# SURVEYS, SAMPLING, AND OBSERVATIONAL DATA

STA304 is an upper-level undergraduate course at the University of Toronto's Department of Statistical Sciences.

---

## Contents


Preamble
- Overview
- How to succeed
- How we'll work
- Advice from past students
- Acknowledgements
- Previous versions

Content
- Week 1
- Week 2
- Week 3
- Week 4
- Week 5
- Week 6
- Week 7
- Week 8
- Week 9
- Week 10
- Week 11
- Week 12

Assessment
- Summary
- Quizz
- Tutorial
- Paper #1
- Paper #2
- Paper #3
- Paper #4
- Final Paper


# Preamble

## Overview

> *The best thing about being a statistician, is that you get to play in everyone's backyard.*

> *The best thing about being a statistician, is that you get to play in everyone's backyard.*
>
> *John Tukey*

The work of applied statisticians, regardless of their specific job title and area of application, is the most important and exciting work in the world right now. The ability to gather data, analyse it, and communicate your understanding of the underlying process is incredibly valuable. In this course you will learn and apply the essentials of this.

We focus on surveys, sampling and observational data. The very stuff of statistical science! We will approach these topics from a practical perspective. You will actually run surveys and learn how messy it is to put one together. You will learn how to think about sampling, how to implement it, and why the details matter. You will forecast an election. And you will conduct original research. More generally, you will learn how to obtain and analyse data and use it to make sensible claims about the world.

To work as an applied statistician requires you to be able to, as part of a small team:

- Gather data in less-than-perfect settings.
- Efficiently prepare and clean data toward some purpose.
- Analyse it in a reproducible, thorough, modern, and statistically-mature manner.
- Communicate your analysis to stakeholders including colleagues and clients with and without formal statistical training.

You likely have some of these skills already. This course will further develop them. At the end of the course you will have a portfolio of work focused on surveying, sampling, and observational data, that you could show off to a potential employer.

Each week you will read relevant papers and books, engage with them through discussion with each other, myself, and the TA. You will bring this all together and show off how much you have learnt through practical, on-going, assessment.

It is important to recognise that putting together everything that you have learnt to this point in this way will be difficult. It is not possible to cover everything that you will need to know. You should proactively identify and address aspects where you are weak through seeking additional information and resources. This course acts as a guide as to what is important, it does not contain everything that is important.

This course is different to many other courses at the University of Toronto. At the end of this course, you will have a portfolio of work that you could show off to a potential employer. You will have developed the skills to work successfully as an applied statistician or data scientist. And you will know how to fill gaps in your knowledge yourself. A lot of scholarships and jobs these days ask for GitHub and blog links etc to show off a portfolio of your work. This is the class that gives you a chance to develop these. It's very important to having something to show that needs to go beyond what is done in a normal class.

## How to succeed

In this course you will work in a self-directed, open-ended manner. Identify relevant areas of interest and then learn the skills that you need to explore those areas.

To successfully complete this course, you should expect to spend a large portion of your time reading and writing (both code and text). Deeply engage with the materials. Find a small study group and keep each other motivated and focused. At the start of the week, read the course notes, all compulsory materials and some recommended materials based on your interest. After doing that, but before the 'lecture' time you should complete the weekly quiz. During 'lectures' I'll live-code, discuss materials in the course notes, talk about an experiment, and you'll have a chance to discuss the materials with me.

You need to be more active in your learning in this course than others - read the notes and related materials - and then go out there and teach yourself more and apply it. You will not be spoon-fed in this course. Each week try to write reproducible, understandable, R code surrounded by beautifully crafted text that motivates, backgrounds, explains, discusses and criticizes. Make steady progress toward the assessment.

This is not a 'bird course'. Typically, after the term is finished, students say that the course is difficult but rewarding. The TAs and I are always available to answer any questions. Please come to office hours!

## How we'll work

This webpage will provide almost all the guiding materials that you need and links to the relevant parts of the notes. The course notes are available here: https://www.tellingstorieswithdata.com. Those contain notes and other material that you could go over. We'll use Quercus really only for assessment submission and grading.

A rough weekly flow for the course would be something like:

1. Read the week's course notes.
2. Read/watch/listen to the required materials.
3. Attend the lecture.
4. Attend the lab.
5. Complete the weekly quiz.
6. Make progress on a paper.

## Advice from past students

Successful past students have the following advice (completely unedited by me):

- "Start reading and writing on a weekly basis, watch some videos on R and RMD but more importantly learn how to use Google."
- "It is not a wise idea to take this course if you did not take any other STA 300 level course before."
- "Start early, find a group of people you trust enough to divide the work up fairly. Let people work to their strengths (people who know R should do the modelling, good writers should write most of the reports, etc.)"
- "Not to worry if you don't do well on the first problem set—the nature of the course is to build up skills overtime, and it's meant to be challenging in the beginning. In the end, it is worth it because you learn very valuable applicable skills on how to write professional reports."
- "Work on your writing and direction following skills."

- “Look at the rubric. There were times that I lost marks because I didn’t follow the rubric properly. Go to office hours, they are very useful as you can ask your own question and also get answers to questions other people ask and you didn’t think of. Also, do the assignments to the best of your ability. You will lose marks if you don’t put in effort and the only person you’re hurting is yourself.”
- “During lectures, focus more on the why the prof is doing what he’s doing. When he runs certain commands in R, figure out why that sequence of code gives what you want, because it’ll help adapt his code into your assignment code. just remembering what he’s doing in lecture becomes useless really quickly since the thought process matters more. also, start everything early.”
- “Do this course when you really want to learn something and have a lot of time to working on it.”
- “you need to be very skillful in RStudio and latex. Otherwise you would be struggling.”
- “Try to incorporate the feedback given and read a looottttttttt. Also start early on the problem sets because they tend to take a lot of time. Don’t give up!”
- “-Find a good group for problem sets”
- “If the assignments stay the same, I would tell students to approach this class from the perspective of ‘storytelling with statistics’ rather than a statistics course. You need to use R, and Markdown, and have a solid understanding of concepts like regression and sampling, but more importantly you need to be able to interpret results and write about them in a way coherent and professional way.”
- “do your readings”
- “Definitely get ready to write reports”
- “Do not take sta304 with Prof Rohan, it is pretty tough”
- “Start your work a bit earlier, make sure to follow the format expected and the rubric exactly.”
- “Read course material. Figure out WHY this paper/video is being shown to you and what you generally learn from it. Surround yourself with people dedicated to putting in the effort to understand material and who are thorough in their work so you can discuss content and/or work together.”
- “1. Be prepared to work extremely hard (8-11 hours a week). 2. Learn RStudio before course begins–STA130 is ideal preparation. 3. Start problem sets as soon as they are released.”
- “learn to code early and extensively use the office hours with the prof.”
- “This course requires lots of time dedicated and is not an”easy bird course" but is an incredibly rewarding course if one wants to learn how statistics is applied in the real world."

## Acknowledgements

Thank you to the following people for generously providing comments, references, suggestions, and thoughts that directly contributed to this outline: Bethany White, Dan Simpson, Jesse Gronsbell, Kelly Lyons, Lauren Kennedy, Monica Alexander and Uzair Mirza. Thank you especially to Samantha-Jo Caetano who influenced all aspects of this and co-taught the first version in Fall 2020.

### Previous versions

2020.

## Content

(Exact coverage will change based on how the class progresses.)

### Week 1

- Content:
  - Introduction
  - Several end-to-end worked examples
- Recording: https://youtu.be/MM8-71Q2rpo

### Week 2

- Content:
  - R essentials
  - Reproducible workflows
- Recording: https://youtu.be/EgOrUk752Fw

### Week 3

'Communicating'.

- Content:
  - Writing
  - Static communication
  - How to make a website and use Shiny.

### Week 4

'Gathering data'.

- Content:
  - Using APIs, scraping, OCR, semi-structured datasets, and text.
  - Observational data; correlation vs. causation, missing data, sources of bias.

## Week 5

'Hunting data'.

- Content:
  - Experiments, sampling and surveys, and A/B testing.
  - Design of surveys, sources of bias, randomized response surveys.
  - Techniques of sampling; stratification, clustering, unequal probability selection.
  - Sampling inference, estimates of population mean and variances, ratio estimation.

## Week 6

'Cleaning and preparing data'.

- Content:
  - Workflow for cleaning data.
  - Effective naming, checks, and testing.

## Week 7

'Storing and retrieving data' and 'disseminating and protecting data'.

- Guest lecture: Chris Henry, Bank of Canada
  - Christopher Henry (Chris) is a Senior Economist at the Bank of Canada. He serves as lead economist for the consumer survey research program on the Currency Department's Economic Research and Analysis team. Chris first joined the Bank as a Research Assistant in 2012, and recently rejoined in 2021 after completing his PhD in Economics. In his role, Chris contributes to the design, implementation, and analysis of a range of surveys that measure the use of cash and alternative methods of payment. He holds an PhD in Economics from Université Clermont Auvergne (France), and an MSc in Mathematics from McMaster University.
- Content:
  - R packages for data, and documentation including datasheets.
  - Personally identifying information, hashing and salting, GDPR and HIPPA, simulated data, and differential privacy.

## Week 8

'Exploratory data analysis'.

- Content:
  - Coming to terms with a dataset and understanding what is in it.

### Week 9

'IJALM - It's Just A Linear Model'.

- Content:
  - Simple linear regression, multiple linear regression, logistic regression, Poisson regression.

### Week 10

'Multilevel regression with post-stratification'

- Content:
  - MRP

### Week 11

'Causality from observational data'.

- Content:
  - Matching and difference in differences.
  - Regression discontinuity and instrumental variables

### Week 12

'Using the cloud' and 'Deploying models'

- Content:
  - R packages for models, Shiny, and Plumber and model APIs

## Assessment

### Summary

| Item | Weight (%) | Due date |
|---|---|---|
| Quiz | 20 | Weekly before the lecture |
| Tutorial | 20 | Weekly the day before the tutorial |
| Paper 1 | 25 | End of Week 3 |
| Paper 2 | 25 | End of Week 6 |
| Paper 3 | 25 | End of Week 8 |
| Paper 4 | 25 | End of Week 10 |

| Item | Weight (%) | Due date |
|---|---|---|
| Paper 4 | 25 | End of Week 10 |
| Final Paper (initial submission) | 1 | Middle of Week 12 |
| Final Paper (peer review) | 4 | End of Week 12 |
| Final Paper | 25 | Two weeks after that |

You must submit Paper 1. And you must submit the Final Paper.

Beyond that, you have scope to pick an assessment schedule that works for you. We will take your best 3 of the 11 tutorials, or your best 8 of 11 quizzes for that 20 per cent—whichever results in a better grade for you (i.e. you can choose to do either quizzes or tutorials). And we will take your two best papers from Papers 1-4 for that 50 per cent (25 per cent for each). The remainder is made up of 1 per cent for submitting a draft of the Final Paper, 4 per cent for peer reviewing other people's drafts of the Final Paper, and 25 per cent for the Final Paper.

Additional details:

- Quiz questions are drawn from those in the Quiz section that follows each chapter of *Telling Stories with Data*. Almost all of them are multiple choice, and you should expect to know the mark within two days of submission.
- Tutorial questions are drawn from those in the Tutorial section that follows each chapter of *Telling Stories with Data*. The general expectation (although this differs from week to week) is about two pages of written content, which the tutor will read, discuss with you, and then provide a mark. You should expect to know the mark within three days of of the tutorial.
- In general papers require a considerable amount of work, and are due after the material has been covered in quizzes and tutorials (i.e. you would draw on knowledge tested in the quizzes, and potentially material could be re-used from the tutorial material). In general, they require original work to some extent. Papers are taken from the Papers appendix of *Telling Stories with Data* and students have access to the grading rubrics before submission.

## Quizz

- You should choose to do either tutorials or quizzes.
- Due date: Weekly before the lecture.
- Weight: 20 per cent. Only best eight out of eleven count and only if that is better for you than counting tutorials.
- Task: Please complete a weekly quiz in Quercus.

## Tutorial

- You should choose to do either tutorials or quizzes.
- Due date: Weekly the day before the tutorial.
- Weight: 20 per cent. Only best three out of eleven count and only if that is better for you than counting quizzes.

- Task: Please complete a tutorial question and submit it via Quercus.
- Rubric:
  - 0 - Any typos, major grammatical errors, other table stakes issues for this level.
  - 0.25 - Grammatical errors, if relevant: tables/graphs not properly labeled, no references, other aspects that affect credibility. Too short.
  - 0.6 - Makes some interesting and relevant points, related to course material (including required materials), but lacking in terms of structure and story/argument.
  - 0.80 - Interesting paper that is well-structured, coherent, and credible.
  - 1 - As with 0.80, but exceptional in some way.

## Paper #1

- You must submit this paper.
- Task: 'Mandatory Minimums' (details will be added to Quercus).
- Due date: End of Week 3.
- Weight: 25 per cent (for Papers #1-#4 the best two of four count).

## Paper #2

- Due date: End of Week 6.
- Task: 'The Short List' (details will be added to Quercus).
- Weight: 25 per cent (for Papers #1-#4 the best two of four counts).

## Paper #3

- Due date: End of Week 8.
- Task: TBA (details will be added to Quercus).
- Weight: 25 per cent (for Papers #1-#4 the best two of four counts).

## Paper #4

- Due date: End of Week 10.
- Task: TBA (details will be added to Quercus).
- Weight: 25 per cent (for Papers #1-#4 the best two of four counts).

## Final Paper

## Final Paper

- Task: TBA (details will be added to Quercus).
- You must submit this paper.
- Due dates:
  - Initial submission: Middle of Week 12.
  - Peer review: End of Week 12.
  - Final Paper: Two weeks after that.
- Weight: 30 per cent
  - Initial submission: 1 per cent
  - Peer review: 4 per cent
  - Final Paper: 25 per cent