STA130H1 S — Introduction to Statistical Reasoning and Data Science

University of Toronto, Spring 2022
Syllabus Version: January 12, 2023
Course Website: https://q.utoronto.ca/courses/296457
Course Email: sta130@utoronto.ca

Teaching Team

Instructor: Prof. Joshua S. Speagle (沈佳士)
Office: AB 206 / OPG 9108
Email: j.speagle@utoronto.ca
Website: https://joshspeagle.com/
Office Hours: Fridays 3:10-5:00pm in AB 206 or by appointment

Mentorship Program Coordinator: Ivan Nguyen, M.Ed.
Email: ivan.nguyen@utoronto.ca

Lead Teaching Assistant (TA)/English Language Learner (ELL) TA:
Quin Xie

Teaching Assistants (TAs):
Hriday Chheda (Office Hours: TBA)
Kevin Sun (Office Hours: TBA)
Tong Su (Office Hours: TBA)
Jerry Yuxuan Wu (Office Hours: TBA)
Yu Zhang (Office Hours: TBA)
Ethelia Choi (Office Hours: TBA)
Benjamin Zhang (Office Hours: TBA)
Lily Thao Nguyen (Office Hours: TBA)
Xiaoxuan Han (Office Hours: TBA)
Lijing Wei (Office Hours: TBA)
Hyung Eun Shin (Office Hours: TBA)
Jaffa Romain (Office Hours: TBA)
Rachael Jaffe (Office Hours: TBA)
Amin Banihashemi (Office Hours: TBA)
Rafael Valencia Sánchez (Office Hours: TBA)
Yupeng Zhang (Office Hours: TBA)
Adeline Leonardi (Office Hours: TBA)

Sahil Patel (Office Hours: TBA)
Ananya Jha (Office Hours: TBA)
Stephanie Ziembicki (Office Hours: TBA)
Yutong Chen (Office Hours: TBA)
Xinyi Yao (Office Hours: TBA)
Xiao Wu (Office Hours: TBA)
Ijeoma Itanyi (Office Hours: TBA)

Course Summary

Data permeates almost every aspect of our modern lives. But how do we make sense of it all? How do sports coaches understand how their team is performing? How does Amazon figure out optimal shipping and distribution routes? How do geneticists uncover genes that can increase our risk of certain diseases? How does TikTok decide what content to show us?

This course is an introduction to statistical reasoning and data science, which are catch-all terms for the interdisciplinary field and associated skillsets that have arisen trying to solve many of these modern-day problems that encompass aspects of acquiring, managing, and analyzing data. It therefore will provide you with some of the basic skills and intuition needed to start your journey into the broader fields of data science.

We will be exploring a number of foundational concepts throughout the course, which will be focused on two main avenues. The primary avenue is a *"question-oriented" strategy*, where we will focus on particular set of questions and then explore ways to address them using particular datasets. This will be explored through much of the direct course materials and generally follows the following pattern:

1. Identify the fundamental *question(s)* we wish to address.
2. Determine what *dataset(s)* we need to address it.
3. Come up with a *strategy* that uses the dataset(s) to try and address it.
4. Try and *implement* our proposed strategy in a robust and computationally feasible way.
5. Apply our code to the dataset(s) and *analyze* the results.
6. *Interpret* the results and consider potential caveats.
7. *Communicate* our findings to others.

The secondary avenue is a *"data-oriented strategy"*, where we start with an existing dataset and try to find interesting questions or patterns within it. This strategy, also known as

"exploratory data analysis", is an integral part of how many areas of modern data science work in practice. This will be primarily explored through your final Capstone Project.

This course, intended for students considering a program in Statistical Sciences, discusses the crucial role played by statistical reasoning in solving challenging problems from natural science, social science, technology, health care, and public policy, using a combination of logical thinking, mathematics, computer simulation, and oral and written discussion and analysis. As such, it is intended to provide a broad introduction to many of the ways we can learn from data, focusing on statistical reasoning, computation, and communication. As such, all three aspects will make up an integral part of the course.

This course will primarily use the **R programming language** and environment.

Course Goals

By the end of this course, you should be able to:

1. Describe how *statistical methods* for description, explanation, and prediction can be used to learn from data.
2. Identify *appropriate uses* of statistical methods to answer questions, including their corresponding strengths and weaknesses.
3. Carry out a variety of statistical *analyses in R*, interpret the corresponding results, and ensure they are robust and reproducible.
4. Clearly *communicate* the results of statistical analyses to both technical and non-technical audiences.

Course Requirements

Corequisites: Single-variable calculus is required as a corequisite (MAT135H1, MAT136H1, MAT137Y1, or MAT157Y1). Some exposure to computer science or programming is strongly recommended (e.g., CSC108H1, CSC110Y1, CSC120H1, or CSC148H1) but not required.

Exclusion: Any of STA220H1, STA255H1, STA238H1, STA248H1, STA261H1, ECO220Y1, ECO227Y1, STAB22H3, STA220H5, STAB57H3, STA258H5, STA260H5, ECO220Y5, ECO227Y5, and/or STAA57H3 taken previously or concurrently.

Distribution Requirement: Science.

Breadth Requirement: The Physical and Mathematical Universes (5).

Course Structure

The course structure is spread across four venues:

- **Class meetings** will be held on Mondays from 09:00-11:00 in PB B150 for Section 1 (LEC0101) and 13:00-15:00 in MC 102 for Section 2 (LEC0201). Attendance is strongly encouraged but *not required* or graded. These will be in person and *will not* be recorded, although slides or other materials will be uploaded to the course website for reference. You are permitted to attend class meetings for either Section except if an in-person exam is taking place (although given capacity restrictions might be tight, you are strongly encouraged to try and attend your original Section whenever possible).

- **Tutorials** will be held on Fridays from 09:00-11:00 for Section 1 (TUT0101-0112) and 13:00-15:00 for Section 2 (TUT0201-0212). **Attendance is required** and graded (see below). These may be hosted online over Zoom if a TA is unable to host your tutorial in person, but this should only occur if all other alternatives have been exhausted. If you are unable to attend your tutorial due to illness or other extraordinary circumstances (e.g., not just generic scheduling conflicts), the you will be permitted to attend remotely. If you know ahead of time that you will be unable to attend tutorials at a particular time due to other immovable conflicts (see also some of the accessibility and accommodation policies below), please send an email to your TA as well as sta130@utoronto.ca with the subject "Tutorial Scheduling Conflict" as soon as you are able so we can work towards a solution.

- **Office hours** will be held regularly from Tuesday through Thursday by the course TAs during from 9:00-17:00 ET. Individual TAs will generally host office hours at the same time and venue every week. These will involve a combination of in-person and online options. Instructor office hours will be held in-person on Fridays as well as by appointment. Outside of TA office hours (which are accessible to all course students), you can also seek help online (see blow). You can also seek help from your assigned TA either during your Tutorial (the preferred option) or via email (if necessary; see additional details below).

- **Online** assignments, quizzes, discussion forums, etc. will primarily be accessed through the Course Website on Quercus. Assignments will be hosted online in an interactive environment so that everyone will be working off of the same code environment and from the same baseline document(s). Course discussions, student forums, etc. will take place on Piazza as well as Quercus (as appropriate). TAs and the Instructor will try to monitor and respond to relevant questions within 48 hours during the week as well as the same day problem sets are due.

Course Materials

There are no required textbooks or other materials. All necessary resources such as slides, datasets, and links will be provided by the Course Instructor as needed (with the exception

of the data for the Capstone Project, as described below).

Grading

Grades will be determined as follows:

- **Mentorship Program: 3%**
- **Tutorials: 12%** (best 7 out of 9)
  - Tutorial Attendance: 4%
  - Tutorial Exercises: 8%
- **Weekly Problem Sets: 15%** (best 7 out of 9)
- **Capstone Project: 25%**
  - Project Proposal: 5%
  - Progress Report: 2%
  - Final Presentation: 2%
  - Final Report: 15%
  - Final Reflections: 1%
- **Exams: 45%**
  - Midterm: 10%
  - Final: 35%

Each of these categories are discussed in more details below.

Missing/Late Policy

No extensions will be granted and no missing or late work will be accepted for weekly problem sets or tutorial exercises. Given the grading scheme and regrading/revision policies, you are instead strongly encouraged to submit assessments even if they are incomplete.

If you have a valid reason that you are and/or will be unable to participate in the mentorship program, attend tutorials, complete the final capstone project, or miss the midterm and/or final, you may request an accommodation or alternative assessment (which may include an oral exam). To do so, please email sta130@utoronto.ca and include "Missing/Alternative Assessment Request" in the subject line.

Re-Grading Policy

Any request to have an assessment re-graded can *only* be made by sending an email to sta130@utoronto.ca (i.e. *not through your TA*) and including "Re-grading Request" in the subject line. These must be requested *within one week* (seven days) of grades being posted, with the exception of the Final Report and Final Project Reflection (which must have a regrade submission within three days of grades being posted). The request will be reviewed by the course instructor and the lead TA and must include the following information:

- Your name and student number.
- A *detailed and written justification* referring to your answer as well as references to the relevant course material to be considered. *It is not enough to just say you believe that your answer deserves higher credit.*

If your request is approved, please note that the *entire assessment* will be re-graded (i.e. the overall grade may increase *or decrease*) by another TA, the head TA, and/or the Instructor.

Revision Policy

Weekly problem sets are subject to a separate "revision" policy to encourage you to correct mistakes and improve understanding of concepts outside of the midterm and final assessments. If your score in a weekly problem set that you completed falls *below 70%*, you can submit a "revised" problem set *within one week* of grades being posted. This must be submitted to your individual TA with "Revised Problem Set" included in the subject line, and include:

- The original solution.
- The new proposed solution.
- A *detailed explanation* for why the original solution was incorrect and why the new proposed solution is correct.

If the new proposed solution *and* the explanation for the original mistake and new proposed solution are correct, you will recover half of the points that they lost associated with that error *up to a maximum of 70%* (i.e. revisions make it possible to raise any score above 40% up to 70%). If either the proposed solution or the explanation is incorrect, no points will be awarded. This revision is not subject to the re-grading policy described above.

Mentorship Program (3%)

Finding community and support on campus not only increases your chances of academic success, but generally improves mental well-being and makes university life more fun. In

addition, since data science is an interdisciplinary and rapidly evolving field, its very nature tends to encourage connection, collaboration, and mentorship. As such, participation in the mentorship program contributes to part of the overall grade.

This program provides a foundation for success as you move through your time at UofT by exposing you to three pillars of learning and support from upper year students who know what you're going through. These are each worth 1% of your grade and involve:

1. **Social and Personal Development** (1%): Attend *at least one* social or personal development activity (event organizers will record attendance) and write a paragraph reflecting on what you learned.

2. **Career Exploration** (1%): Attend *at least one* career exploration activity (event organizers will record attendance) and write a paragraph on what you learned.

3. **Peer-to-peer mentorship** (1%): Connect with a STA130 peer mentor (your mentor will affirm your meeting/conversation) and write a paragraph reflecting on what you learned.

More details will be made available on Quercus/SharePoint, including a link to the events calendar and additional biographies of STA130 peer mentors. As grading for these events will be ongoing throughout the semester, **please allow for up to two week from the date of the activity before the grade appears in Quercus**. Note that while you will only receive credit for one event/meeting in each category, you are highly encouraged to attend more if you find them helpful!

For questions about the program, please contact the Mentorship Program Coordinator Ivan Nguyen (ivan.nguyen@utoronto.ca) and CC the course email (sta130@utoronto.ca) with "STA130 Mentorship Program" in the subject line.

Tutorials (12%)

Each week, you will earn a tutorial grade for attending Tutorials (4%) along with participating in and completing tutorial exercises (8%), which will be combined into a weekly Tutorial grade. There will be a tutorial exercise each week (outside of the first week of class, reading week, the week of the midterm assessment, and the last week of the semester) that will be due **Fridays at 11:59 pm ET** (i.e. later that day, if you aren't able to complete all exercises during the tutorial). This gives a total of 9 tutorial exercises. Your lowest two combined Tutorial grades will be dropped (i.e. including both attendance and performance), so only your top 7 combined Tutorial grades will count towards your final grade.

Problem Sets (15%)

A problem set will be assigned each week (outside of the first week of class, reading week, the week of the midterm assessment, and the last week of the semester) due **Thursdays at 11:59 pm ET** (i.e. the night before Tutorials). This gives a total of 9 problem sets. Problem sets will typically be made available by Friday the week prior (i.e. before the next Monday class meetings), although you are not expected to work on them over the weekend. *Your lowest two problem set scores will be dropped*, so only your top 7 problem sets scores will count towards your final grade.

Problem sets will typically consist of two parts:

1. *Programming in R*: You'll be asked to complete a set of questions using R to generate figures, tables, or other statistical analyses and interpret your results. The first questions will be mandatory while the others will be optional (although you are still encouraged to complete them to help prepare for upcoming larger assessments). **Your code will need to successfully compile and run to receive full marks.** You are heavily encouraged to include comments in your code to make it easier for your TA to award potential partial credit should any issues arise.

2. *Communicating Results*: You will be asked to submit a short written or spoken piece communicating your results along with other statistical ideas to a non-technical audience using the vocabulary introduced in the course.

Submissions will only be accepted through Quercus, so please do not send any homework submissions by email to the course website, Instructor, or any TA except under extraordinary circumstances (which will need to be described in the email). While you can submit as many times as you would like to Quercus, only the last submission before the deadline will be usually graded (unless you have specifically left comments letting us know otherwise).

Capstone Project (25%)

The Capstone Project involves synthesizing many of the topics discussed in the course and includes an analysis of a real-world dataset. As a central component of the course, it is worth around a third (31%) of the total grade. A general description can be found below. Additional details will be provided on Quercus.

Group Policy

The Capstone Project will be a **group project**, with up to four students in a particular group. *These groups will be randomly assigned within each Tutorial* (for a total of up to six groups per tutorial). Students who would like to work together can express their preferences to their TA along with some provide justification, but there is *no guarantee* that we will be able to accommodate all individual student preferences. Once groups are formed and

the Project Proposals (see below) have been accepted, *students are not permitted to change groups* except under extraordinary circumstances (and will require permission from your TA and the Instructor). To accommodate students who may add/drop the class later in the semester, the overall expectations all aspects of the Capstone Project will be adjusted based on the final overall group size.

To ensure that individual student efforts within each group can be evaluated fairly, you will also be required to submit an **individual contribution statement** that describes your particular role(s) and/or contribution(s) to the overall project. **If the levels of individual contributions vary substantially within a particular group, TAs and the Instructor reserve the right to assign *individual* grades to one or more students within the group** related to any submission related to the Capstone Project. Justification for this decision will be provided and individual grades will remain subject to the course re-grading policy (see above).

Data Policy

Groups may work with any dataset of their choosing for the Capstone Project. Note that *at least one* common dataset will be made available to all students in the class specifically for the Capstone Project. There is no penalty or reward for working with a different dataset other than the common datset(s) provided to you. If your choose to work with one or more alternate datasets, you are *entirely responsible* for obtaining both that dataset as well as any permissions necessary to use that dataset as part of the course. A copy (or a link to a copy) must also be submitted along with the Project Proposal (see below) and the Progress Report and Final Report as well (if particular aspects of the data change between assessments). If you have any additional questions, please email sta130@utoronto.ca along with your individual TA with "Capstone Project Dataset" in the subject line.

Project Proposal (5%)

As part of the Capstone Project, each group will have to submit an initial **Project Proposal** in **mid-February**. This is expected to be around 1-2 pages (single-spaced) and will include details on three interesting research questions that final project will try to answer, a rough description of the proposed methods, and some details regarding the expected contributions of each group member. This will be a group submission. After all proposals are submitted, the teaching team will work together with groups to modify proposals and provide guidance as needed.

Progress Report (2%)

Around a month after the Project Proposal has been submitted, each individual/group will have to submit an interim **Progress Report** in **mid-March**. This is expected to be 2-3 pages (single-spaced), including some figures and/or data tables, and directly address

progress on the topics and associated timelines outlined in the original Project Proposal. This will be a group submission but will also include individual contribution statements submitted by each group member (see above).

Final Presentation (2%)

The final week of the class will be devoted to the **Final Presentation** that is currently scheduled to take place on **April 3** during the usual class time. The current plan is that this will involve designing and presenting a **conference-style poster** at a STA130 poster fair, where the poster will be graded on specific information content as well as overall style. This will be a group submission but will also include individual contribution statements submitted by each group member (see above).

Final Report (15%)

After the Final Presentations have concluded, each individual/group is required to submit a **Final Report** that will be due on **April 6** (the last day of class). This is expected to be around 6-10 pages (single-spaced and including figures) and include a title, abstract, introduction, overview of data and methods, description of the results, interpretation and discussion of the final findings, and conclusion. This will be a group submission but will also include individual contribution statements submitted by each group member (see above).

Final Reflections (1%)

You will be asked to provide some small feedback on other group posters as well as your own experience working on the Capstone Project (e.g., challenges faced, changing plans, particular methodology details learned, etc.). This will be an individual submission.

Exams (45%)

There will be an **in-class midterm** assessment on **Monday, February 27** following reading week that will be worth 10% of the overall grade. An **in-person final exam** will also be held at the end of the course on [**Exam Date TBD**] worth 35% of the final grade.

Academic Integrity Policy

Academic integrity is one of the cornerstones of the University of Toronto. It is critically important both to maintain our community which honours the values of honesty, trust, respect, fairness and responsibility and to protect you, the students within this community, and the value of the degree towards which you are all working so diligently.

As such, while collaboration is heavily encouraged when planning activities to accompany

weeks leading discussion, **all submitted assignments are expected to be your own work**. In particular, **it is not permitted to share answers or to directly share R code or written answers for anything that is to be handed in (e.g., week problem sets)** or to plagiarize text written by another student. Doing so would be considered an **academic offense**, which are treated extremely seriously. If necessary, you should cite any books, articles, websites, lectures, and other resources outside of the reading materials using appropriate citation practices. While not necessary, it is also a good habit to explicitly acknowledge classmates that you collaborated with if their input helped you with the final material that you are submitting.

Note that for assignments such as problem sets that require R code, the code you submit *must* have been used to generate the document. If the submitted code does not match the submitted output and an explicit reason is not given upon submission, this would also constitute an academic offense.

For more information, please see: `https://www.academicintegrity.utoronto.ca/`.

Accessibility

Students with diverse backgrounds, learning styles, and needs are welcome in this course. In particular, if you need academic adjustments or accommodations because of a documented disability, health consideration, or other any other reason, please reach out to the Instructor and/or the Accessibility Services Office (ASO) at accessibility.services@utoronto.ca and/or Accomodated Testing Services (ATS) at ats.info@utoronto.ca as soon as possible. The sooner we know your needs, the faster we can work together towards helping you achieve your learning goals in this course.

If you have an accommodation letter from your accessibility advisor that is relevant to this course, please do the following:

- Email your letter to sta130@utoronto.ca with "Accomodation Letter" as part of the subject line, cc your advisor, and include anything else that you wish us to know and/or any additional questions you may have as soon as you can.
- Ensure that you register any assessment that you require an accommodation for with the ASO/ATS at least two weeks before the assessment date.
- Confirm any accommodations for *each* specific assessment with the teaching team at least one week before the assessment date (i.e., if you receive extra time for timed assessments, confirm this with the teaching team at least one week prior to the midterm assessment and the final assessment, even if this was already discussed at the beginning of the semester).

For more information, please see: `https://www.accessibility.utoronto.ca/`.

Accomodation for Religious, Indigenous, and/or Spiritual Observances

The University of Toronto supports reasonable accommodation of the needs of students who observe religious holy days other than those already accommodated by ordinary scheduling and statutory holidays. Please email sta130@utoronto.ca with "Accomodation for Observances" as part of the subject line at least three weeks in advance if you require accommodations or expect absences so we can work together to make alternative arrangements.

For more information, please see: `https://www.viceprovoststudents.utoronto.ca/policies-guidelines/accommodation-religious/`.

Course Schedule and Topics

Subject to change as the semester progresses. Topics will be updated based on what is/will be covered in class.

Week 1 (Jan 9): Overview and logistics

In the first class (Week 1), will go through the syllabus (this document) together and address any questions that you might have.

Week 2 (Jan 16): Introduction to R and statistical reasoning

What is R? What do we mean when we say "statistical reasoning"? In week 2, we will cover the basics of R, along with the online environment that will be used for all upcoming problem sets and other assessments. We will also begin exploring some basic statistics questions, including distributions and summary statistics.

Week 3 (Jan 23): Additional R topics

How can we use R to start exploring and analyzing our data? While last week we got started with the basics, this week will continue our exploration of R with a focus on input data types, data manipulation (often through large tables), utility functions, and plotting/visualizing results. We will also explore how these can be used to start exploring some of the statistical concepts we began addressing last week.

Week 4 (Jan 30): Hypothesis testing and quantifying uncertainties

We now know how to compute summary statistics, but how can we use them to test out hypotheses? How can we figure out how confident (or uncertain) our results are? In week 4, we will begin exploring how to test hypotheses when comparing one dataset to an assumed answer (1-sample test) or when comparing two datasets to each other (2-sample test). We'll also discuss how to assign our conclusions appropriate levels of confidence along with asso-

ciated uncertainties.

Week 5 (Feb 6): Capstone Project datasets (Guest Lecture)

In Week 5, we will have a guest lecture! Our guest will provide an overview of the common dataset that can be used for the capstone projects. It will also be an opportunity to answer any questions regarding details related to the final project.

Week 6 (Feb 13): Sampling, bootstrap, jackknife errors

How can we use computational methods to approximate randomness and real-world behaviour? Can we use these to derive uncertainties? In Week 6, we will explore how we can use (pseudo-)randomness and data simulations to estimate uncertainties. We'll first explore methods for the case where we want to assume we know something about the data itself (similar to 1-sample hypothesis tests), and then two (the bootstrap and the jackknife) in more general cases where we don't (or can't) make those assumptions (similar to 2-sample hypothesis tests).

Week 7 (Feb 20): READING WEEK

No class!

Week 8 (Feb 27): MIDTERM

We will have an in-class midterm assessment in Week 8.

Week 9 (Mar 6): Linear regression I

How can we uncover relationships between variables? What's the difference between a dependent and independent variable? In Week 9, we'll cover the most common statistical analysis method used in almost all applications today: linear regression. We'll cover the basics of how it works, how it can be modified, and how we can interpret the results.

Week 10 (Mar 13): Linear regression II

How can we extend our analysis to include multiple independent variables? How do we interpret the results? How can we be confident if a variable is really informative? In Week 10, we'll continue where we left off by exploring how linear regression can be generalized to include many variables. We'll also start going into more practical usage, including how to modify the framework, validate your approach, and interpret the results.

Week 11 (Mar 20): Classification, logistic regression, and decision trees

What are we supposed to do if we're trying to predict particular classes or yes/no answers,

instead of just a number? Are there other tools we can use to help us out? In Week 11, we'll go through the basic problem of classification and how it differs from what we've look at before. We'll discuss how we can to build on our linear regression knowledge to come up with a logistic regression model that does the trick. Afterwards, we'll use the opportunity to start exploring more "machine learning"-oriented methods such as decision trees.

Week 12 (Mar 27): Ethics

In Week 12, we'll go through a series of additional advanced topics. These will be focused around responsible and ethical practices and include discussions of issues such as being cognizant of confounding variables and the importance of study design, among others.

Week 13 (April 3): FINAL PRESENTATIONS

The final class (Week 13) will be dedicated to the STA130 final poster session for the Capstone Projects!