# STA304 (L0201): Surveys, Sampling and Observational Data[1]

## Calendar items

**Description:** 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. Observational data; correlation vs. causation, missing data, sources of bias.
**Prerequisite:** ECO220Y1/ ECO227Y1/ GGR270H1/ PSY201H1/ SOC300H1/ SOC202H1/ STA220H1/ STA255H1/ STA261H1/ STA248H1/ STA238H1/ STA288H1/ EEB225H1/ STAB22H3/ STAB57H3/ STA220H5/ STA258H5/ STA260H5/ ECO220Y5/ ECO227Y5.
Exclusion: STAC50H3, STAC52H3, STA304H5.

## Details

**Instructor:** Rohan Alexander - rohan.alexander@utoronto.ca - (you are welcome to call me 'Rohan' ('row'-'hun')).
**TAs:** Kevin Zhang, Marija Pejcinovska, and Xiaomeng Ma.
**Term:** Fall 2020.
**Mode:** Online with all recordings made available via the course website and Quercus for students unable to join.
**Course website:** https://rohanalexander.com/sta304.html and Quercus. Please check these regularly.
**Course discussion forums:** Quercus/Piazza.
**Lecture times:** 9am-11am on Thursdays (lectures recorded and posted to the course website/Quercus).
**Lab times:** 11am-noon (labs recorded and posted to Quercus).
**Office hours:**

- Rohan: I am available between noon and 2pm on Thursdays. Please use the following link to book an appointment: https://calendly.com/rohanalexander/10min.

[1] 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, and Monica Alexander. Thank you especially to Samantha-Jo Caetano who influenced all aspects of this.

- TAs: Office hours will occur online for an hour immediately following the lab i.e. between noon and 1pm on Thursdays.

# 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.

# Purpose

*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 the 2020 US Presidential 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.

# 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 Quercus/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're in the middle of a pandemic.

# Late policy

Please try to manage your time effectively. If no extension has been granted for medical reasons, 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. For instance, 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.

There are some assessment items where late submissions (apart from where an accommodation applies) cannot be accepted. This is because it would be unfair to other students (in the case of peer review) or it would be unfair to yourself (trying to forecast the US election *ex post* is a very different task).

# Accommodations

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 apply in these situations.

## Problem sets

If a problem set is missed for a valid reason, you may ask to be excused from that problem set. Extensions will not be given. If approved, the weight of the missed problem set will be shifted to the other problem sets. To request to be excused from a problem set, please email rohan.alexander@utoronto.ca. For consideration, your email:

- must be received within one day of the due date for the missed problem set (ideally before),
- must include your full name and student number,
- must specify the problem set missed including the date, and
- must include the following two sentences:
    1. 'I affirm that I am experiencing an illness or personal emergency and I understand that to falsely claim so is an offence under the Code of Behaviour on Academic Matters.'
    2. 'I understand that the weight of this problem set will be shifted to the other problem sets.'

No more than two of the problem sets will be accommodated in this way.

## Test

If the test is missed for a valid reason, you may ask to be excused from the assessment. Extensions cannot not be given as it would not be fair to the other students. If approved, the weight of the missed test will be shifted to the final

paper. To request to be excused from the test, please email rohan.alexander@utoronto.ca. For consideration, your email:

- must be received within one day of the missed test (ideally before),
- must include your full name and student number,
- must specify that the test was missed including the date, and
- must include the following two sentences:
    1. 'I affirm that I am experiencing an illness or personal emergency and I understand that to falsely claim so is an offence under the Code of Behaviour on Academic Matters.'
    2. 'I understand that the weight of the test will be shifted to the final paper.'

### Final paper

The final paper is a critical piece of assessment. Extensions for valid reasons may be granted for a maximum of three days. The exact extension granted will be at the discretion of the instructor.

To be considered, an extension request must be sent to rohan.alexander@utoronto.ca by the business day before the due date.

Where possible, please alert me to potential issues as early as you can. This will allow me to work with you to find a suitable solution.

### Minimum submission requirement

If you are going to miss more than 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. If such an amount of work were missed, even for valid reasons, then an oral exam may be required to calculate a fair mark, at the discretion of the instructor. Please ensure you and your registrar get in touch with me as early as possible if this may be the case for you.

## 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.

# Writing

Communication and especially writing is a critical aspect of the statistical workflow. Papers 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. You will be heavily penalised for papers that do not meet these basic requirements.

Papers should demonstrate your understanding of the material you have learnt and your confidence in drawing on the terms, techniques, and issues you have considered. Your work must be thoroughly referenced.

If you have concerns about your ability to do any of this, then please make use of the writing support provided to students - https://writing.utoronto.ca/. The services are designed to target the needs of both native and non-native speakers and the programs are free. I have used similar services in the past at other universities and always found them very helpful.

# COVID-19

We are in the middle of a pandemic. This term will be a difficult one for you, but also for everyone involved - your TAs and me, faculty, staff, and of course the other students in your courses. Nonetheless, in this course I want to provide you with an opportunity to do the best work of your life, to learn, and to contribute. Some degree of flexibility and good faith is needed from all of us. If you need accommodations then please be as proactive as possible in asking for them.

# Assessment

This is a guide to the assessment. More details about each of these will be released on the course website.

Problem Sets 2, 3, and 4, occur in groups. You are welcome to form your own groups, but if you cannot then you will be allocated a group two weeks before the due date. Everyone in the group will get the same mark. The size of the group is not taken into account in marking.

If there are technical difficulties that prevent you from completing the test then the test accommodation will apply and the weight will be transferred to the Final Paper.

| **Item** | **Weight (%)** | **Due date** |
|---|---|---|
| Problem Set 1 | 10 | 27 September 2020 |
| Problem Set 2 | 15 | 7 October 2020 |
| Problem Set 3 | 15 | 16 October 2020 |
| Problem Set 4 | 20 | 2 November 2020 |
| *9-13 November 2020, Fall reading week.* | | |
| Test | 10 | 19 November 2020 |
| Final Paper | 30 | During Final Assessment Period - exact date TBA |

# Core texts

There is no one textbook and there is no exact prescription for what you must learn. You should be guided by the topics, your interests, and background, of course within the constraint of the assessment that you must complete, although that is fairly broad. The main textbooks that we draw on are:

1. Gelman, Andrew, Jennifer Hill and Aki Vehtari, 2020, *Regression and Other Stories*, Cambridge University Press.
2. Kohavi, Ron, Diane Tang, and Ya Xu, 2020, *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*, Cambridge University Press.
3. McElreath, Richard, 2020, *Statistical Rethinking*, 2nd Edition, CRC Press.
4. Wu, Changbao and Mary E. Thompson, 2020, *Sampling Theory and Practice*, Springer.

If you only have a little statistics, then get Gelman, Hill, and Vehtari. You should be able to get it for a little less than $50 and it covers most of the basics. If you have some statistics or have plans to get into it, then get McElreath. McElreath is very expensive, at around $100 or so. Unfortunately there's no way around it, as it is a very strong textbook at the moment. If you can get a cheap first edition then that is fine and the lecture notes will detail the main differences. We only use Kohavi, Tang, and Xu for the week focused on A/B testing, but it's quite popular

in industry at the moment and if you want to buy it you should be able to get it for around $40. Finally, Wu and Thompson provides the statistical basis for the sampling in the course. It is also very expensive, but you can get a free PDF legally through the U of T library (search Springer, go to the link, then search for the book).

Additionally, for each topic, in the lecture notes I will include additional material that you may like to consider if a topic is of particular interest.

# 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; studentlife.utoronto.ca/as.

# Plagiarism and integrity

Please do not plagiarise. I will not tolerate it, and will pursue it to the full extent that I can.

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

As a general rule, I 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 R code or written answers for anything that is to be handed in. For example, "For question 2.1 what R function did you use?" is a fair question when discussing course material with others in the class; "Please show me your R code for question 2.1" 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. When asked to hand in code and a problem set or project document, the code you submit must have been used to generate the document. If it does not (i.e., the submitted code does not match the submitted output), this is also considered an academic offense.

I will not tolerate any academic offenses. This includes (but is 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 very seriously and dealt with accordingly. If you

have any questions about what is or is not permitted in this course, please do not hesitate to contact your instructor.

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.

Note that when an assignment is required to be completed as a team (e.g., project), you may discuss and share answers and code with other members of your team, but not with another team in the class or anyone outside the course.

# Intellectual Property Statement

Course material that has been created by your instructor (i.e. lecture slides, term test questions/solutions and any other course material and resources made available to you on Quercus) 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. While recordings of class meetings will be made available to you on the course website, these are intended only for students registered in the course. You are not authorized to copy these materials or distribute them to individuals who are not registered in the course. If you would like to record any course activities in this course, you MUST ask permission from your instructor in advance. According to intellectual property laws, not asking permission constitutes stealing.

# Recognized Study Groups

Recognized Study Groups (RSGs) are small study groups of 3 to 6 students from the same course who meet weekly to learn course content in a collaborative environment. Each group is made up of students from the same course. One student volunteers to be the RSG Leader and helps organize and plan weekly activities. The RSG Leader is a student who is trained in group facilitation and effective learning techniques. RSG Leaders are not tutors – they are learning

along with group members. A student staff member is also assigned to each group to help connect you to academic resources and support your group's goals. While not compulsory for this course, I would recommend you get involved with an RSG.