Surveys, Sampling, and Observational Data

Rohan Alexander

31 January 2022

Contents

1 Preamble

1.1 Overview

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.

1.2 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!

1.3   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.

1.4   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."

1.5 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.

2 Content

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

2.1 Week 1

- Content:
  - Introduction
  - Several end-to-end worked examples

2.2 Week 2

- Content:
  - R essentials
  - Reproducible workflows

2.3 Week 3

'Communicating'.

- Content:
  - Writing
  - Static communication

2.4 Week 4

'Gathering data'.

- Content:
  - How to make a website and use Shiny.
  - Using APIs, scraping, OCR, semi-structured datasets, and text.
  - Observational data; correlation vs. causation, missing data, sources of bias.

2.5   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.

2.6   Week 6

'Cleaning and preparing data'.

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

2.7   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.

2.8   Week 8

'Exploratory data analysis'.

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

2.9   Week 9

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

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

2.10   Week 10

'Multilevel regression with post-stratification'

- Content:
    - MRP

2.11 Week 11

'Causality from observational data'.

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

2.12 Week 12

'Using the cloud' and 'Deploying models'

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

3 Assessment

3.1 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 4
Paper 2	25	End of Week 6
Paper 3	25	End of Week 8
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 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.

3.2 Quiz

- 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.

3.3 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.

3.4 Paper #1

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

3.5 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).

3.6 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).

3.7 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).

3.8 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

4 Other

4.1 Learning objectives

- Design a survey or sample that appropriately gathers information of interest.
- Carry out a variety of statistical analyses in R to make inference on the data collected from a survey/sample.
- Identify and implement different sampling techniques and different study designs and the trade-offs involved in each.
- Identify sources of bias within a study and comment on a study's design, including weaknesses, strengths, and appropriate analyses.
- Clearly communicate results of statistical analyses to technical and non-technical audiences.

4.2 Communication

If you have a question, there is a decent chance that others have the same question or, at least, will benefit from the answer. Please post all questions to Piazza so that everyone in the course can benefit from your questions and our answers. You are encouraged to post answers to the questions of other students, where appropriate. Of course, if you have a concern of a personal nature then please email the TAs or me and you should begin your subject line with the course code 'STA304', and then an appropriate subject.

Emails and the message board are not checked or responded to by either the TA or me after hours or on the weekend.

Please be polite. We continue to be in a pandemic.

4.3 Accommodations with regard to assessment

You do **not** need to reveal your personal or medical information to me. I understand that illness or personal emergencies can happen from time to time. The following accommodations to assessment requirements exist to provide for those situations.

Straight-forward (will automatically apply to all students - there's no need to ask for these):

- Quiz: Worst three quizzes are dropped.
- Tutorial: Worst eight tutorials are dropped.
- Papers #1-#4: Worst two are dropped.

So for those (with the exception of Paper #1), if you have a situation, then just don't submit.

Slightly more involved:

- Paper #1: I'm open to a day without penalty to account for situations. Beyond that it begins to slow down the class. You must submit something for Paper #1.
- Peer review: No accommodation or late submission is possible for this because it would hold up the rest of the class. If you cannot submit then email me before the deadline and the weight will be shifted to the final paper.
- Final paper: The final paper is a critical piece of assessment. It's also up against deadlines for submission of grades. Extensions for valid reasons may be granted for a maximum of three days, however this isn't possible for all students (i.e. there may be restrictions around graduating students). Hence, the

exact extension needs to be at my discretion. To be considered, an extension request must be sent to rohan.alexander@utoronto.ca by the business day before the due date so there is time to get advice from the Department and your college about your particular circumstance.

4.4 Minimum submission requirement

If you are going to not be able to submit at least two problem sets, and/or be unable to submit the final paper then it would be unfair on the other students to allow you to pass the course. Please ensure you and your registrar get in touch with me as early as possible if this may be the case for you.

4.5 Re-grading

Requests to have your work re-graded will not be accepted within 24 hours of the release of grades. This is to give you a chance to reflect. Similarly, requests to have your work re-graded more than seven days after the release of the grades will not be accepted. This is to ensure the course runs smoothly.

Inside that 1-7 day period if you would like to request a re-grade, please email rohan.alexander@utoronto.ca with a subject line that starts with 'STA304'. You must specify where the marking error was made in relation to the marking guide. Your entire assessment will be re-marked and it is possible that your grade could reduce.

4.6 Plagiarism and integrity

Please do not plagiarize. In particular, be careful to acknowledge the source of code - if it's extensive then through proper citation and if it's just a couple of lines from Stack Overflow then in a comment immediately next to the code.

You are responsible for knowing the content of the University of Toronto's Code of Behaviour on Academic Matters.

Academic offenses include (but are not limited to) plagiarism, cheating, copying R code, communication/extra resources during closed book assessments, purchasing labour for assessments (of any kind). Academic offenses will be taken seriously and dealt with accordingly. If you have any questions about what is or is not permitted in this course, please contact me.

Please consult the University's site on Academic Integrity http://academicintegrity.utoronto.ca/. Please also see the definition of plagiarism in section B.I.1.(d) of the University's Code of Behaviour on Academic Matters http://www.governingcouncil.utoronto.ca/Assets/Governing+Council+Digital+Assets/Policies/PDF/ppjun011995.pdf. Please read the Code. Please review Cite it Right and if you require further clarification, consult the site How Not to Plagiarize http://advice.writing.utoronto.ca/wp-content/uploads/sites/2/how-not-to-plagiarize.pdf.

4.7 Late policy

You are expected to manage your time effectively. If no extension has been granted and no accommodation applies, then the late submission of an assessment item carries a penalty of 10 percentage points per day to a maximum of one week after which it will no longer be accepted, e.g. a problem set submitted a day late that would have otherwise received 8/10 will receive 7/10, if that same problem set was submitted two days late then it would receive 6/10.

4.8 Writing

Papers and reports should be well-written, well-organized, and easy to follow. They should flow easily from one point to the next. They should have proper sentence structure, spelling, vocabulary, and grammar. Each point should be articulated clearly and completely without being overly verbose. Papers should demonstrate your understanding of the topics you are studying in the course and your confidence in using the terms, techniques, and issues you have learned. As always, references must be properly included and cited. If you

have concerns about your ability to do any of this then please make use of the writing support provided to the faculty, colleges and the SGS Graduate Centre for Academic Communication.

4.9   Accessibility needs

Students with diverse learning styles and needs are welcome in this course. In particular, if you have a disability/health consideration that may require accommodations, please feel free to approach me and/or Accessibility Services at 416 978 8060 or visit studentlife.utoronto.ca/as.

4.10   Intellectual Property Statement

Course material that has been created by your instructor is the intellectual property of your 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.