

STA303: Methods of Data Analysis II
Course guide

Prof. Liza Bolton

Winter 2022

Contents

1	How to use this course guide	3
2	Syllabus	4
2.1	Course information	4
2.2	Land acknowledgment	5
2.3	Prerequisites	6
2.4	Course format and organization	6
2.5	Learning objectives	6
2.6	Textbooks	6
2.7	Computing and minimum technical requirements	7
2.8	Course outline	7
2.9	Assessments	8
2.10	Hours expectations	9
2.11	Marking concerns / regrade requests	9
2.12	Missed work policies	10
2.13	Communication policy	10
2.14	Accommodations and accessibility	11
2.15	Recognized study groups	12
2.16	Meet to complete	12
2.17	Feeling distressed?	12
2.18	Intellectual property statement	12
2.19	Academic integrity	13
2.20	Course design principles	14
3	Start here!	15
3.1	Introductions	15
3.2	How this course works	16
3.3	Hours expectations	17

<i>CONTENTS</i>	3
Assessments	19
4 Assessment overview	19
4.1 Graduate student modification (1002H)	19
5 Mini-portfolio	20
6 Portfolio	21
7 Mini-mixed assessment	22
8 Mini-portfolio	23
9 Mixed assessment	24
10 Knowledge basket: Writing and peer feedback	25
10.1 General instructions	25
10.2 Module 1 writing task	26
11 Knowledge basket: Professional development task	28
11.1 Professional development proposal	29
11.2 Professional development evidence and reflection	34
12 Knowledge basket: Other	38
12.1 ‘Getting to know you’ survey	38
12.2 Pre-knowledge check	38
Modules	41
13 Module 1	41
13.1 Instructor information	41
13.2 Upward management tips	41
13.3 Recap of linear models	43
13.4 Why model?	43
13.5 Linear models	43
13.6 Linear regression assumptions	43
13.7 What makes it a <i>linear</i> model?	43
13.8 Optional refresher reading	44
13.9 Common statistical tests as linear regression	44
14 Module 2	55
15 Module 3	56

16 Module 4	57
17 Module 5	58
References	59
 Appendix	 61
18 Resources	61
18.1 Course tools overview	61
18.2 Using RStudio with the JupyterHub	62
18.3 Zoom, Zoom, Zoom, Zoom...	62
18.4 Student support services and resources	63
 19 FAQs and Errata	 65
19.1 Frequently asked questions	65
19.2 Other	67
19.3 Errata	67
 20 Bits and pieces	 68
20.1 Code to generate course art	68
20.2 M1 supporting information on matrices (not assessed)	69

THIS SITE IS STILL IN PROGRESS! The information is not yet official.

1

How to use this course guide

This course guide has been created using [bookdown](#). You can download a PDF version of the whole guide with the download button above (down arrow into a tray). Note that many part of this guide will be updated as the course proceeds. You can also put the website into dark mode and changed the font style, if you find a different display preferable.

If you would like a PDF copy of the slides, you can ‘Print to PDF’ in your browser. Shortcut: Cmd+P or Ctrl+P, and select ‘Save as PDF’ (or similar).

1.0.1 Communication policy reminder

All content and logistics questions must be asked on [Piazza](#). Personal or private course matters should be emailed to sta303@utoronto.ca. Quercus mail or emails sent directly to teaching team members will not be answered. If you’ve missed an assessment due to illness or emergency, please fill out the appropriate form.

2

Syllabus

You can get a more visual PDF version of the syllabus [here](#). This version is hopefully better suited to searching and screen readers.

University of Toronto, Department of Statistical Sciences
STA303/1002: Methods of Data Analysis II

2.1 Course information

2022

HALF YEAR, HALF CREDIT

2.1.1 Course description

The course focuses on using and interpreting advanced statistical methods with applications in a number of different areas. The overall theme of this course is dealing with situations where the assumptions of the regression models developed in STA302 may not apply. The course is a mixture of theory and application. Assignments will involve computing with R and there is a significant focus on written and oral communication.

2.1.2 Class times

Class: Wednesdays

- L0101: 10:00 a.m.–12:00 p.m. ET
- L0201: 3:00–5:00 p.m. ET

Tutorials: Alternating Thursdays

- 12:00–1:00 p.m. ET or
- 5:00–6:00 p.m. ET

Delivery

- Wednesday classes will be **online** and recorded.
- The Thursday tutorial time will alternate between group activities and TA office hours. They will be **online** for January, and an optional in-person stream for the fortnightly group activities is planned after that.
- **Synchronous attendance** is recommended but not required to complete the course.

2.1.3 Materials

All **materials** will be posted on [Quercus](#) and/or the [course guide](#).

Course discussion board: [Piazza](#)

2.1.4 Teaching team

2.1.4.1 Instructor

Liza Bolton

Pronouns: [she/her](#)

Email: sta303@utoronto.ca

Office hours: 2nd half of Wednesday classes

Please call me: Liza or Prof. Bolton/Prof. B

How do you pronounce that?

- Liza: [a video...](#)
- Bolton: like the words “bowl” + “tonne”

2.1.4.2 Head TA

Amin Banihashemi ([he/him](#))

Please call me: Amin (A-meen [æ mi n or ə mi n] [bæni h mi] decode [here](#))

Email: sta303@utoronto.ca

2.1.4.3 Teaching assistants

Vedant Choudhary, Shuang Di, Sonia Markes, Ian Richter, Xiaochuan Shi, Lei Sun, Liam Welsh, Dongyang Yang, Kevin Zhang, Robert Zimmerman

Office hours: Alternating Thursdays, see [Quercus](#)

Email: sta303@utoronto.ca

2.2 Land acknowledgment

The land on which our University operates is the traditional lands of the Anishinaabe, the Haudenosaunee, and the Mississaugas of the Credit. With the Dish With One Spoon treaty, these peoples agreed to share and protect this land, and all those who have come here since, both Indigenous and non-Indigenous, are invited into this treaty in a spirit of respect and peace. This land is also, more recently, subject to Treaty 13, a treaty between the Mississaugas and the British Crown.

In this course, we are coming together to discuss statistics, a field that has been part of historical and ongoing colonization, oppression, and harm of Indigenous peoples. Let us remind ourselves of our responsibilities to this land, its original peoples, and to each other and work to be ethical and culturally competent practitioners in our chosen fields.

We encourage you to consider the history of the land wherever you are. <https://www.whose.land/en/>

2.3 Prerequisites

STA302/1001

I.e., we will assume that you are familiar with running linear regression analyses, including checking assumptions and some of the mathematical reasoning behind the models. Material from the second-year statistical theory courses which are prerequisites to STA302 will be drawn on extensively. Knowledge of programming with R is essential.

2.4 Course format and organization

This course is composed of **five two-week modules** and **two assessment focus weeks**.

Each **module** has an “I show you, you show me” structure. It starts with you watching videos and/or doing readings and a guided code demonstration or other relevant class topics (“I show you”). This course is flipped, so the expectation is that you engage with some or all of the asynchronous content before Wednesday in the first week. Then you will do some or all of the following: a group problem-solving activity, practice quizzes, assessment activities.

During **assessment focus weeks**, no new content will be released.

2.5 Learning objectives

By the end of the course, you will be able to:

- **Wrangle** and **explore** a dataset
- Create appropriate data **visualizations**
- Describe **ethical considerations** in data analysis
- Understand the assumptions and appropriate use cases for **linear mixed models**, **generalized linear models**, **generalized linear mixed models**, and **generalized additive models**
- **Write** and **execute R code** for the model types covered
- Accurately and appropriately **interpret** the results of the model types covered and communicate these to a range of **audiences**

2.6 Textbooks

You do not have to purchase a textbook for this course. There are three texts that we will use extensively, and they are all freely available to you. Additional readings will be assigned as appropriate.

[Wickham. R for Data Science. 2019.](#)

[Legler and Roback. Broadening Your Statistical Horizons. 2019.](#)

[Wood. Generalized Additive Models: An Introduction with R, 2nd Edition. 2017.](#) (requires you to log in with your UTORid)

2.7 Computing and minimum technical requirements

We will be using [RStudio](#) to make reproducible data analysis reports using [R](#) and [R Markdown](#).

You can use RStudio on your personal machine or through the U of T JupyterHub: jupyter.utoronto.ca.

To participate in synchronous classes and office hours you will need a **U of T Zoom account**. If you do not yet have one, go to <https://utoronto.zoom.us/> to set one up. To participate fully, you will need Desktop client or mobile app: version 5.3.0 or higher or ChromeOS: version 5.0.0 (4241.1207) or higher. You can check your desktop client or mobile app version by following [these instructions](#).

All students should consult the [minimum technical requirements](#) for participation in online learning. If you are facing financial barriers to obtaining the required technology, please contact your [College Registrar's Office](#) to obtain information regarding your potential eligibility for a need-based bursary.

2.8 Course outline

The topics listed below are subject to change.

Module/focus	Important dates
Jan 10–21 Module 1: Welcome! <ul style="list-style-type: none"> How this course works, assessments, admin, upward management Common tests as linear regression and linear regression recap Reading strategy: previewing and skimming [Optional] Get ahead for module 2 if unfamiliar with R Markdown, ggplot and dplyr 	JAN 20 Prerequisite knowledge check due JAN 23 Last day to enrol in S courses
Jan 24–Feb 4 Module 2: Professional skills for data analysis <ul style="list-style-type: none"> Data visualization Data cleaning, merging, and exploration Statistical communication Ethical professional practice 	FEB 1 Lunar New Year FEB 2 Prerequisite check workshop FEB 3 Professional dev proposal due FEB 3 Mini-portfolio due
Feb 7–Feb 18 Module 3: Linear mixed models <ul style="list-style-type: none"> Identifying correlated data Fixed and random effects Fitting, visualizing and interpreting correlated data Likelihood ratio tests 	FEB 14 Valentine's Day FEB 17 Portfolio due
Feb 21–25 Reading Week: No class!	FEB 21 Family Day (U of T closed)
Feb 28–Mar 11 Module 4: Generalized linear models <ul style="list-style-type: none"> Theory for GLMs Fitting models, checking assumptions and interpreting logistic and Poisson regression 	MAR 9 Mini-mixed assessment (12 hour window, 8 a.m.–8 p.m.)

Module/focus	Important dates
Mar 14–18 Mixed assessment focus week	MAR 13 Daylight savings begins MAR 14 Last day to cancel course MAR 16 Mixed assessment (12 hour window, 8 a.m.–8 p.m.)
Mar 21–Apr 1 Module 5: GLMMS and GAMS	MAR 31 Professional dev evidence and reflection due
<ul style="list-style-type: none"> Generalized linear mixed models Generalized additive mixed models 	
Apr 4–8 Project Focus week	APR 7 Final project due (bonus) APR 11 Project (no bonus)

2.9 Assessments

There are two special elements of assessment/grade calculation in this course that are important to be aware of in planning your approach to it.

- *Two roads diverged in a yellow wood*¹: You can opt for Path A or Path B to get your final mark. I will calculate your marks along both paths and then assign you the higher of the two as your final mark.
- A-tisket, a-tasket², fill up your **knowledge basket**³:
 - Personalize which of these assessments you do based on your interests and skills you want to develop.
 - You can ‘max out’ your basket: just keep putting grades in until you get to 10% (Path A) or 5% (Path B).

Assessment	Path A	Path B	STA1002 ONLY	Due dates
Mini-portfolio	5	0	0	Feb 3
Portfolio	20	25	25	Feb 7
Mini-mixed assessment	5	0	0	Mar 9 (12-hour assessment window, 8:00 a.m.–8:00 p.m. ET)
Mixed assessment	20	25	25	Mar 16 (12-hour assessment window, 8:00 a.m.–8:00 p.m. ET)
Final project	45	45	50	Apr 7 (5% pt bonus) Apr 11 (no bonus)
Knowledge basket	5	5	0	Multiple
Total	100	100	100	

Everything ⁴ in STA303 is due at 3:03 p.m. ET

Note: All times in this course are in Eastern Time (Toronto time). Please note that daylight savings will begin in Canada on March 13. If you are not based in Canada this may change the time conversion for you. Please keep this in mind.

¹The Road Not Taken, by Robert Frost. But, dear traveller, you can in fact take both roads and then get whichever gives you the better mark. <https://www.poetryfoundation.org/poems/44272/the-road-not-taken>

²A-tisket, a-tasket by Ella Fitzgerald, <https://www.youtube.com/watch?v=1bgFkeDLpSI>

³“Whaowhia te kete mātauranga” is a Māori proverb meaning “Fill up the basket of knowledge”. Mātauranga is specifically traditional Māori knowledge. You can listen to the pronunciation here <https://www.massey.ac.nz/student-life/m%C4%81ori-at-massey/te-reo-m%C4%81ori-and-tikanga-resources/te-reo-m%C4%81ori-pronunciation-and-translations/whakatauk%C4%AB-m%C4%81ori-proverbs/>

⁴Mixed assessment windows are 8-8, I could make the 303 work out fairly.

2.9.1 Knowledge basket

The following components are guaranteed knowledge basket assessment options. Additional opportunities may be offered throughout the course (Team Up!, speaker series reflections, etc.)

Assessment	%	Due date
Pre-knowledge check: completion	0.5	Jan 20
Pre-knowledge check: 80%+ or workshop	0.5	Jan 20 (80%+ score) Feb 2 (workshop)
Professional development task proposal	1	Feb 3
Professional development task evidence & reflection	3	Mar 31
Writing & peer review (Create-Assess-Reflect) x 5	0.5 X 5	Friday-Tuesday-Friday each Module
Module check-in x 5	0.1 X 5	Last Friday of each Module

2.10 Hours expectations

While everyone has different work styles and learning needs, I want to provide some guidance around how I expect this course to look for you.

Plan to be doing 6–8 hours of work on STA303 each week. In a **two-week module**, this may look like:

- 2–4 hours on videos and readings
- 1–3 hours of attending synchronous class or reviewing the recording and activities
- 1–2 hours on knowledge basket assessments
- 2–6 hours on other assessments
- Remaining time attending reading announcements, office hours, checking Piazza, revision, etc.

2.11 Marking concerns / regrade requests

Any request to have an assessment remarked must be submitted to [the appropriate form](#) on the Forms page on Quercus under the following conditions:

- **Wait** 24 hours after the release of grades. Use this time to go over sample solutions and course materials. Wait 24 hours after the release of grades. Use this time to go over any sample solutions and feedback, the instructions, and relevant course materials.
- After the 24-hour period has finished, you will have one week to submit your regrade request. (I.e., one week + 24 hours total.)
- Your request must include a detailed and thoughtful justification referring to your answer and the relevant course material to be considered. Please note that I reserve the right to review the grading of all questions or parts when you re-submit an assessment for reconsideration (i.e., your grade could go down).
- You will receive a confirmation email upon submitting the form. Allow for two weeks for processing after the request window closes before following up.
- The specific timeline and requirements for **final project regrade requests** will be announced later.

Only answers in English (or appropriate code, mathematical symbols) can be accepted in this course. Answers submitted in a different language will receive a 0 and will not be eligible for regrading. If you have an autotranslation extension on your browser, be very careful about how this can interact with Quercus.

Please note that I reserve the right to review the grading of all questions or parts when you re-submit an assessment for reconsideration (i.e., your grade could go down).

2.12 Missed work policies

In general, late work is not accepted, without either:

- an extension approved **48 hours** before the due date, or
- a personal illness/emergency declaration no more than 3 days after the due date.

Please note that technical difficulties knitting an Rmd or getting the due time wrong do **not** constitute personal emergencies.

The following assessments are eligible:

- Mini-portfolio
- Portfolio
- Mini-mixed assessment
- Mixed assessment
- Professional development proposal
- Professional development evidence and reflection

Upon receipt of your request, we will contact you via email within 2 business days to confirm an accommodation, as appropriate.

2.12.1 Exceptions

- **Knowledge basket** assessments (other than the professional development task) are not eligible for extensions.
- There are no routine extensions granted for the **final project**. In exceptional circumstances, you can work with your College Registrar and me on this.

2.12.2 IMPORTANT NOTES

- If too much work is missed, even for valid reasons, an **oral exam** may be required to calculate a fair mark, at the discretion of the instructor. Please ensure you and/or your College Registrar get in touch with me as early as possible if this may be the case for you.
- If you have accommodation letters from an accessibility advisor, make sure you read the instructions in the ‘Accommodations and accessibility’ section.
- Unless discussed with your instructor first and an agreement is come to, if you submit an assessment, it will be assumed that you deemed yourself fit enough to do so and your grade will stand as calculated. No accommodation will be made based on claims of medical, physical, or emotional distress after the fact.

2.13 Communication policy

AKA How to get your questions answered

Following our course communication policy helps ensure you receive answers and supports in a timely fashion, and also shows respect for the teaching team’s time and effort. We reserve the right to ignore any correspondence that does not conform to this policy.

Course logistics? e.g.,

- What is the deadline for the final project?

- Where do I submit the assignment?

Course content? e.g.,

- Why do we sometimes use `glm()` and sometimes use `glmer()`?

- My code won't run for question #1

Info to share with class? e.g.,

- I have a link/resource/opportunity to share with my classmates

Missed an assessment due to illness or personal emergency?**Want to request a regrade of an assessment?****Personal/sensitive**

circumstances? (i.e., something which is not appropriate to share with the whole class)

PIAZZA FORUM

- Link in Quercus' course navigation menu.
 - Posts can be anonymous for your classmates, but instructors and TAs will be able to see your name.
 - Before posting a question, double-check the syllabus AND search to see if someone else has already asked a similar question (you can edit the question to add yours or post a follow-up at the bottom).
- Try to answer your classmates' questions—this is a great way to reinforce your own understanding while also helping your classmates! Don't worry if you aren't 100% sure of the answer—answers will be reviewed/endorsed/completed by the teaching team!

FORMS

- Use the appropriate form linked on the [Forms](#) page on Quercus.
- If you cannot meet a deadline because you are ill, please also refer to the **Missed Work** section in this syllabus. A doctor's note is not required, but if you have one you can upload it as supporting documentation.
- If you wish to request a regrade, please also refer to the **Marking Concerns** section in this syllabus. Be prepared to provide a detailed justification and possible supplementary materials. "I worked hard on this so I should get a better mark" is not an appropriate justification (yes, I do receive emails like that).

COURSE EMAIL: sta303@utoronto.ca

- Send emails from your utoronto.ca email address to ensure they don't automatically go to a Junk folder.
- Include your full name and UTORid.
- This account will be monitored by the head TA and course instructor; if you want to reach Prof. Bolton only please include [Prof. Bolton] in the **subject line**; do not email her directly about course matters.
- Allow at least 24 hours for a response during the week (Monday to Friday, ET) and do not expect responses on the weekend. Do not send a follow-up email until at least two business days (Toronto time) later.

NEVER send Quercus mail to the STA303 teaching team

NEVER use the 'Add Comment' option on Quercus. They will not read.

Please.

2.14 Accommodations and accessibility

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

- Email your letter to sta303@utoronto.ca with "Accommodation letter" as part of the email subject, CC your

advisor and let us know anything else you wish us to know/any questions you have. Please do this as soon as possible after you enroll in the course/receive this syllabus.

- Confirm any accommodations for each specific assessment **one week** before the assessment. (I.e., if you receive extra time for timed assessments, confirm this one week prior to the mixed assessments, even if we have already discussed this at the beginning of the semester.)

2.14.1 Accessibility services

The University of Toronto is committed to accessibility. If you require accommodations for a disability or have any accessibility concerns about the course or course materials, please contact Accessibility Services as soon as possible: email accessibility.services@utoronto.ca or visit the website at [<http://accessibility.utoronto.ca>](http://accessibility.utoronto.ca).

2.14.2 Religious Accommodation

At the University of Toronto, we are part of a diverse community of students, staff, and faculty from a wide range of cultural and religious traditions. For this course, I have sought to avoid scheduling compulsory activities in ways that will clash with religious holy days (not captured by statutory holidays). If you anticipate missing a course activity due to a religious observance, please let me know as early in the course as possible. With sufficient notice—ideally at least three weeks—we can work together to make alternate arrangements.

2.15 Recognized study groups

I would highly recommend you [get involved with an RSG](#). 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.

2.16 Meet to complete

Meet to Complete is an online “study with me” space where you can study alongside other students. To join Meet to Complete, enroll in the [Meet to Complete course on Quercus](#). Learning, even online, doesn’t need to be lonely!

2.17 Feeling distressed?

You may find yourself feeling overwhelmed, depressed, or anxious. Lots of people feel the same way. There is help available from mental health professionals 24 hours a day via online and phone-based services listed on [this page in the course guide](#), as well as a range of other helpful U of T and community resources.

Accessibility Services (see above) also provides supports for mental health concerns.

2.18 Intellectual property statement

Course material that has been created by your instructor (i.e., lecture slides, questions/solutions, and any other course material and resources made available to you) is the intellectual property of your instructor (or the credited holder of the copyright) 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. 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.

2.19 Academic integrity

2.19.1 Plagiarism

You may be at risk of plagiarizing if you do not understand the rules and your responsibilities. You must not present the work of others as your own. This includes, but is certainly 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:

- <https://www.academicintegrity.utoronto.ca/perils-and-pitfalls/>
- <https://www.academicintegrity.utoronto.ca/smart-strategies/>
- [This video](#) will be assigned later in the course.

YOU are responsible for knowing the content of the [University of Toronto's Code of Behaviour on Academic Matters](#). The University of Toronto treats cases of academic misconduct very seriously. Academic integrity is a fundamental value of learning and scholarship at the U of T. Participating honestly, respectfully, responsibly, and fairly in this academic community ensures that your U of T degree is valued and respected as a true signifier of your individual academic achievement.

Other potential offences include, but are not limited to:

- Looking at someone else's answers.
- Letting someone else look at your answers.
- Misrepresenting your identity.
- Falsifying or altering any documentation required by the University.
- Falsifying institutional documents or grades.

All suspected cases of academic dishonesty will be investigated following the procedures outlined in the Code of Behaviour on Academic Matters. If you have any questions about what is or is not permitted in this course, please do not hesitate to contact me.

2.19.2 Specific advice on untimed assessments

As a general rule, for untimed assessments, I encourage you to discuss course material with each other and ask others for advice. However, it is **not permitted** to share R code or written answers for anything that is to be handed in. For example, "For question 2 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" 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 offense. **When asked to hand in code and the output it creates, 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.

2.19.3 Rules for timed assessments (e.g., mixed assessments)

While all timed assessments in this course are open-book, they are not "open-person". You **MUST NOT** discuss any details of the assessment with anyone else during the assessment window, regardless of your completion status. This includes, but is not limited to, current classmates, friends, and tutors. For example, even asking someone "which slide did you look at to answer question 3" is not appropriate for timed assessments.

2.19.4 NOTE: BE CAREFUL ABOUT PRIVATE TUTORING COMPANIES

You may have been contacted by private tutoring companies trying to sell their services to you for statistics courses. Please be extremely careful with these services as some forms of tutoring can pose an academic offence risk. A good tutor helps you understand the subject area and supports your learning. A good tutor does not give you answers. **There are no shortcuts to learning. Learning takes time and effort.**

Be cautious about giving money to companies whose motivation is profit. They may tell you they have ‘insider information’. They don’t. They may even offer you the opportunity to commit academic offenses. Please do not put your University of Toronto education at risk by participating in these kinds of unacceptable behaviours. If you have any questions or concerns about what is okay and what is not in your course, please ask!

2.20 Course design principles

Here are some of the principles around which I have designed this course. I hope they might provide some useful insight into why some things are the way they are, and help you think about how to navigate this course and make the best of it.

The teaching team really wants you to have a great time in this course, learn lots of delicious statistics, and become confident, competent, and useful statistical thinkers. Please approach us and this course with an open mind and help us make it a good experience for you by providing thoughtful, constructive feedback.

2.20.1 Humans learn better ‘little and often’ but everyone is burnout from two years of a pandemic

I know some students absolutely hate weekly tasks and how common they have become with online learning, and that all the little tasks and deadlines can become overwhelming. I also know that cramming is the absolute worst way to learn and actually retain that learning.

This is why in this course:

- I have taken things out or made them baskets options, based on student feedback
- There is a knowledge basket that lets you approach frequent low-stakes practice in the way that suits you best and aims to facilitate and reward spaced repetition, the way science says is the best way to learn ([this free course is fantastic for learning more about how to learn](#)).
- Has been converted to have a two-week module structure to give you more flexibility.
- Has two assessments pathways with different numbers of assessments.

2.20.2 Writing is good for statisticians

Writing not only helps you explain yourself to others, it can also be a fulfilling act of creative personal expression and a way to clarify your own understanding of a concept. Writing is an important part of this course because it is an important skill for your future careers/next steps in education. Lots of support and information here: [<https://writing.utoronto.ca/>](https://writing.utoronto.ca/).

2.20.3 Course content is Accessible

My intention is to make this course accessible as possible with captions for all video and audio and Quercus/course guide design that is easy for folks using screen readers to navigate. If there is something I could do differently in this area that would make your life easier and you’re comfortable to tell me, please do! One thing I know isn’t great but, regretfully, don’t have the resources to change, is the ‘quality assurance’ of the autogenerated captions for videos. Please reach out on Piazza if you’re ever unsure about something they say.

3

Start here!

3.1 Introductions

Hi folks,

Welcome to STA303! We're excited you're joining us on this statistical voyage. I look forward to introducing myself to you in our first class on Wednesday, but for now, there are basic introductions below for me and our Head TA Amin. Feel free to skip to [How this course works](#), I know there is a lot to read in the module!

Looking forward to a great semester! See you in class on Wednesday.

3.1.1 Professor Liza Bolton, Instructor

Email: sta303@utoronto.ca (Put "[Prof. Bolton]" in the subject line to email me directly)

Pronouns: she/her

Before moving (back) to Canada in 2019, I had lived more than half my life in New Zealand. (I still mention New Zealand a lot in class...) My current research areas are in statistics education and online learning, as well as health disparities across ethnic groups. I used to run a small consulting company and called myself a Data Ambassador. Why? Well, lots of people are consultants. I even did an internship in management consulting once upon a time. But it wasn't a satisfying title for what I wanted my work with people to look like. I wanted something that focused on the communication and interpersonal side, not just high quality and appropriate analysis. People who aren't confident in their ability to analyse their own data need a go-between, someone who can be an ambassador for their data! While I don't do consulting any more, I love helping students build their technical and professional skills so they can go out into the world and be excellent ambassadors for data themselves.

Last movie I cried in: Kiki's Delivery Service

Favourite food: Corn. Popped, on the cob, in a chip, Mmmmm.

Book most often given as a gift: [A Matter of Fact: Talking Truth in a Post-Truth World](#) by Jess Berentson-Shaw

3.1.2 Amin Banihashemi, Head TA

Email: sta303@utoronto.ca (Amin will often be the one responding to your emails)

Pronouns: he/him

I'm a fourth-year PhD student at the Institute of Medical Science. I have been a TA for STA130 in DoSS for the past 3 years and this is my second semester as Head TA of STA303.

My area of research is clinical Neuroscience, something I am passionate about. I analyze images of brain and eye structures in neurodegenerative diseases. I investigate possible associations of these structures with each other

and with the ability to remember well and carry out goal-oriented tasks successfully. I love creating reproducible statistical analysis workflows in R. I also like audiobooks, candlelight, and apple pie (which I make myself!)

\includegraphics[width=10d0%]{images/headers/map}

3.2 How this course works

This course is organized into five two-week modules of learning + two one week assessment-focus weeks.

All course material will be made available through this course site in Quercus (any links to outside sites will be found here). Take a moment to familiarize yourself with some of the tools and content areas found in the left navigation bar. You can move through content in a module by selecting the “Next” button at the bottom right of the page. You can also visit the Module section in the left navigation menu so see all the available modules and their contents.

All times listed are ‘Toronto time’, i.e. Eastern Time. Note that Daylight Savings Time begins Sunday, March 13, 2022. You may find this time converter helpful: <https://www.timeanddate.com/worldclock/meeting.html>

3.2.0.1 In most modules there will be:

- A weekly module released on Monday morning.
- A quiz based on the content released in the module due on Tuesday at 6:00 p.m ET.
 - Special note for the [Week 1 quiz](#): this quiz is available until January 26 at 6:00 p.m. ET. (with no penalty) because I know there is a lot to get used to in the first week.
- A synchronous class on Wednesday at 12:00 p.m. ET (L0101) and 3:00 p.m. ET (L0201).
 - Both sessions will be the same, you only need to attend one.
 - Please review the [steps for attending a class](#).
 - Synchronous classes will be recorded. You’re expected to watch the recording if you cannot attend live. They will be posted on the [course overview](#) page.
- Weekly writing Create phase due Thursday at 6:00 p.m. ET.
- Weekly writing Assess phase due Friday at 6:00 p.m. ET.
- Weekly writing Reflect phase due (next) Monday at 6:00 p.m. ET.
- Office hours.
 - Prof office hours will occur after the Tuesday synchronous classes, i.e. 11:10–12:00 p.m ET and 3:00–4:00 p.m., in the same [Zoom call](#).
 - TA office hours: will begin in Week 2 (TBC) and the schedule will be updated [here](#).

3.2.0.2 Students joining off the waitlist

You don’t have to submit missed quizzes or writing activities or alert me that you joined the course late, these will be covered by the associated ‘best of’ policies. See the [Syllabus](#) for more information.

If you have a *friend* on the waitlist, they can sign up to receive materials here.

3.3 Hours expectations

While everyone has different work styles and learning needs, I want to provide some guidance around how I expect this course to look for students.

Plan to be doing 6–8 hours of work on STA303 each *week*. In a two-week module, this may be comprised of:

- 2–4 hours on videos and readings
- 1–3 hours of attending synchronous class or reviewing the recording and activities
- 1–2 hours on knowledge basket assessments
- 2–6 hours on other assessments
- Remaining time attending reading announcements, office hours, checking Piazza, revision, etc.

3.3.1 Module flow

3.3.2 Communication

- Our course discussion board on [Piazza](#) is to be used for all content and administration questions. *Only* sensitive or personal issues/questions should be sent to sta303@utoronto.ca. We reserve the right not to respond to emails that should be Piazza posts.
 - Please ensure all course-related emails include your **UTORID**.
- There are several important **forms** that you may need if you miss an assessment due to **illness or emergency** or wish to request a **regrade** of an assessment.
- I will use Quercus [announcements](#) to share course information and updates. **Please make sure you read these**. I may also occasionally email or Quercus message you about things that relate specifically to you.

3.3.3 To do now

Press the next button below to continue through this module. In the following pages you will:

- Read the [Syllabus](#).
- Join the [Piazza discussion board](#).
- Understand the [tools](#) we will be using in this course.
- [Optional] Introduce yourself in the [Introductions discussion board](#).
- Learn about some of the [services and supports](#) available to you as a U of T student.
- Make sure you have a U of T Zoom account <https://utoronto.zoom.us/>.

Header photo by [Andrew Stutesman](#).

Assessments

4

Assessment overview

There are two special elements of assessment/grade calculation in this course that are important to be aware of in planning your approach to it.

- *Two roads diverged in a yellow wood*¹: You can opt for Path A or Path B to get your final mark. I will calculate your marks along both paths and then assign you the higher of the two as your final mark.
- A-tisket, a-tasket², fill up your **knowledge basket**³:
 - Personalize which of these assessments you do based on your interests and skills you want to develop.
 - You can ‘max out’ your basket: just keep putting grades in until you get to 10% (Path A) or 5% (Path B).

Assessment	Path A	Path B	STA1002 ONLY	Due dates
Mini-portfolio	5	0	0	Feb 3
Portfolio	20	25	25	Feb 7
Mini-mixed assessment	5	0	0	Mar 9 (12-hour assessment window, 8:00 a.m.–8:00 p.m. ET)
Mixed assessment	20	25	25	Mar 16 (12-hour assessment window, 8:00 a.m.–8:00 p.m. ET)
Final project	45	45	50	Apr 7 (5% pt bonus) Apr 11 (no bonus)
Knowledge basket	5	5	0	Multiple
Total	100	100	100	

4.1 Graduate student modification (1002H)

There is no difference in the grading scheme or assessment for graduate students enrolled in STA1002, other than an additional ‘path’ to your final grade where you may opt out of the ‘basket’ assessments, if you wish. This only applies to graduate students enrolled in STA1002, not to any students enrolled in STA303.

You don’t need to advise me of your choice, I will calculate you mark all three ways above and give you the highest of those marks.

¹The Road Not Taken, by Robert Frost. But, dear traveller, you can in fact take both roads and then get whichever gives you the better mark. <https://www.poetryfoundation.org/poems/44272/the-road-not-taken>

²A-tisket, a-tasket by Ella Fitzgerald, <https://www.youtube.com/watch?v=1bgFkeDLpSI>

³“Whaowhia te kete mātauranga” is a Māori proverb meaning “Fill up the basket of knowledge”. Mātauranga is specifically traditional Māori knowledge. You can listen to the pronunciation here <https://www.massey.ac.nz/student-life/m%C4%81ori-at-massey/te-reo-m%C4%81ori-and-tikanga-resources/te-reo-m%C4%81ori-pronunciation-and-translations/whakatauk%C4%AB-m%C4%81ori-proverbs/>

5

Mini-portfolio

Information	Note
Name	Mini-portfolio
Type (Main, Mini or Basket)	Mini
Value	5% (Path A) 0% (Path B)
Due	Thursday, February 3, 2022 at 3:03 p.m. ET
Submission instruction	Submission: Via Markus
Accommodations and extension policy	In the case of a personal illness/emergency, a declaration can be made , but must be submitted no more than 3 days after the due date. Extensions may be requested through the same form up to 48 hours before the due date.

5.0.1 Instructions

6

Portfolio

Information	Note
Name	Portfolio
Type (Main, Mini or Basket)	Main
Value	20% (Path A) 25% (Path B)
Due Submission instruction	Thursday, February 17, 2022 at 3:03 p.m. ET Submission: Via Markus
Accommodations and extension policy	In the case of a personal illness/emergency, a declaration can be made , but must be submitted no more than 3 days after the due date. Extensions may be requested through the same form up to 48 hours before the due date.

6.0.1 Instructions

7

Mini-mixed assessment

Information	Note
Name	Mini-mixed assessment
Type (Main, Mini or Basket)	Mini
Value	5% (Path A) 0% (Path B)
Due	Wednesday, March 9, 2022; assessment window from 8:00 a.m. ET - 8:00 p.m. ET
Submission instruction	Submission: Via Quercus quiz (50 minutes, 1 attempt, no pausing) and Markus (10 percentage point penalty for not submitting required files)
Accommodations and extension policy	In the case of a personal illness/emergency, a declaration can be made , but must be submitted no more than 3 days after the due date. Extensions may be requested through the same form up to 48 hours before the due date.

7.0.1 Instructions

8

Mini-portfolio

Information	Note
Name	Final project
Type (Main, Mini or Basket)	Main
Value	45% (Path A) 45% (Path B)
Due	Thursday, April 7, 2022 at 3:03 p.m. ET for 2% pt bonus. Submission accepted until Monday, April 11, 2022 at 3:03 p.m.
Submission instruction	Submission: Via Markus
Accommodations and extension policy	There are no routine extensions granted for the final project. In exceptional circumstances, you can work with your College Registrar and me on this.

8.0.1 Instructions

9

Mixed assessment

Information	Note
Name	Mixed assessment
Type (Main, Mini or Basket)	Main
Value	20% (Path A) 25% (Path B)
Due	Wednesday, March 16, 2022; assessment window from 8:00 a.m. ET - 8:00 p.m. ET
Submission instruction	Submission: Via Quercus quiz (2 x 50 minute components, 1 attempt, no pausing) and Markus (10 percentage point penalty for not submitting required files)
Accommodations and extension policy	In the case of a personal illness/emergency, a declaration can be made , but must be submitted no more than 3 days after the due date. Extensions may be requested through the same form up to 48 hours before the due date.

9.0.1 Instructions

10

Knowledge basket: Writing and peer feedback

10.1 General instructions

The module writing and peer feedback activities have three stages.

- **Create** phase due the first Friday of the module at 3:03 p.m. ET
- **Assess** phase due the second Tuesday of the module at 3:03 p.m. ET
- **Reflect** phase due the second Friday of the module at 3:03 p.m. ET

10.1.1 Create phase

The **Create** phase is due the first Friday of the module at 3:03 p.m. ET

- Spend ~30 minutes writing a response to the prompt.
- Write about 200–500 words. The word count isn't strict, but the submission requirements listed below ARE.
 - The prompt should be clearly and comprehensively addressed.
 - Your writing should be in full sentences and be broken into paragraphs as appropriate.
 - Any grammatical or word choice errors should be minimal and not obstruct the meaning.
 - There is a clear central idea that is well summarized in a concluding sentence(s).

10.1.1.1 Submission requirements

Your submission should be:

- typed (not handwritten)
- no more than one page
- single-spaced
- size 12 font
- margins should be no larger than 1 inch
- saved as a PDF

10.1.2 Assess phase

The **Assess** phase is due the second Tuesday of the module at 3:03 p.m. ET

You will need to assess TWO of your peers in this phase. You will be asked:

- if they have met the submission requirements,
- to rate them on the rubric,
- to make a short comment about a strength of this piece of writing,
- to make a short comment about a way this piece of writing could be improved.

10.1.3 Reflect phase

The **Reflect** phase is due second Friday of the module at 3:03 p.m. ET

- Read the feedback received from your peers.
- Rate the usefulness on a 3-point scale.

10.1.4 General instructions

These assessments can be used to make up your knowledge basket. They are also useful in helping you prepare for your portfolio writing samples and final project writing.

Warning: MAKE SURE YOU CLICK SUBMIT! Check that each phase is showing as submitted. Students have occasionally struggled with this in peerScholar's interface and no regrades/adjustments will be possible after the fact.

Your mark for this assessment will be based on participation. There will be separate marks for each phase, i.e., you can get part marks overall. 45% for completing create, 45% for completing assess, 10% for completing reflect. Please note that participation not in the spirit of the assessment (e.g. just putting the Lorem Ipsum text, or giving feedback like "write better" and nothing else) will not get you marks.

10.2 Module 1 writing task

Information	Note
Name	Module 1 writing task
Type (Main, Mini or Basket)	Basket
Value	0.5% (0.245%, 0.245%, 0.01%) Completion
Due	Create phase: Friday, January 14, 2022 at 3:03 p.m ET; Assess phase: Tuesday, January 18, 2022 at 3:03 p.m ET; Reflect phase: Friday, January 21, 2022 at 3:03 p.m ET;
Submission instructions	Submission: Via peerScholar
	Marked for completion
Late submissions, accommodations, and extension policy	No late submissions, accommodations, or extensions.

10.2.1 Instructions

Make sure you are familiar with the [general instructions](#) for these types of tasks.

10.2.1.1 Prompt

Discuss what you consider the most important dos and don'ts when giving peer feedback. What will make your peers' feedback most valuable to you?

10.2.1.2 Rubric

	Poor or Missing	Adequate	Good	Excellent
Address prompt	No response OR does not address one of the prompts for this week.	While the prompt is somewhat addressed, there is a lot missing and/or much of the response is not relevant/off-topic.	Prompt is addressed, though may go somewhat off-topic at points, or lacks some depth in its coverage.	Prompt is clearly and comprehensively addressed.
Structure	No response OR there is no structure, very difficult to follow.	Some structure but difficult to follow	The organization follows some logical structure	Well organized, follows a logical structure.
Writing mechanics	No response OR considerable writing and grammatical issues that completely obscure the meaning OR lots of slang and inappropriate word choice.	Multiple sections are difficult to read, but is otherwise understandable.	Slight difficulty in understanding one or two sections.	Can read and follow along with minimal effort. Some grammatical or word choice errors are allowable, but they must not obstruct meaning.
Conclusion	No response OR there is no concluding sentence(s).	The conclusion is weak not well supported.	A conclusion is present but does not completely summarise the central idea.	There is a clear central idea that is well summarised in a concluding sentence(s).

11

Knowledge basket: Professional development task



Image source: <https://www.nsta.org/q-if-tree-falls-forest-and-theres-no-one-around-hear-it-does-it-make-sound>

If a tree falls in a forest, and there's no one around to hear it, does it make a sound? Or, more relevant to this course, if you do a data analysis and can't share it with anyone in helpful ways, did you *really* do anything? With this in mind, you will have the opportunity to choose an area of relevant professional development to pursue over the course of the semester. This could include technical skills that make you better at collaborating with others (version control, Git, GitHub), creating things others will find useful (an R package) or practising communication (oral or written).

There is a 1% **proposal** due fairly early in the semester, and then the final submission of **evidence** of, and **reflection** on, your activity is worth 3% and due toward the end of the semester.

Example professional development tasks include:

- Learning how to set up and use Git and GitHub (this might come in handy if collaborating on the final project with a group)
- Setting up a personal profile website (a bit like a digital CV, GitHub provides free hosting for simple sites)
- Participating in weekly TidyTuesday activities
- Writing a stats blog
- Developing an R package and sharing it on GitHub
- A public speaking based activity like a debating society or Toastmasters
- Conducting a series of interviews with industry professional or academics and publishing videos/write-ups

- Create a wildly successful stats memes TikTok à la [Chelsea Parlett-Pelleriti](#) (okay, maybe not this one...but she's well worth checking out)

Note: the task must be related to **communication or collaboration in some way**. For example, 'learning SQL' would not be sufficiently directly related to communication or collaboration, but developing an R package and sharing it with others requires communication (writing the documentation) and is a great way to contribute to the collaborative and supportive R community. If you are not sure, please ask!

You will use the SMART goals framework (see image below) when setting out your proposal. The more thought you put into this upfront, the easier collecting evidence and reflecting on your progress will be at the end. More information will be available on the respective assessment pages when they go live.

11.1 Professional development proposal

Information	Note
Name	Professional development proposal
Type Main, Mini or Basket	Basket
Value	1%
Due	Thursday, February 3, 2022 at 3:03 p.m. ET
Submission instructions	Submission: PDF via Markus
Late submissions, accommodations, and extension policy	In the case of a personal illness/emergency, a declaration can be made , but must be submitted no more than 3 days after the due date. Extensions may be requested through the same form up to 48 hours before the due date.

11.1.1 Instructions

There is a general overview of this task on the [professional development overview page](#).

Example professional development tasks include:

- Learning how to set up and use Git and GitHub (this might come in handy if collaborating on the final project with a group)
 - Possible resource: <https://happygitwithr.com/>
- Setting up a personal profile website (a bit like a digital CV, GitHub provides free hosting for simple sites)
 - Possible resource: <http://jmcglone.com/guides/github-pages/>
 - Possible resource: <https://uoft-doss-issc.github.io/website-workshop/>
- Participating in weekly TidyTuesday activities
 - Possible resource: <https://github.com/rfordatascience/tidytuesday>
- Writing a stats blog
- Developing an R package
 - Possible resource: <https://r-pkgs.org/index.html>
- A public speaking based activity like a debating society or Toastmasters
- Conducting a series of interviews with industry professional or academics and publishing videos/write-ups

- Create a wildly successful stats memes TikTok à la [Chelsea Parlett-Pelleriti](#) (okay, maybe not this one...but she's well worth checking out)
1. Choose a professional development task that you can devote at least 5–10 hours to over the next several weeks.
 2. Work through the SMART goals framework to describe what you will do. The more thought you put into this upfront, the easier collecting evidence and reflecting on your progress will be at the end. More information will be available on the respective assessment pages when they go live.
 3. Explain WHY the goal is a good choice for you and your career/further education path ('Relevant' criteria).

11.1.2 Submission requirements

Your proposal should be:

- typed (not handwritten)
- one page¹
- single-spaced
- size 12 font
- margins should be no larger than 1 inch
- saved as a PDF

With these specifications, your proposal will be approximately 500 words.

11.1.3 Rubric

¹References, if relevant, may be included on a second page, but a simple hyperlink may also be sufficient. References are not required.

Component	Missing 0%	Poor 25%	Adequate 50%	Good 75%	Excellent 100%	Points
Specific	There is no goal stated.	Goal is too generic and poorly specified.	Goal is stated but it is not specific. At least one resource may or may not be listed.	Goal is specific but no/unclear resource is mentioned.	Goal is specific. At least one resource that will be used to help is mentioned.	1
Measurable	There is no measure for attaining the goal.	Minimal evidence of consideration for measuring progress towards the goal or success.	Definition of success lacks clarity. Measures of progress may or may not be listed.	What success will look like is clearly defined but there no/unclear statement of measure of progress.	Both definition of success and a measure of progress are clearly defined.	1
Attainable	There is no explanation for how the goal is to be attained.	Inappropriate scope, not believable that the goal is attainable.	Some issues with the scope of the goal, may not be attainable. Understanding of the steps and potential problems very limited.	Goal has a mostly appropriate scope but may lack some clarity around steps to take and potential problems.	Goal has an appropriate scope, and a strong understanding of the steps to take and potential problems is shown.	1
Relevant	Goal is not related to communication or collaboration appropriate for statisticians/data-related roles AND no discussion of why this goal was chosen.	Goal is not related to communication or collaboration appropriate for statisticians/data-related roles but some discussion of why this goal was chosen OR goal is appropriate, but no discussion of personal relevance.	Goal is tenuously related to communication or collaboration appropriate for statisticians/data-related roles but good discussion of why this goal was chosen OR goal is appropriate, but limited discussion of personal relevance.	Goal is related to communication or collaboration appropriate for statisticians/data-related roles AND there is a reasonable discussion of personal relevance.	Goal is related to communication or collaboration appropriate for statisticians/data-related roles AND there is a clear discussion of personal relevance.	2
Time-bound	There is no timeline or evidence of consideration of time-bounding the goal.	Minimal evidence of consideration time-bounding the goal.	Timeline is not listed according to days/weeks or dates. Time for delays and troubleshooting may or may not have been considered.	Timeline is listed based on days/weeks or dates. No/unclear time for delays and troubleshooting is considered.	Timeline is listed based on days/weeks or dates. Time for delays and troubleshooting is considered.	1
Structure	There is no structure, very difficult to follow.	Minimal evidence of an attempt to structure the proposal logically.	Some structure but difficult to follow.	The organization follows some logical structure.	Well organized, follows a logical structure.	1
Writing mechanics	No response OR response is largely unintelligible.	Considerable writing and grammatical issues that completely obscure the meaning OR lots of slang and inappropriate word choice.	Multiple sections are difficult to read but it is otherwise understandable.	Slight difficulty in understanding one or two sections.	Can read and follow along with minimal effort. Some grammatical or word choice errors are allowable, but they must not obstruct meaning.	2
Conclusion	There is no concluding sentences.	Minimal evidence of a concluding statement.	The conclusion is weak not well supported.	A conclusion is present but does not completely summarise the central idea.	There is a clear central idea that is well summarised in a concluding sentences.	1

11.1.4 Checklist

Before submitting your proposal, check the following:

- Your goal is specific. (**Specific**)
- At least one resource you will use to help you is identified. (**Specific**)
- It is clear how you will define success. (**Measurable**)
- It is clear how you will measure your progress. (**Measurable**)
- Your goal has an appropriate scope and is attainable. (**Attainable**)
 - This is shown through a description of the steps you’ll need to take and what potential problems you might face. Related to timeline below.
- The goal is related to communication and/or collaboration appropriate for a statistician/data-related role. (**Relevant**)
- It is clear why this goal is relevant to you personally. (**Relevant**)
- The steps you want to complete can be completed in 7 weeks. (**Attainable** and **Time-bounded**)
- A timeline with dates/weeks for the required steps is included. (**Time-bounded**)
- The timeline shows some accommodation for troubleshooting/delays. (**Time-bounded**)
- There is a clear concluding sentence or sentences that wrap(s) up the proposal.
- You have proofread your proposal and made sure the structure is logical and there are not intrusive grammatical or word choice errors.
- You have written in full sentences.
- Your submission is typed (not handwritten), one page, single-spaced with size 12 font and the margins are no larger than 1 inch.
- Your final version is saved as a PDF.

11.1.5 Things to keep in minds as you start working towards your professional development goal

- Track your time
 - A time sheet should be part of your evidence submission
- Create a work log document or file to store screenshots in, make notes of tasks you’ve completed
 - This will make writing your reflection much easier

11.1.6 Example smart goal

Note: This is *not* a collaboration/communication goal, but provides some examples of how to approach a SMART goal. Where there are ellipses (“...”) we are suggesting there would be more of the same, but this example should be enough to give you the idea. You should *not* use ellipses in this way in your own proposal.

11.1.6.1 Specific

Goal: *“Climb to the peak of Mount Robson over 2 weeks in January and write one blog post for my website about the journey.”*

Bad example: “Climb a mountain and write about it.” Which mountain? What type of writing and with what purpose?

Resources: *“I will need climbing equipment, my camera, notebook ...”*

11.1.6.2 Measurable

Defining success: *“Reach the peak of Mount Robson and return and write at least 1 blog entry with a picture about the journey, all in 2 weeks.”*

Bad example: “Climbed Mount Robson.” Needs more details about what this success looks like. Walk up a few metres and turn back? Relates to time-bounded also.

Measuring my progress: *“I will measure my progress by how many steps of my goals I have accomplished and the altitude of my climb relative to the elevation of Mount Robson (3,954 m). I will have taken a picture”*

11.1.6.3 Attainable

You have or can you learn the required skills: *“I am sufficiently fit to make this climb safely and have climbed a similar mountain recently. I have a camera and am currently taking an online course on blogging.”*

Possible steps I need to take:

1. *Arrive at Mount Robson Visitors Center with my climbing plan.*
2. *Climb until ... on first day.*

...

6. *Write one blog entry with one picture.*

Bad example: “Make progress with my goal by 7% each day.” Not useful to you, what does 7% actually mean here? Will it really be the same amount of progress every day?

Potential problems:

1. *Weather might be bad, and I may have to delay my climb by a few days.*

Bad example: “Something might not work.” What? In what way? Could you solve it?

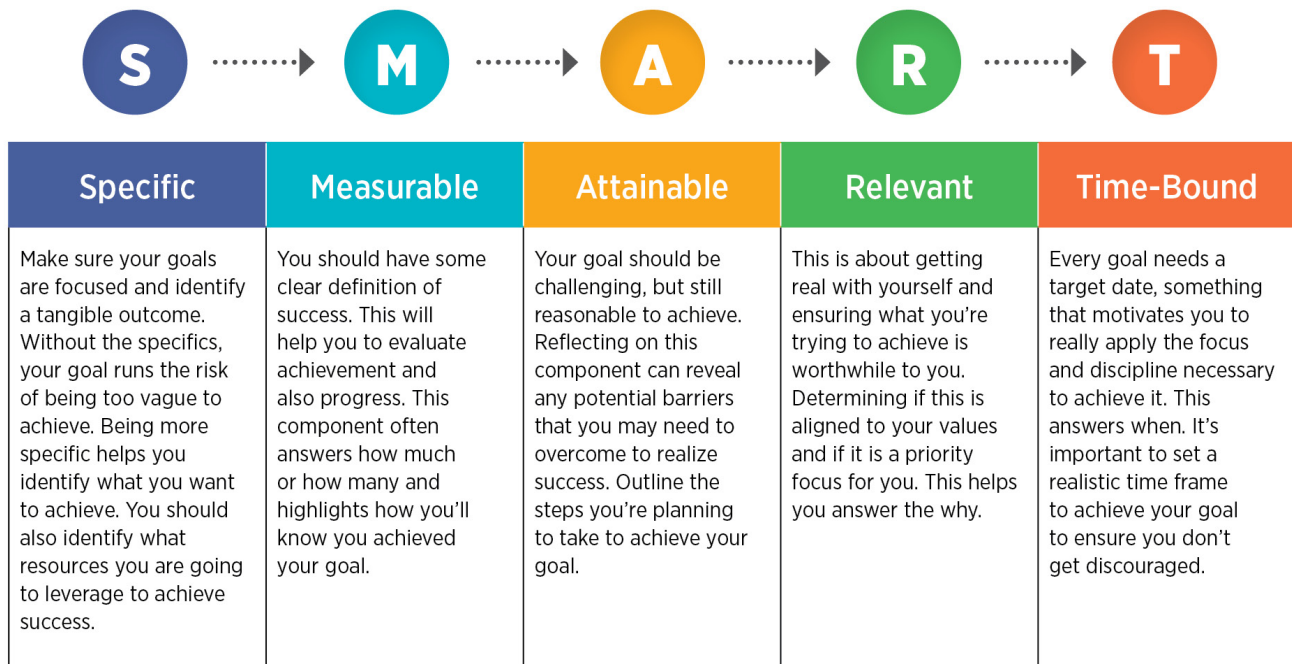
11.1.6.4 Relevant

Personal relevance: *“The climbing experience will help me learn new skills like ... which will help me work towards becoming a professional climbing guide.”*

Bad example: *“Seems like fun.”*

11.1.6.5 Time-bounded:

“A 2-week journey from Jan 25th:Days 1-3to climb;Days4-6return;5 days for delays;3days to rest.”



Source: Canadian Management Center, <https://cmcoutperform.com/setting-smart-goals>

11.2 Professional development evidence and reflection

Information	Note
Name	Professional development evidence & reflection
Type Main, Mini or Basket	Basket
Value	3%
Due	Thursday, March 31, 2022 at 3:03 p.m. ET
Submission instructions	Submission: PDF via Markus
Late submissions, accommodations, and extension policy	In the case of a personal illness/emergency, a declaration can be made , but must be submitted no more than 3 days after the due date. Extensions may be requested through the same form up to 48 hours before the due date.

You may wish to use the provided template to complete this task. This is *not* required to get full marks, but we believe this will be a helpful structure for you to ensure you have addressed all parts of the task.

11.2.1 Templates

- [template.docx](#)
- [template.Rmd](#), [template.pdf](#), [template_rmd.zip](#)

11.2.2 Reflection and evidence components

11.2.2.1 Activity, alignment and lessons

Activity: Describe what you did and what you learned. Be **specific** but you don't need a step-by-step guide or dates, save that for your timesheet. Assume your reader has *not* read your proposal and that you are introducing your goal and plan to them for the first time. (It may help to imagine how you might explain this task and your progress in a job interview.)

11.2.2.1.1 Examples

- *I wrote 4 blog posts on The first one was on ... One unanticipated activity I had to do was ...*
- *This task helped me learn how to write for general audience...*
- *Through this task I improved my...*

Alignment: How well did your activity and progress align with your proposal?

11.2.2.1.2 Examples

- *My plan changed as my goal was not specific enough. I think the reason for this is ... I can set realistic goals next time by ...*
- *Most of my attainable steps and potential problems were correctly identified because*
- *This is a good method which I will use again when I need to ...*
- *My way of measuring progress worked because ... This is useful for next time because ...*

Lessons: What did you learn from using the SMART goal-setting experience?

11.2.2.1.3 Examples

- *An interesting thing I learned from the SMART goal-setting exercise was that ...*
- *I prefer to modify the format of SMART goals to match my work habits by ...*

11.2.2.2 Evidence

Provide links **and/or** copy/paste a screenshot of a blog, website, Github repository, Rmd document, ... Provide commentary on what the image/link is and what it demonstrates. I.e., Describe what we should be seeing and understanding from what you've included and how it relates to your goals.

11.2.2.2.1 Examples

- *The screenshot shows ... which is evidence of partially completing [goal] ...*
- *The provided link goes to ... which relates to [goal] ...*
- *Notice that there are separate buttons on the website for each of the ...*

11.2.2.3 Timesheet

Include a timesheet that briefly describes how you used your time (no specific format required, but something like the below is fine). Hours spent can be approximate.

Describe what you learned from filling your timesheet. This can be short and could include how you'd track your time differently in future, what you noticed about your work habits from tracking your time

11.2.2.3.1 Examples *About ...% of my time was spent preparing, maybe because the task was new. This means for future tasks that are new I need to consider*

The timesheet can be better suited to my work habits if ...

Week	Week starts on (Monday)	Time spent (Hrs)	Activity (Brief description of what was done, e.g. what you read, what you tried on GitHub or state.)
5	Feb 7	1.5	Example: Worked through chapters 1 & 2 of ...
6	Feb 14 Feb 21	0	Example: No progress this week <i>Reading week</i>
7	Feb 28	1	Example: Set up postcard home page and pushed to GitHub ...
8	Mar 7	2	Example: Wrote one blog entry about ...
9	Mar 14	1.5	Example: Fixed coding error that ...[reason for error] .. by [fix] ...
10	Mar 21	2	Example: Edited the video on [topic] ... using [software] ...

11.2.3 Recommended structure

You do not *have* to set out your writing in this way to get full marks, but we believe this will be a helpful structure for you to ensure you have addressed all parts of the task.

#	Section heading	Format
1	Activity, alignment and lessons	~3 paragraphs: <ul style="list-style-type: none"> • activity, • alignment, • SMART lessons
2	Evidence	Selected screenshots of work done, links, ...
3	Timesheet	Fill provided template or create your own <ul style="list-style-type: none"> • 1-3 sentences of what you learned

11.2.4 Rubric

Component	Missing 0%	Poor 25%	Adequate 50%	Good 75%	Excellent 100%	Points
Activity description	Missing	Limited description of task, insufficient detail to understand the scope of activity.	Some description of the activity, but too general and/or would require the reader to have some previous knowledge of the goal/activity chosen to fully under it.	Mostly clear description of activity but lacking some specificity or does not fully introduce the task for a reader with no previous knowledge of the tasks.	Clear and specific description of the activity undertaken. Appropriately introduced for a reader with no background knowledge of the task.	2
Alignment	Missing	Limited description of alignment with original proposal. Lacks specifics and clear reasoning.	Some description of alignment with original proposal, though lacking reflective depth.	Reflection on alignment with proposal provides some specific details and reasoning for changes/successes.	Reflection on alignment with proposal is insightful and specific, detailing what went to plan/what didn't and any changes that needed to be made, as appropriate.	2
SMART goal-setting lessons	Missing	Limited description of lessons learned from the SMART goal setting approach.	Some description of lessons learned from the SMART goal setting approach, though lacking reflective depth.	Reflection on lessons learned from the SMART goal setting approach provides some specific details	Reflection on lessons learned from the SMART goal setting approach is insightful and specific.	2
Evidence	Missing	Insufficient evidence provided. No explanation or relation to goal was not clear.	Some evidence provided but aspects of explanation/commentary were not clear and/or the relationship to the goal not well described.	Evidence provided with commentary and the relationship to the goal described.	Relevant evidence provided with clear commentary on what it demonstrated and how it related to the goal described.	1
Timesheet	Missing	Some hours or activity descriptions missing/insufficient. No description of what was learned from the time tracking activity.	Some hours or activity descriptions missing/insufficient. Some description of what was learned from the time tracking activity.	Timesheet complete but insufficient description of what was learned from the time tracking activity OR timesheet somewhat complete but good description of what was learned from the time tracking activity.	Timesheet complete and a clear and relevant description of what they learned from the time tracking activity.	1
Structure	Missing	Minimal evidence of an attempt to structure logically.	Some structure but difficult to follow.	The organization follows some logical structure. Paragraphs either don't have topic sentences or have unrelated ideas.	Well organized, follows a logical structure. Paragraphs have topic sentences and contain related ideas.	1
Writing mechanics	Missing	Considerable writing and grammatical issues that completely obscure the meaning OR lots of slang and inappropriate word choice.	Multiple sections are difficult to read but it is otherwise understandable.	Slight difficulty in understanding one or two sections.	Can read and follow along with minimal effort. Some grammatical or word choice errors are allowable, but they must not obstruct meaning.	1

12

Knowledge basket: Other

12.1 ‘Getting to know you’ survey

Information	Note
Name	‘Getting to know you’ survey
Type (Main, Mini or Basket)	Basket
Value	0.1%
Due	Thursday, January 13, 2022 at 3:03 p.m ET
Submission instructions	Submission: Via Quercus survey
	Untimed survey, marked for completion
Late submissions, accommodations, and extension policy	No late submissions, accommodations, or extensions.

12.1.1 Instructions

Answer the following questions to help me get to know the class. Some of the questions might seem a little random, but I’m hoping to use them for some future class activities.

(Some questions are from factfulnessquiz.com, you are not being marked on correctness, so please answer to the best of your ability without looking anything up.)

12.2 Pre-knowledge check

Information	Note
Name	Pre-knowledge check
Type (Main, Mini or Basket)	Basket
Value	0.5% for completion + 0.5% for a score of 80%+ OR active attendance at the prerequisite knowledge workshop 2022-02-02
Due	Thursday, January 20, 2022 at 3:03 p.m ET (Workshop: Wednesday, February 2, 2022 in class time)
Submission instructions	Submission: Via Quercus survey 60 minutes, 1 attempt, no pausing
Late submissions, accommodations, and extension policy	No late submissions, accommodations, or extensions.

12.2.1 Instructions

Answer the pre-knowledge check quiz questions to the best of your ability. In Quercus, you will see two assignment entries, one for the quiz and for storing your final score that takes into account completion of the quiz and the 80%+ score or workshop engagement/attendance.

12.2.1.1 Grade structure

- Completion of the quiz (all questions attempted) will earn you 0.5%.
- If you score 80% or more on the quiz, you earn an additional 0.5%.
- If you *do not* score 80% or more, you can still earn this additional 0.5% by attending and actively engaging with the workshop on Wednesday, February 2, 2022 in class time.
- The maximum points to earn here is 1% (i.e., you cannot get BOTH the 80%+ and workshop attendance bonuses)

Modules

13

Module 1

13.1 Instructor information

Prof. Liza Bolton

Course email: sta303@utoronto.ca

Help hours: During the second half of Wednesday classes: 11:10—12:00 p.m. ET and 4:10—5:00 p.m. ET

Please do not email my individual email with STA303 questions or use Quercus mail. It makes it harder to keep track of and address your questions. Sending me messages on three different platforms will NOT speed up my response. I will not respond messages that don't follow the course communication policy.

You can see the latest version of [my email autoresponder here](#).

13.2 Upward management tips

'Upward management' is basically managing your manager.¹ If you make their life easier and help them be more effective, this should also make *your* life easier and is a good investment in your own career and skills building.

While our course isn't a business, some of the basic parts of this concept apply well to your time at univeristy AND can set you up for success in graduate studies and your future career.

But why should you put effort into upward managing me? Well, while I will always seek to treat all students fairly and to listen to your feedback, YOU can make this easier or harder for me, and thus make this course better or worse for yourselves.

For example, if I have to use all my time and energy following up on unclear emails for more information and dealing with students who haven't followed instructions etc. etc....I won't have that energy to put into writing TeamUp! activities to give you extra practice and opportunities to earn bonus points.

13.2.1 Communicate using the tools your manager prefers.

- **In business**, this means knowing who likes face-to-face vs email, or whether your manager would rather receive an instant message than an email for a quick question.
- **In STA303**, this means:
 - Using Piazza for all course admin and content questions.
 - Using the appropriate forms for accommodations and regrade requests.

¹I used to haaaaate this concept because I learned it while doing a consulting internship at a large international professional services firm and I really struggled with my manager.

- Emailing sta303@utoronto.ca for private issues not otherwise covered by the other tools, e.g. emailing me your Accommodation Services letter, requesting an extension to an assessment that conflicts with essential travel.
- Asking questions in office hours.

13.2.2 Write good emails (when emails are appropriate)

- **In business**, this might mean:
 - Choosing who should be the main recipients vs CCed/BCCed.
 - Ensuring your contact details are clear in your signature.
 - Make sure the subject line is informative and short
 - Making the text of the email as clear and concise as possible.
 - Use proper grammar and punctuation.
 - Don't use emoji in formal emails. If in doubt, leave 'em out.
- **In STA303**, this looks like:
 - Everything in the right-hand column, plus the following.
 - Starting an email with “Hi Prof. Bolton,” or “Hi Liza,”.²
 - Sign off the email with your **preferred name** (i.e. what should I call you when I reply) and if you have different official name, include that and your UTORid below your name.
 - Subject line including [Prof. Bolton] or [TA name] if your email is for a specific person.

13.2.3 Understand your manager's goals.

- **In business**, this might look like understanding their KPIs and how you can help make sure these are met.
- **In STA303**, most of *my* goals are *for you*, like that you learn useful statistical skills, improve your writing skills etc. I also personally want to improve as an instructor and have fun talking about something I love.
 - You can help me with these goals by working on this course every week, trying your best, asking for help early and often, engaging with feedback gathering mechanisms and providing constructive feedback if something is not working.

13.2.4 Demonstrate self-management and resilience while also asking for help and flagging problems early.

- **In business**, this might look like:
 - Being proactive about addressing possible problems before they occur. Managers like a ‘no surprises’ policy.
 - Preparing a list of questions to cover in a meeting or to compile in an organised fashion into one email (instead of ten).
 - Searching for answers yourself before asking your manager and improving your strategies for finding available information.
- **In STA303**, this looks like:
 - Putting in a little effort into find answers before posting on Piazza/asking in class. (*Have you searched Piazza? have you re-read/watched the assigned materials? have you checked the syllabus and recent announcements?*)
 - Come to office hours often, lists of questions very welcome!
 - Getting in touch (or asking your registrar to) **early** if you might hate to miss a lot of class. It is usually easier to find a solution if I know things ahead of time, or as soon as possible after.

²“Dear Madam” and “To my esteemed professor” make me uncomfortable.

13.3 Recap of linear models

13.4 Why model?

- The goal of a model is to provide a (relatively) simple summary of a dataset.
- We can describe data AND make predictions.

13.5 Linear models

In a linear model,

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \epsilon_i$$

The response is predicted by a linear function of explanatory (or predictor) variables plus an error term.
a.k.a.

$$\text{DATA} = \text{MODEL} + \text{ERROR}$$

13.6 Linear regression assumptions

L: your model is **L**inear.

I: Errors are **I**ndependent (usually satisfied if observations are independent).

N: Errors are **N**ormally distributed with expected value zero, $E[\epsilon_i] = 0$

E: **E**qual/constant variance (homoscedasticity), $\text{var}[\epsilon_i] = \sigma^2$.

We can express “I N E” above as assuming the errors are i.i.d Normal with mean of zero and variance σ^2 ,

$$\epsilon_i \sim N(0, \sigma^2)$$

13.7 What makes it a *linear* model?

A model is **linear** if it is *linear in the parameters*. That is, all the β enter the model in a linear way. It is totally fine if the predictor variables enter the model in a non-linear way.

13.7.1 Linear

- $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$
- $y_i = \beta_0 + \gamma_1 \delta_1 x_{1i} + \beta_2 \exp(x_{2i}) + \epsilon_i$

13.7.2 NOT linear

- $y_i = \beta_0 + \beta_1 x_i^{\beta_2} + \epsilon_i$
- $y_i = \beta_0 \exp(\beta_1 x_{1i}) + \epsilon_i$

Internally screaming “DON’T LET THE BETAS TOUCH” often helps me remember what is not linear. Additionally, don’t let them do anything *weird*, like get exponentiated.

13.8 Optional refresher reading

13.8.1 Brief discussion of assumptions with examples

Section 1.3 of *Broaden your Statistical Horizons* on the assumptions of OLS <https://bookdown.org/roback/bookdown-bysh/ch-MLRreview.html#ordinary-least-squares-ols-assumptions> [freely accessible]

13.8.2 Fitting linear models in R

Section 1.6 of *Broaden your Statistical Horizons* on Multiple linear regression (bootstrapping not assessed) <https://bookdown.org/roback/bookdown-bysh/ch-MLRreview.html#multreg> [freely accessible]

13.8.3 Delicious mathematics

Chapter 1 of Wood, S. N. (2017). *Generalized additive models : An introduction with r, second edition* <http://go.utlib.ca/cat/13435628> [you will need to log in to the U of T library for access]

13.9 Common statistical tests as linear regression

13.9.1 Introduction

Technical note: I have had to remove code answer checking, but if you get stuck the final ‘Hint’ is the solution. This is ungraded, so don’t worry, just slightly inconvenient.

You have probably encountered several statistical tests in your studies so far.

13.9.1.1 Parametric

E.g. one-sample t-tests, paired t-tests, two-sample t-tests, one-way ANOVA, two-way ANOVA

Parametric tests make assumptions about the distribution of the population from which our sample data have been drawn.

13.9.1.2 Non-parametric

E.g. Wilcoxon signed rank, Mann Whitney-U, Kruskal-Wallace

Non-parametric tests do not assume that our outcome is Normally distributed. They are sometimes called ‘distribution-free’, but note that this is because they have fewer assumptions than parametric tests, not because they have no assumptions at all.

13.9.1.3 Aside: But why are there two types of tests?

Parametric tests are more **powerful**, i.e., they have a better chance of detecting an effect if there is one there to find. So why would you ever use a less powerful test? Well, with great power comes ~~great responsibility~~ more assumptions that must be valid to proceed.

There is a GIF in the web version. Not required content. <https://tenor.com/view/spider-man-uncle-ben-with-great-power-comes-great-responsibility-its-true-just-saying-gif-24193883>

Non-parametric tests are a great choice when your outcome is an ordinal variable, is ranks, or there are problematic outliers.

For the purposes of this lesson, we're going to focus more on parametric tests, but also take a look at the corresponding non-parametric tests with the slight white lie that they are just ranked versions of their parametric companions. This approach is pretty good as long as you have a reasonable sample size.

Imagine this:

*You're on a ship trying to spot land. **Parametric** tests are the crew member with the best eyesight, but they can be fussy and the conditions have to be right for them to work in or they will breakdown.*

***Non-parametric** tests are the crew member with not quite as good eyesight, but they're more laid back about the conditions you make them work in.*

In the following sections we'll explore several of these tests.

13.9.2 One-sample t-test

I am assuming you've seen this in a 200-level statistics course or equivalent. Brief recap below.

13.9.2.1 Use case

You want to know if it is believable that the population mean is a certain value (our 'hypothesized value' below).

13.9.2.2 Assumptions

1. The data are continuous.
2. The data are normally distributed.
3. The sample is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample

(Do these sound familiar from linear regression?)

13.9.2.3 Hypotheses

$$H_0 : \mu = \text{hypothesized val}$$

$$H_1 : \mu \neq \text{hypothesized val}$$

What are we doing? Finding the strength of evidence against the claim that the population mean is some hypothesized value.

The test statistic, t , is calculated as follows:

$$t = \frac{\bar{x} - \text{hypothesized val}}{s/\sqrt{n}}$$

We then compare this t value to the t -distribution with degrees of freedom $df = n - 1$ and find the area under the curve that represents the probability of values like ours or more extreme.

13.9.2.4 Example

Suppose existing research suggests that the average weight of penguins is 4000 grams. You want to see if this makes sense for your new penguins data.

$$H_0 : \mu = 4000$$

$$H_1 : \mu \neq 4000$$

The `penguins` dataset is already loaded, you don't have to run any libraries. Use the `t.test()` function run a one-sample t -test.

```
##
## One Sample t-test
##
## data:  penguins$body_mass_g
## t = 4.6525, df = 341, p-value = 4.7e-06
## alternative hypothesis: true mean is not equal to 4000
## 95 percent confidence interval:
##  4116.458 4287.050
## sample estimates:
## mean of x
##  4201.754
```

13.9.2.5 Now as a linear model

First, consider the following, what would a linear regression with no predictor variables and just an intercept tell you?

Create a linear regression model called `mod1` (replace the blank below) that is an ‘intercept only model’ with `body_mass_g` as the response.

```
##
## Call:
## lm(formula = body_mass_g ~ 1, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1501.8  -651.8  -151.8   548.2  2098.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4201.75      43.36   96.89  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 802 on 341 degrees of freedom
## (2 observations deleted due to missingness)
```

It turns out the estimate from this linear regression is the same as the sample mean.

```
mean(penguins$body_mass_g, na.rm = TRUE) #na.rm = TRUE removes missing values
```

```
## [1] 4201.754
```

Now, recall that with the t-test, we calculate our test statistic by subtracting the hypothesized value from the mean. Let’s run the linear model again, but on the left-hand side of the formula, subtract the hypothesized value.

```
##
## Call:
## lm(formula = body_mass_g - 4000 ~ 1, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1501.8  -651.8  -151.8   548.2  2098.2
##
## Coefficients:
```



```
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)   201.75     43.36   4.652  4.7e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 802 on 341 degrees of freedom
## (2 observations deleted due to missingness)
```

Compare the results of this `summary(mod2)` and your earlier t-test. You should see that the t value, degrees of freedom and p-value are the same for both analyses.

Thus, our one sample t-test hypotheses,

$$H_0 : \mu = \text{hypothesized val}$$

$$H_1 : \mu \neq \text{hypothesized val}$$

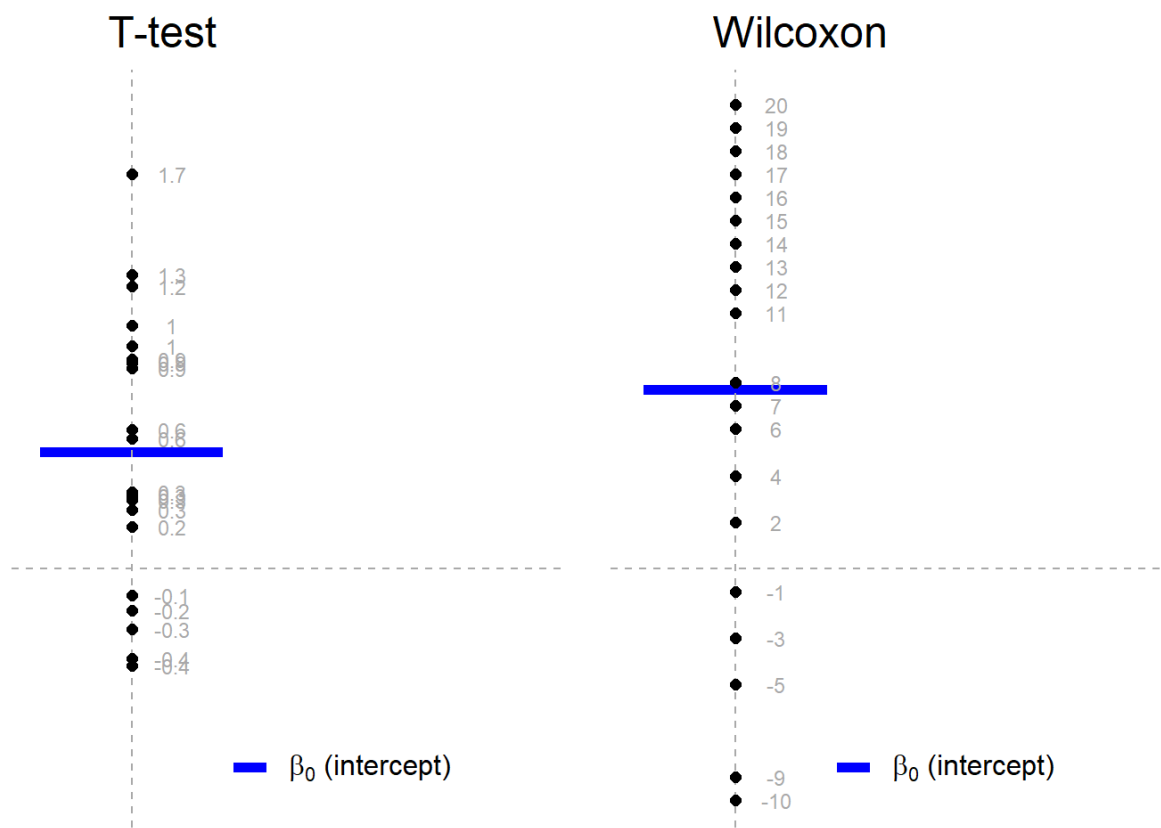
are equivalent to our linear regression hypotheses about the intercept,

$$H_0 : \beta_0 = \text{hypothesized val}$$

$$H_1 : \beta_0 \neq \text{hypothesized val}.$$

13.9.2.6 Wilcoxon signed-rank test

While the linear regression approach to the one-sample t-test is exact, we can also approximate the Wilcoxon rank-sign test with linear regression. See below.



Note: The above is just example from some toy data, but aims to illustrate how a t-test is treating the data and how the Wilcoxon test is treating the data.

```
# Function to get signed rank of each observation
signed_rank = function(x) sign(x) * rank(abs(x))

# The wilcoxon test function
wilcox.test(penguins$body_mass_g, mu = 4000)

##
## Wilcoxon signed rank test with continuity correction
##
## data: penguins$body_mass_g
## V = 34723, p-value = 0.0004829
## alternative hypothesis: true location is not equal to 4000
```

```
# Equivalent linear model
mod3 <- lm(signed_rank(penguins$body_mass_g-4000) ~ 1)
summary(mod3)

##
## Call:
## lm(formula = signed_rank(penguins$body_mass_g - 4000) ~ 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -334.63 -173.13  -22.13  187.87  305.37
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    36.63      10.53   3.479 0.000569 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 194.7 on 341 degrees of freedom
## (2 observations deleted due to missingness)
```

[Optional] Check out the theory behind the rank transformation in section 3.0.2 https://lindeloev.github.io/tests-as-linear/#3_pearson_and_spearman_correlation

13.9.2.7 Paired sample t-test and Wilcoxon matched pair

A paired t-test is equivalent to a one sample t-test if you just consider $x_{\text{diff } i} = x_{1i} - x_{2i}$, i.e., $x_{\text{diff } i}$ is the difference of the paired values for each observation, and proceed with $x_{\text{diff } i}$ as you would in the one sample case. Likewise for the Wilcoxon

The R code (not evaluated here) would be as follows:

```
# Built-in Wilcoxon matched pairs
wilcox.test(x1, x2, paired = TRUE)

# Equivalent linear model:
summary(lm(signed_rank(x1 - x2) ~ 1))
```

13.9.3 Dummy variables

Let's take a quick detour before we explore the next tests. We'll need to understand the concept of dummy variables and contrasts first.

13.9.3.1 The matrices we use for linear regression

Recall that we can express our linear regression in matrix form:

$$\mathbf{y} = X\beta + \varepsilon$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

and

$$X = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

We often talk about \mathbf{X} as the **model matrix** (or design or regressor matrix) and it will be the focus of this section.

13.9.3.2 Getting our model matrix in R

Let's start by fitting a model with `body_mass_g` as the response and `flipper_length_mm` and `species` as the predictor variables.

(Note: Users of statistics use a lot of different words to refer to the same thing. Can you think of other terms people might use instead of *response* and *predictor*?)

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + species, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -927.70 -254.82  -23.92   241.16 1191.68
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4031.477    584.151  -6.901 2.55e-11 ***
## flipper_length_mm    40.705      3.071   13.255 < 2e-16 ***
## speciesChinstrap  -206.510     57.731  -3.577 0.000398 ***
## speciesGentoo     266.810     95.264   2.801 0.005392 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 375.5 on 338 degrees of freedom
## (2 observations deleted due to missingness)
## Multiple R-squared:  0.7826, Adjusted R-squared:  0.7807
## F-statistic: 405.7 on 3 and 338 DF,  p-value: < 2.2e-16
```

Now we can use the `model.matrix()` function to extract the model matrix for `mod4`. I've applied `head()` to stop the entire thing being printed.

```
##      (Intercept) flipper_length_mm speciesChinstrap speciesGentoo
## 1             1             181             0             0
## 2             1             186             0             0
## 3             1             195             0             0
## 5             1             193             0             0
## 6             1             190             0             0
## 7             1             181             0             0
## 8             1             195             0             0
## 9             1             193             0             0
## 10            1             190             0             0
## 11            1             186             0             0
## 12            1             180             0             0
## 13            1             182             0             0
## 14            1             191             0             0
## 15            1             198             0             0
## 16            1             185             0             0
## 17            1             195             0             0
## 18            1             197             0             0
## 19            1             184             0             0
## 20            1             194             0             0
## 21            1             174             0             0
```

You'll notice that even though we only had an intercept and two variables, we have four columns in our model matrix. You should also notice that R has given the columns helpful names, and that we have a column for the Chinstrap species and the Gentoo species, but not the Adelie species.

Further, recall that when we are working with a categorical variables we call the different values the the variables can take “**levels**”. I may also refer to these as factor variables, and talk about the “levels of the factor”.

What R is doing is dropping the first level (alphabetically) of the categorical variable and then creating **dummy variables** for each of the other levels.

The dropped level becomes our **reference level** and this should be familiar from interpreting summary output in previous courses where you have conducted multiple linear regressions with categorical variables.

A dummy variable is also called an indicator variable, and it *indicates* whether or not the given observation takes that level or not. I.e., if the 40th penguin in this dataset had a 1 in the speciesGentoo column, then I know it is a Gentoo penguin, and that it won't have a 1 in the speciesChinstrap column because each penguin can only have one species.

More generally, the sum across the row of the dummy variables for one categorical variable will either be 0 (if that observation has the reference level) or 1 (not the reference level) but you will never have more than one ‘one’ amongst the dummies for a given categorical variable.

13.9.3.2.0.1 [Unassessed aside] Why do we have to drop one of the levels? You may recall that for the matrix calculations required to get our vector of β s, we need to be able to invert X our matrix. We can only invert

matrices for which all the columns are linearly independent and if we have the intercept AND dummies for all the levels of the categorical variable, our matrix will be linearly dependent.

Additional optional discussion [here](#).

13.9.4 Two means

Back to tests!

Independent t-tests let you compare two means. I am assuming you've seen this in a 200-level statistics course or equivalent. Brief recap below.

13.9.4.1 Use case

You want to know if it is believable that two independent groups have the same population mean.

13.9.4.2 Assumptions

1. The data are continuous.
2. The data are normally distributed (in each group).
3. Each group is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample
4. The variances for the groups are equal.

Notice that these are the same assumptions as the one-sample t-test, but with the equality of variances assumption added.

13.9.4.3 Hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

What are we doing? Finding the strength of evidence against the claim that the population means for both groups are the same. This differs from the one sample test because we have uncertainty about BOTH values here. Both are population parameters that we don't know.

The test statistic, t, is calculated as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

We then compare this t value to the t-distribution with degrees of freedom $df = n_1 + n_2 - 2$ and find the area under the curve that represents the probability of values like ours or more extreme.

13.9.4.4 Example

Conduct an independent t-test to test if the mean of `body_mass_g` is the same for male and female penguins (`sex`). Add your code below. Note: you must set `var.equal = TRUE` as one of the arguments for it to be the independent t-test. If you don't set this we are conducting a *Welch's t-test*. I won't be covering this, but it is covered in the source credited at the end of this activity.

```
##
## Two Sample t-test
##
## data:  body_mass_g by sex
## t = -8.5417, df = 331, p-value = 4.897e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -840.8014 -526.0222
## sample estimates:
## mean in group female    mean in group male
##           3862.273           4545.685
```

Now, based on what we've learned, write a linear model using the `lm()` function to do the same this as our independent t-test. Save the model as `mod5`.

```
##
## Call:
## lm(formula = body_mass_g ~ sex, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1295.7  -595.7  -237.3   737.7  1754.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3862.27      56.83   67.963 < 2e-16 ***
## sexmale      683.41      80.01    8.542 4.9e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 730 on 331 degrees of freedom
## (11 observations deleted due to missingness)
## Multiple R-squared:  0.1806, Adjusted R-squared:  0.1781
## F-statistic: 72.96 on 1 and 331 DF,  p-value: 4.897e-16
```

Take a moment to match up parts of the outputs that are the same. There is a difference here in that the sign of the test statistics differs. That does not matter as our t-distribution is symmetrical and we're doing a two-tailed test.

13.9.4.5 Mann-Whitney U

Similar idea to before, except for this test it is just rank not signed rank.

```
# Wilcoxon / Mann-Whitney U (multiple names)
wilcox.test(body_mass_g ~ sex, data = penguins)

##
## Wilcoxon rank sum test with continuity correction
##
## data:  body_mass_g by sex
## W = 6874.5, p-value = 1.813e-15
## alternative hypothesis: true location shift is not equal to 0
```

```
# As linear model with our dummy-coded group_y2:
summary(lm(rank(body_mass_g) ~ sex, data = penguins))

##
## Call:
## lm(formula = rank(body_mass_g) ~ sex, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -183.68  -74.13  -16.63   91.32  162.87
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  128.630      6.948  18.512  <2e-16 ***
## sexmale      86.051      9.783   8.796  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 89.25 on 331 degrees of freedom
## (11 observations deleted due to missingness)
## Multiple R-squared:  0.1895, Adjusted R-squared:  0.187
## F-statistic: 77.37 on 1 and 331 DF,  p-value: < 2.2e-16
```

13.9.5 ANOVA

13.9.5.1 Use case

You’ve probably seen ‘ANOVA’ in the context of model comparison, but it is also a popular test in psychology and other disciplines.

Let’s look specifically at one-way ANOVA (or the F-test). It tests if all the means for several groups (more than 2) are the same or if at least one is different.

I hope this sounds a bit like the next evolution from the independent t-test...

(+ 1000 stats respect points to anyone who draws Pokemon-esque evolutions of these three tests...with regression as the mega-evolution...)

13.9.5.2 Assumptions

And it just so happens that the assumptions for the one-way ANOVA (also called the F-test) are EXACTLY the same as for the independent t-test

There is a GIF in the web version. Not required content. <https://gifest.blogspot.com/2019/03/snl-hi-saturday-night-live-hey-kate.html>

1. The data are continuous.
2. The data are normally distributed (in each group).
3. Each group is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample
4. The variances for the groups are equal.

13.9.5.3 Hypotheses

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

H_1 : at least one μ differs from the others

13.9.5.4 Example

Let's now look at body mass across species. Suppose we wanted to know if was believable that the the means body mass in grams was the same across all three species. This is when we could fit a quick ANOVA to test this. The `aov()` allows us to do this.

```
summary(aov(body_mass_g ~ species, data = penguins))

##              Df    Sum Sq Mean Sq F value Pr(>F)
## species        2 146864214 73432107   343.6 <2e-16 ***
## Residuals     339  72443483   213698
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2 observations deleted due to missingness
```

That looks a lot like the output from calling `summary` on `lm()`...in fact, `aov` is just a wrapper for `lm`! Which means it has been linear regression the whole time.

13.9.5.5 We'll stop here and talk further about this particular topic in class this week.

Your next topic is [reproducible examples \(reprexes\)](#).

13.9.6 Credits

Credit to **Jonas Kristoffer Lindeløv** for the excellent resource this resource is based on. [There are more examples there than we will cover in this course.](#)

14

Module 2

15

Module 3

16

Module 4

17

Module 5

References

Appendix

18




Resources

18.1 Course tools overview

While we've tried to keep things as streamlined as possible, there are still several different tools we'll be using this semester. Your U of T login should work with all of them. The below PDF file provides an overview of how you'll be interacting with each one.

- At the bottom of the page is an embedded slideshow introducing you to the JupyterHub.
- You can always access Piazza from the Navigation Menu on the left.
- Instructions for setting up your U of T Zoom are on the Zoom page and links are in the Navigation menu and on the home page.

18.1.1 Admin

Logo	Description
	Quercus will be used for timed assessments, some submissions and announcements.
	Synchronous classes and office hours will be hosted via Zoom . You MUST join using your U of T Zoom account to be admitted. Get your account: utoronto.zoom.us
	Microsoft Forms will be used for several important administrative forms. You will need to be signed in to your U of T account in the same browser to access these.

18.2 Using RStudio with the JupyterHub

We will be using R through RStudio to conduct analyses in this course. If you have a local installation of R you are welcome to continue using that, but, for this course, you do not need to have R and RStudio installed. Instead, assessments and activities will be shared through the U of T JupyterHub. This gives you access to RStudio in your browser through your U of T login on any internet-connected device. It means you don't have to fight package installations and we can instead focus on the good stuff.

Please read through the following slides, experiment with the example sharing link, make sure you know how to knit an Rmd to pdf + export the pdf, and practice navigating and moving files.

Link: <https://rstudio-with-jupyterhub-uoft.netlify.app>.

18.3 Zoom, Zoom, Zoom, Zoom...

Access to STA303 synchronous meetings and office hours is restricted to our students.

[Set up your U of T Zoom account](#)

18.3.1 Make sure your Zoom is up to date

To participate fully, you will need Desktop client or mobile app: version 5.3.0 or higher. You can check your desktop client or mobile app version by following [these instructions](#).

18.3.2 Customize!

Once you have logged in, [please customize your profile](#):

- Update your name to your **preferred name** (what you would like us to call you in class) Note: this may not be allowed with your U of T settings, so don't worry if this doesn't work.
- Add a **profile picture** (please make it a photo of YOU or an avatar that looks like you...we don't want Snoopy or Joe Biden¹ in class)

18.3.3 VPN

There is a [University of Toronto VPN \(UTORvpn\)](#) that you have access to as a student. It may help with video quality and access to U of T resources.

If you are based in mainland China, the [Alibaba Cloud Enterprise Network \(CEN\)](#)[Links to an external site.](#) service should help with your Quercus access.

18.3.4 Notes:

1. Please always use your real name and face for this course, and be cautious about changing them and your virtual background for other meetings. A joke background for a call with family or friends may not be appropriate for class.
2. For class meetings, the settings will always be that your camera and microphone are off to begin with so you have the control to check these things first.

¹Yes, these are real images students have used.

3. We do ask that, when possible, you use your microphone in office hours, breakout groups and any other small group meetings and strongly prefer that you use your camera AND microphone. We trust you to make the best choice for your environment, comfort and learning.
4. You may get a “**This meeting is for authorized participants only message**”. Choose the “Sign in with SSO” option to sign in.

18.3.5 Changing your profile picture on Zoom and Quercus

Follow these instructions to add a profile picture (or bitmoji style avatar if you’d prefer) to [Quercus](#) and [Zoom](#). I want this experience to be more social and less faceless. Please don’t use photos of cartoon characters, etc. A good photo will be a close-up of your face so we can see who you are even when the photo is small.

18.3.6 What to do if you experience technical difficulties during class?

First, (if possible) send me a chat note that you’re having technical difficulties and are working to resolve them.

Second, leave the meeting and re-enter. This often resets things and resolves the problem. Before entering the meeting, make sure all of your devices are properly plugged in and Bluetooth devices are connected.

If that doesn’t fix things, exit the meeting again and update your Zoom Client. This is the Zoom software that should be on your computer. Here’s a short video tutorial explaining how to update the software: <https://www.youtube.com/watch?v=E7zERcVLUBM>.

After updating, enter the meeting again to see if this resolved your problems.

Our synchronous classes are recorded, so if your technology is just going catastrophically wrong, go get a cup of tea/coffee/water and relax, you can catch up with the recording when it is posted on Quercus.

18.3.7 What to do if your instructor or TA is experiencing technical difficulties on Zoom

First, check the **chat** to see if the instructor or the project mentor have said what is going on and what they are doing to fix things and follow any instructions they give.

Second, if they have disappeared completely, wait 10 minutes (or until the end of the meeting time, whichever comes first) before closing the call. (You can do other things in the meantime, but be ready to jump back in).

Third, expect to see an announcement on Quercus afterwards telling you what to do (e.g. it might be to watch a video I’ll record later, to review some slides or perhaps there is nothing to do and i’ll see you next time).

18.4 Student support services and resources

18.4.1 Mental health support

You may find yourself feeling overwhelmed, depressed, or anxious. Lots of people feel the same way. There is help available from mental health professionals 24 hours a day via online and phone-based services. Here are some that are available to U of T students:

- [MySSP - My Student Support Program](#) 1-844-451-9700, or outside of Canada call 001-416-380-6578
- [Good2Talk Student Helpline](#) 1-866-925-5454, or text GOOD2TALK to 686868
- [Distress Centres of Greater Toronto](#) 416-408-4357, or text 45645

There is also the new Navi tool for U of T students, it is a chatbot and your questions are totally anonymous. <http://uoft.me/navi>

The student union are also curating a list here: <https://www.utsu.ca/mental-health/>

18.4.2 General University resources

The following are some important links to help you with academic and/or technical service and support:

- **Health & Wellness** can help with appointments with a range of clinicians, nutrition, immunizations, sexual and reproductive health and much more. Many of their services continue to be available online.
- **Arts & Sciences** student resources through [Sidney Smith Commons Online](#)
- **General** student services and resources at [Student Life](#)
 - Tips for dealing with [multi-choice questions](#) (MCQs)
 - Book an appointment with a [learning strategist](#) (they can help you with strategies for MCQs also)
- Full **library** service through the [University of Toronto Libraries](#)
- Resources on **academic support** from the [Academic Success Centre](#)
- Learner support at the [Writing Centre](#)
- Information about [Accessibility Services](#)
- Quercus Information in the [Canvas Student Guide](#)
- Logistical and social support for **international students** at the [Center for International Experience](#)

Visit the A&S [online resources for students](#) page for resources available to support you through your online studies. If you have further questions, please email ask.artsci@utoronto.ca.

18.4.3 Financial support

A list of University financial supports, work-study opportunities, as well as provincial and federal government programs is available on the University's [Financial Support & Funding Opportunity directory](#).

18.4.4 Arts & Science COVID19 FAQ

The [Arts & Science Undergraduate FAQ page](#) addresses frequently asked questions that are specific to undergraduate students taking courses with the Faculty of Arts & Science. On this page you will find information for:

Messages from Dean Woodin can be found on the [A&S latest updates](#) page.

19

FAQs and Errata

19.1 Frequently asked questions

While [Piazza](#) is our main class question and answer board, this page will have some static Questions and Answers for frequently asked questions. Use Cmd + F (Mac) or Ctrl + F (PC) to search for keywords on this page.

Additionally, make sure you're familiar with the [Syllabus](#). I've tried to explain as much as I could there.

19.1.1 Course admin

19.1.1.1 STA303 pre-requisites

I didn't take STA302 or an equivalent course, can I still take STA303?

No, sorry. We enforce pre-reqs strictly. This isn't up to me. You will be removed from the course. Please **reach out to the UG stats team** (ug.statistics@utoronto.ca) about questions of this nature.

- [Book an appointment during their office hours.](#)
- For grad courses, please email grad.statistics@utoronto.ca.

19.1.1.2 Recorded lectures

Where can I find recorded lectures and how soon can I expect them to be available?

Recorded lectures will be linked on the [Course overview](#) page. I aim to have the links up within 24 hours of class. They take some time to process.

19.1.1.3 Sections

Do L0101 and L0201 cover the same materials?

Yes. The only difference is when your synchronous class is. See the below question in “Attending synchronous class”.

19.1.1.4 Attending synchronous class

Can I attend the synchronous class for the other section?

Yes!*

*If the number of attendees is getting too close to the call cap of 300 (this is a Zoom license thing), preference will be given to those enrolled in the session and others will be asked to leave. Any Team Up! activity bonuses will also be applied-section/session does not matter.

Why can't I access the class/office hour Zoom meeting?

You must have and be signed in with your University of Toronto Zoom account. Use the 'SSO' (single sign-on) option.

“Will STA303/STA1002 be able to be completed online-only?” Yes.

- **STA303** will be a flipped course, with content delivered online and opportunities for activities both in-person (subject to health advice) and online. All assessments will be completed and submitted online.
- If you are enrolled in an **in-person tutorial**, you will have an option to attend online instead.
- Synchronous attendance is NOT required to pass this course, but being able to attend (online) synchronously at the times in the timetable may make things easier for you.

“Where is your office?”/“Can I come see you in-person?” At this stage, I cannot offer in-person office hours nor student meetings in my office. Drop-ins are not currently allowed for any instructors or TAs in Statistical Sciences, unless they have told you they have organized another meeting space for this purpose.

19.1.2 Team Up!

19.1.2.1 Troubleshooting & FAQ advice from U of T CTSI

My screen seems to be lagging compared to those of my group members (eg. I'm not on the same question as the driver/members, the answer I chose does not appear on my group members' screens, etc.)

As long as you are logged in, you will receive your grade, so just continue to participate by communicating with your team. Almost always, your device will re-sync with the rest of your team's devices within a minute or two. If not, refresh your screen once. Repeatedly refreshing is not usually helpful!

I've been disconnected. Will my quiz progress be saved? Can I re-join my group?

Enter the Team Up! session again and you will automatically be put back into your group. Your progress will be saved, and you will return to the question you or your group was working on. This is true for the Driver as well as any team member.

How do I send my completed quiz results?

Click the large red “Submit to Quercus” button at the end of the Team Up! quiz.

Can we change our group driver?

You can request a driver change by pressing the red exclamation mark in the top right of your Team Up! quiz.

How is the driver for our group chosen, and how can the driver pass the group ID to others in remote classes?

Group members can decide who will be the Driver (ideally someone with a good internet connection). The Driver can pass the group ID to other members verbally through chat or microphone in breakout rooms.

The Driver of my group has to leave unexpectedly or their device has stopped functioning. How do we proceed with the quiz?

Request for a Driver Change using the red button at the top right of your Team Up! Session. This button only provides help for Driver Changes. You may also need to speak with your instructor or TA.

19.2 Other

19.2.1 References

Can you (Prof. Bolton) write me a reference? I'm desperate!

Please read my personal policy [here](#) to get a sense of under what circumstances I could write for you, but basically, a good mark in one class is not sufficient and you need to have at least two ‘activities’ with me. If you believe you meet my basic criteria, you can request a reference from me [here](#). I will then accept or decline based on the the information provided.

Do you (Prof. Bolton) have any research opportunities available?

Research/work study/teaching assistant opportunities:

- See the [Department website](#). The main round of TA recruitment occurs during the summer but there are occasionally emergency postings.
- Some information about opportunities with me on my [website](#).
- **Reading courses:** STA496/497: Readings in Statistics must be registered for as part of [special enrolment during July](#). I will not be taking on any further students.
 - If you’re thinking about the future, I usually consider taking on a small number of STA497 students for a **half-credit, year long version**.
 - There is much more information about my past students, research interests and what a course with me might be like on my website: [<https://www.lizabolton.com/reading_courses.html>](https://www.lizabolton.com/reading_courses.html).

19.3 Errata

ID	Location	Note

20

Bits and pieces

20.1 Code to generate course art

```
# install.packages('readr')
# install.packages('tidyverse')
# install.packages("devtools")
# devtools::install_github("BlakeRMills/MetBrewer")

library(readr)
library(MetBrewer)
library(tidyverse)

course_code <- "STA303"

my_colours <- c(met.brewer("Cross", n = 8), met.brewer("Cross", n = 9))

set.seed(parse_number(course_code))

ngroup=17
names=paste("G_",seq(1,ngroup),sep="")
DAT=data.frame()

for(i in seq(1:30)){
  data=data.frame( matrix(0, ngroup , 3))
  data[,1]=i
  data[,2]=sample(names, nrow(data))
  data[,3]=prop.table(sample( c(rep(0,100),c(1:ngroup)) ,nrow(data)))
  DAT=rbind(DAT,data)
}

colnames(DAT)=c("Year","Group","Value")
DAT=DAT[order( DAT$Year, DAT$Group) , ]

ggplot(DAT, aes(x=Year, y=rev(Value), fill=Group )) +
  geom_area(alpha=1 )+
  theme_bw() +
  scale_fill_manual(values = my_colours)+
  theme(
    text = element_blank(),
    line = element_blank(),
    title = element_blank(),
    legend.position="none",
```

```

panel.border = element_blank(),
panel.background = element_blank(),
plot.margin = margin(0, -2.7, 0, -2.7, "cm"))

ggsave(paste0(course_code, "-base.png"), width = 24, height = 2)

```

20.2 M1 supporting information on matrices (not assessed)

20.2.1 Background

20.2.1.1 Some true things about matrices

- The **rank** of a matrix is the number of linearly independent columns your matrix has.
- If the number of columns = the rank of the matrix all the columns are **linearly independent**. If the number of columns is $>$ the rank of the matrix, all the columns are *not* linearly independent.
- You can only **invert** a square matrix if all its columns are linearly independent. (Determinant non-zero).

Why do we care? In linear regression, to estimate β , our vector of coefficients, we calculate $(X^T X)^{-1} X^T Y$. The elements of β can't be estimated if $X^T X$ (a square matrix) isn't invertible.

Clarification to what I said in the tests activity: We usually perform regression with an intercept because we don't want to assume our line passes through the origin, **0**. So, if there is an intercept, (column of 1s in the model matrix) we must convert every categorical variable with k levels into $k - 1$ dummy variables to have the intercept and still satisfy linear independence. If we ditch the intercept, we can have k dummies, but this is only usually useful in the specific case of ANOVA.

20.2.2 Example

```

## # A tibble: 12 x 2
## # Groups:   species [3]
##   body_mass_g species
##   <int> <fct>
## 1      3900 Adelie
## 2      3700 Adelie
## 3      3450 Adelie
## 4      3175 Adelie
## 5      4500 Chinstrap
## 6      3850 Chinstrap
## 7      3650 Chinstrap
## 8      4550 Chinstrap
## 9      5150 Gentoo
## 10     5000 Gentoo
## 11     5700 Gentoo
## 12     4600 Gentoo

## # A tibble: 12 x 4
##   body_mass_g species.Adelie species.Chinstrap species.Gentoo
##   <int>         <dbl>         <dbl>         <dbl>
## 1      3900           1           0           0
## 2      3700           1           0           0
## 3      3450           1           0           0
## 4      3175           1           0           0

```

```
## 5      4500      0      1      0
## 6      3850      0      1      0
## 7      3650      0      1      0
## 8      4550      0      1      0
## 9      5150      0      0      1
## 10     5000      0      0      1
## 11     5700      0      0      1
## 12     4600      0      0      1
```

20.2.2.1 Classic regression, k-1 dummies (intercept)

```
##
## Call:
## lm(formula = body_mass_g ~ species, data = pengwings)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -512.50 -310.94  -34.37   348.44   587.50
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3556.3      206.8   17.199 3.42e-08 ***
## speciesChinstrap    581.2      292.4    1.988 0.07809 .
## speciesGentoo    1556.2      292.4    5.322 0.00048 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 413.6 on 9 degrees of freedom
## Multiple R-squared:  0.7627, Adjusted R-squared:  0.71
## F-statistic: 14.46 on 2 and 9 DF, p-value: 0.001545

##      (Intercept) speciesChinstrap speciesGentoo
## 1             1             0             0
## 2             1             0             0
## 3             1             0             0
## 4             1             0             0
## 5             1             1             0
## 6             1             1             0
```

Does the rank of the model matrix equal the number of columns?

```
## [1] TRUE
```

Okay, linearly independent, we're good to go!

20.2.2.2 Classic regression but trying to force k dummies (with an intercept)

```
##
## Call:
## lm(formula = body_mass_g ~ species.Adelie + species.Gentoo +
##      species.Chinstrap, data = wider)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```



```
## -512.50 -310.94 -34.38 348.44 587.50
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4137.5      206.8   20.010 9.04e-09 ***
## species.Adelie  -581.2      292.4   -1.988 0.07809 .
## species.Gentoo   975.0      292.4    3.334 0.00874 **
## species.Chinstrap    NA         NA      NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 413.6 on 9 degrees of freedom
## Multiple R-squared:  0.7627, Adjusted R-squared:  0.71
## F-statistic: 14.46 on 2 and 9 DF,  p-value: 0.001545

##      (Intercept) species.Adelie species.Gentoo species.Chinstrap
## 1           1           1           0           0
## 2           1           1           0           0
## 3           1           1           0           0
## 4           1           1           0           0
## 5           1           0           0           1
## 6           1           0           0           1
## 7           1           0           0           1
## 8           1           0           0           1
## 9           1           0           1           0
## 10          1           0           1           0
## 11          1           0           1           0
## 12          1           0           1           0
## attr(,"assign")
## [1] 0 1 2 3
```

```
## [1] 3
```

Does the rank of the model matrix equal the number of columns?

```
## [1] FALSE
```

Not linearly independent. Why?

Well, this intercept column is a linear combination of the three species columns!

```
##      1      2      3      4      5      6      7      8      9     10     11     12
## TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

20.2.3 Regression NOT classic, actually ANOVA! (no intercept)

```
##
## Call:
## lm(formula = body_mass_g ~ 0 + species.Adelie + species.Gentoo +
##     species.Chinstrap, data = wider)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -512.50 -310.94 -34.38  348.44  587.50
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## species.Adelie    3556.2      206.8   17.20 3.42e-08 ***
## species.Gentoo     5112.5      206.8   24.73 1.39e-09 ***
## species.Chinstrap  4137.5      206.8   20.01 9.04e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 413.6 on 9 degrees of freedom
## Multiple R-squared:  0.9932, Adjusted R-squared:  0.9909
## F-statistic: 435.8 on 3 and 9 DF,  p-value: 4.659e-10

##      species.Adelie species.Gentoo species.Chinstrap
## 1              1              0              0
## 2              1              0              0
## 3              1              0              0
## 4              1              0              0
## 5              0              0              1
## 6              0              0              1
## 7              0              0              1
## 8              0              0              1
## 9              0              1              0
## 10             0              1              0
## 11             0              1              0
## 12             0              1              0
## attr("assign")
## [1] 1 2 3
```

```
## [1] 3
```

Challenge question for +100 stats respect points: How do you interpret these coefficients?

Does the rank of the model matrix equal the number of columns?

```
## [1] TRUE
```

Great, back to being linearly independent.

20.2.4 Further reading (if you want it)

As I said at the beginning, I'm not planning to assess you on any of this. If you're interested in knowing more, or just think matrix algebra is delicious, I think this is a delightfully approachable walk through. <https://online.stat.psu.edu/stat462/node/132/>