended period, spanning most of the course) in which students are assigned to groups and asked to propose, plan, execute, and report on a data analysis or data science project, including identifying and/or collecting relevant data, performing data wrangling/munging/cleaning/tidying, exploratory data analysis, data visualization, statistical inference, and/or prediction/forecasting, and interpret/critique/evaluate the results of their analysis
- Introduction to the "Data Science Lifecycle" (e.g. https://www.datascienceethics.com/ and https://www.ischoolonline.com/blog/what-is-data-science-lifecycle) and to various "computational thinking" concepts (e.g. https://www.cs.cmu.edu/~15-110-spring-2015/schedule/index.html) including "algorithmic thinking" and "computational reproducibility"
- Introduction/review of basic arithmetic through Pre-calculus to support the course project and exposition involving a quasi-presentation
- Introduction to "Data Ethics" concepts (e.g. https://www.datasciencecourse.org/notes/ethics/) including "algorithmic fairness" and "algorithmic transparency" and "algorithmic accountability" and "algorithmic ethics" and "algorithmic bias" and "algorithmic harm" and "algorithmic transparency and

to me (the professor) that I spend my time demonstrating how to practically go about doing something, rather than just discussing "theoretically" what the different aspects and considerations of a given topic are.

> The weekly coding practice through homeworks and office hours are meant to provide ample reinforcement and support as you learn to practically apply coding to perform statistical and data science analyses. **Building coding comfort is a primary objective of the course**. *The approach to the course being taken here is directly intended as strategy designed to better address this objective*.

Even with all of the above in mind, the conceptual understanding of a topic is of course certainly relevant and often highly useful in facilitating actual practical applied analysis; so, we still need to spend time on the coneptual topics of the methodologies examined in the course. Some of the relevant conceptual topics related to the practical analyses of the course has been moved into the homework that explore the topics and homework exercises. But the majority of the effort in the course put towards understanding statistical and data science concepts is now going to take place as part of the STA130 tutorials.

> STA130 TAs begin the week preparing for ~60 minutes for 60-120 minutes of office hours; then, they spend ~90 minutes marking the weeks homework and then another ~60 minutes preparing to the presentations and activities they'll give during the ~120 minutes of the STA130 tutorial. They'll then spend another ~150 minutes marking student tutorial activity and assignment submissions into the following week.

STA130 tutorials will review the topics of the course, but with the overall processes of the methodologies having already been covered in class, the tutorials can now more deeply discuss and explore the conceptual understandings related to these analyses. Further, since the STA130 TAs have more recently experienced the learning process that STA130 students are going through, STA130 TAs will be well positioned to highlight and emphasize the key explainations that they feel are most effective useful for understanding the topics at hand. STA130 tutorials are thus viewed as a critical and essential component of STA130, and this is reflected in their contribution to the overall course grade.

> STA130 tutorials use presentations and activies specifically designed to faciliate exploring and learning the most important topics and concepts in STA130. The sequencing of the STA130 tutorial presentation topics and activies are intentional and tightly regulared in terms of allotted time expectations.

TA-Student Communication

TA-student Communication is limited to in-person exchanges in OH and Tutorial.

- TAs should NOT share personal emails with students or communicate with students through personal emails.
- If TAs need to communicate with their students, they can do so through a Quercus Announcement to this Tutorial Sections.
- If students feel they need to contact their TAs they should email sta130@utoronto.ca and the head TA will address their inquiry and facilitate follow up as necessary.

Marking

Marking is expected to be completed within one week of submission with delays expected to be a relatively minor occurance.

Homework

TAs have been alloted ~90 minutes to mark the Homework for the ~24 students. This means that TAs have ~4 minutes to review and enter the mark for each students Homework. This will not be enough time to provide any substantive feedback on homework submissions; however, homework feedback is highly automated through MarkUs. So, it is not the TAs job to provide feedback on the homework; rather, it is the TAs job to assign a reasonably fair scores for demonstration of effort and correct completion on the homework assignment submission.

Homework marks are thus assigned though the following two-stage process:

1. A rough “proportion completed” score is quickly assigned based on a quick characterization of student submissions suggested by a glance assessment of the MarkUs autotesting results

0	no substantive submission
1/4:	about 1/4 of the problems completed with reasonable effort
1/2:	about 1/2 of the problems completed with reasonable effort
3/4:	about 3/4 of the problems completed with reasonable effort
1:	all problems completed with reasonable effort

> where “completed with reasonable effort” means that MarkUs autotests are correct "more often than not".
>
> - The TAs will carefully note which MarkUs autotests are intentionally designed to automatically fail for the purposes of provided example solutions to students and do not count these against the students as "wrong/red"
> - So, if the autotests are "mostly green", then the "proportion completed" just roughly estimates where "mostly green" stops and "all the rest are wrong/red" begins (ignoring automatically failing autotests).
> - Students can run autotests; so, if a student homework submission is not working, the student is responsible for noticing this and correcting any problem with the coding that's causing the error.

1. Secondly, there will be a handful of problems which are assigned manual review by the TAs. If a TA does not think an answer demonstrates sufficient understanding of the topics in question in any question they are directed to manually review, they will deduct 1/4 point from the initial "proportion completed" score for each problem they are not satisfied with (down to a lowset score of 0, as assigned HW scores will not be negative).
   - TAs should only deduct points when an answer fails to reasonably demonstrate an understanding of the topics of its contents, or fails to understand the prompt of a question

Tutorial Assignment

TAs have been alloted ~120 minutes to mark the Tutorial Assignment for the ~24 students which means that TAs have ~5 minutes to review, provide feedback, and enter the mark for each students Tutorial Assignment. **The Tutorial Assignment itself is expected to take ~60 minutes for students to complete outside of tutorial after tutorial has finished; typically involves a written submission of somewhere between 200 to 500 words; and, is typically based on an electronic submission through Quercus. Tutorial Assignments are made at the end of Tutorials on Friday; are due by 9 AM the following Monday of the next week; and may be submitted without attending tutorial.** The default Rubric for a Tutorial Assignment generally conforms to the sum of the following two columns, with allowances for adjustments of 1/8 for submissions that fall somehwere in between the levels indicated below.

	General Impression of Written Communication	Substantive Correctness of Answer Content
Clear and Understandable	1/2	1/2
Mostly makes sense and can generally be followed	1/4	1/4
Hard to follow and not comfortably communicated	0	0

Tyically, recieving points for the "General Impression of Written Communication" would be expected to in some sense be prerequesite for recieving points for the "Substantive Correctness of Answer Content". There may quite possibly not be any single specific answer which is correct. In which case, "Impression" and "Correctness" are assessed on the basis of internal coherrency, consistency, and sensibility of the provided arguments and statements.

Tutorial Activity

The Tutorial Activity takes place during tutorials and is a participation and engagement score that requires attendance during the tutorial. **Typically the Tutorial Activity requires students to hand in written submissions on a physical sheet of paper during the tutorial.** The TAs will not review the answers of these Tutorial Activity submissions, however; and, they will only use them for entering marks for completion of the tutorial activity. The answers to Tutorial Activity will typically be conversationally reviewed as part of the tutorial presentation. TAs have been alloted ~30 minute to enter Tutorial Activity marks into Quercus; but, since this is more time than is needed, the TAs will use the excess time in completing the other marking activities as needed.

JupyterHub and MarkUs

- Quick Demos
- Some conversation about how Autotesting works in MarkUs

Legal Notices

Academic Integrity

You are responsible for knowing the content of the University of Toronto’s Code of Behaviour on Academic Matters. We encourage you to discuss course material with each other and ask others for advice; however, it is not permitted to share answers or to directly share code or written work for anything that is handed in as a course assignment. For example, “For question X what function did you use?” is a fair question when discussing course material with others in the class; but, “Please show me your code for question X” is not an appropriate question. If writing or code is discovered to match another student’s submission or outside source, this will be reported as an academic offence. If it is found that submitted code and its corresponding output document do not actually match, this will be reported as an academic offence. Academic offences will be taken seriously and dealt with accordingly. If you have any questions about what is or is not permitted in this course, please seek advice from your tutorial section TA.

> Exception: you may discuss and share answers and code with other members of your project team where the effort is understood to be group work.

All students, faculty and staff are expected to follow the University’s guidelines and policies on academic integrity. For students, this means following the standards of academic honesty when writing assignments, collaborating with fellow students, and writing tests and exams. Ensure that the work you submit for grading represents your own honest efforts. Plagiarism -- representing someone else’s work as your own or submitting work that you have previously submitted for marks in another class or program -- is a serious offence that can result in sanctions. Speak to me or your TA for advice on anything that you find unclear. To learn more about how to cite and use source material appropriately and for other writing support, see the U of T writing support website at http://www.writing.utoronto.ca. Consult the Code of Behaviour on Academic Matters for a complete outline of the University’s policy and expectations. For more information, please see https://www.artsci.utoronto.ca/current/academic-advising-and-support/student-academic-integrity and http://academicintegrity.utoronto.ca.

AI tools such as ChatGPT

Students may use an AI tool such as ChatGPT for enhancing their learning objectives, however, they must not share the course materials and assessments as all content entered may become part of the tool’s dataset and may inadvertently resurface in response to other prompts. While it's possible that the use of generative artificial intelligence tools or apps for assignments in this course, including tools like ChatGPT and other AI writing or coding assistants could be helpful, simply copy-pasting will do very little to support learning. Read and understand and at most paraphrase helpful responses from any generative artificial intelligence applications.

Plagiarism

You must not present the work of others as your own. This includes, but is not limited to, copying text and including it in your writing without a citation and quotation marks. There are many resources to help you learn more here: https://www.academicintegrity.utoronto.ca/smart-strategies/

Intellectual Property Statement

Course material that has been created by STA130 instructors (lecture slides, course activities, quizzes and examinations, etc.) are the intellectual property of the STA130 instructor and is made available to you for your personal use in this course. Sharing, posting, selling or using this material outside of your personal use in this course is not permitted under any circumstances and is considered an infringement of intellectual property rights.