R Notebook

Course guide

Prof. Liza Bolton

Winter 2022

Contents

1

How to use this course guide

This course guide has been created in R using the `bookdown` package.

In the web version, you can put this guide into dark mode and change the font style, if you find a different display preferable. Play around with the buttons at the top of the screen and see what they do.

You can download a PDF version of the whole guide with the download button (on the web version, look for the PDF icon button at the top). Note that *many* parts of this guide will be updated as the course proceeds, so the PDF might not be that useful to you unless you are separating out relevant portions for your annotations/notes, etc. Alternatively, to get just a specific chapter, you can 'Print to PDF' from your browser, and this is also how you can get a PDF copy of any linked/embedded slides. Shortcut: Cmd+P or Ctrl+P, and select 'Save as PDF' (or similar).

1.1 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.

See the communication policy in the syllabus for more information.

1.2 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.

1.2.1 Tutoring companies

Tutoring companies do not have any right to suggest they are associated with this course. There is extensive support available within the course, department and university for all students. **It is the opinion of the author that many tutoring companies are a predatory scam and a potential risk to students' academic integrity.** This statement should be prominently included by any such company seeking to share this resource.

1.3 Contributors

Big thanks to Amin Banihashemi and Ian Richter for their contributions to this course guide.

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 p.m.—5:00 p.m. ET

Tutorials: Alternating Thursdays

- 12:00 p.m.—1:00 p.m. ET or
- 5:00 p.m.—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! • 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 • 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 7–18 Module 3: Linear mixed models • 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 • 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 • Generalized linear mixed models • Generalized additive mixed models	**MAR 31** Professional dev evidence and reflection due
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. (Grad students have a third option)
- A-tisket, a-tasket[2], 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 5%.

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 (2% 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.

[1] The 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

[2] A-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/

[4] Mixed assessment windows are 8-8, I couldn't 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 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 requested **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 components) 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*	**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!
Missed an assessment due to illness or personal emergency? **Want to request a regrade of an assessment?**	**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).
Personal/sensitive circumstances? (i.e., something which is not appropriate to share with the whole class)	**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>.

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/>.

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 is 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 the 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!)

3.2 A few things to know upfront

- **R** is the programming language used in this course.
- I use and teach a lot of `tidyverse`, especially `dplyr` for data wrangling and `ggplot` for visualization.
  - If you're not familiar with these, consider getting ahead by checking out:
    - chapters 3 and 5 of R for Data Science by Hadley Wickham & Garret Grolemund, and
    - the DoSS Toolkit, especially the modules 'Hand me my plyrs' and To ggplot or not to ggplot.
- The entire course is online until Jan 31. While some elements of the delivery of STA303 may be in person after that, the course will have a large online component regardless and **all assessments will be able to 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.
- I *cannot* add students to a course, waive prerequisites or move them off a waitlist. If you or a friend have any questions of this nature, please reach out to the UG stats team (ug.statistics@utoronto.ca).
  - Waitlisted STA303 students may access course materials by filling out this form, but they will not be able to complete assessments during this time. The course is designed to accommodate this.
- Check out my autoresponder message post for FAQs and other useful information.

3.2.1 Students joining off the waitlist

You do not have to submit any assessments you missed while on the waitlist. They are all knowledge basket assessments and you will have multiple future opportunities to make up those grades. See the Syllabus for more information.

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

3.3 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 guide and/or in Quercus.

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

3.3.0.1 In most modules there be will be:

- Asynchronous content released at the beginning.
- A synchronous class on Wednesday at 10:00 a.m. ET (L0101) and 3:00 p.m. ET (L0201).
  - Both sessions will be the same, you only need to attend one.
  - You must be logged in with your U of T Zoom account (utoronto.zoom.us) to access the 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.

- A knowledge basket writing task.
  - Create phase due first Friday at 3:03 p.m. ET.
  - Assess phase due second Tuesday at 3:03 p.m. ET.
  - Reflect phase due second Friday at 3:03 p.m. ET.

- Office hours.
  - **Prof office hours** will occur during the second half of the Wednesday synchronous classes, i.e. approximately 11:10–12:00 p.m ET and 4:10–5:00 p.m., in the same Zoom call.
  - **TA office hours** will take place during the Thursday tutorial times (12:00 p.m. and 5:00 p.m. ET) on the *second* Thursday of each module.

3.4 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.

3.4.1 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.

3.5 To do now

- Read the Syllabus (pdf version here)
- Join the Piazza discussion board.
- Understand the tools we will be using in this course.
- 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/. Update it if needed, see more here.

Assessments overview

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. (Grad students have a third option, see below.)
- A-tisket, a-tasket[2], 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 5%.

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 (2% 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.

[1]The 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

[2]A-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.

Portfolio assessments aim to help you demonstrate your technical coding, statistical thinking, communication and reflection skills. This mini-portfolio also aims to recap and refresh knowledge from your previous statistics courses as well as building your ability to create quality data visualizations.

5.1 General instructions

- Be very careful to follow instructions on variable naming. If you do not, your code won't pass autograding and you will not receive the grades. This will not be eligible for regrading requests.
- Comment your code! In an R code chunk comments start with a # (pound sign or hashtag symbol). Don't confuse this with the use of # to denote different levels of headings in the text parts (governed by Markdown syntax) of an R Markdown document.
- You should neatly format your code. No strict style is required, but make sure it is easy to read your work. (Optional) You may find Google's R Style Guide or the Tidyverse Style Guide interesting.
- DO include your code in the body of the PDF itself (don't set echo=FALSE, don't hide/suppress etc.). Note that this is *different* to what you will be asked to do in the final project or in professional reporting. This is a demonstration of your skills.
- If asked to write and answer in words, write in full sentences.
- KNIT EARLY AND OFTEN! Don't leave things till the last minute, your Rmd not knitting is not an emergency for which an extension will be granted.

5.1.1 Template

You can access the template for this assessment here.

There is currently a lot of 'filler text' and 'filler code' in the template that you will want to **delete.** Fun fact: All filler text sourced from Hipster Ipsum, which Katy Wang in the UG Stats office introduced me to.

5.2 Submission instructions

- Submit both your Rmd (must be called: sta303-w22-mini-portfolio.Rmd) and PDF (must be called: sta303-w22-mini-portfolio.pdf) on MarkUs.
- You do not need to submit any data or tex files.

5.3 Cover page

You don't have to use the provided template, but you DO need to write your mini-portfolio in RMarkdown and include a cover page. The cover page must have:

- A title and subtitle (you can use my examples in the template or update them as you see fit, no points for changing, but it might make it a better part of your own website or GitHUb portfolio, in the future)
- Your name
- Date (assessment submission date is fine)

In the template, you can change the colour of this cover to any colour you would like by replacing 6C3082 in the YAML (line 11) to another hex code. You could use this tool to help you: https://htmlcolorcodes.com/color-picker/

5.4 Introduction

Write this section last or second to last (before the refection).

In the introduction section, write a brief summary of the skills you have demonstrated in this mini-portfolio, across the statistical skills sample, writing sample and reflection sections. Think of it like a **cover letter** for this document. It should be appropriate for a fairly general audience—imagine a future employer reading this. You may want to briefly explain the course context, as you understand it. What is STA303/1002 about? (Consider the learning objectives in the syllabus)

Your introduction should be **no longer than 300 words** and must **fit on one page**.

5.5 Statistical skills sample

5.5.1 Task 1: Setting up libraries

Set up a chunk called `setup` where you load the `tidyverse` and `readxl` libraries. Set your chunk options to `message=FALSE` so all the package loading information isn't included in your our output. You will need to make sure you run this chunk each time you start a new session so you can use many of the functions required.

5.5.2 Task 2: Visualizing the variance of a Binomial random variable for varying proportions

Goal: Show visually that for a fixed value of n, $p = 0.5$ will result in the largest variance for a Binomial random variable.

- Choose two appropriate values of $n > 0$ for your demonstration and save them as `n1` and `n2`
- Create a vector of proportions, `props`, from 0 to 1, in steps of 0.01 (Tip: use the `seq()` function).
  - If I suggest a function you haven't seen before, you can search its documentation in your console by typing a ? in front of the function name, e.g. ?seq.
- Create a tibble (a data type in R, like a dataframe), `for_plot`, with the vector of `props` as the first variable, and two additional variables calculating the variance for each of your two chosen n values. (Call these `n1_var` and `n2_var`)
- Create **two** plots, one for each of your values of n, using `ggplot` and apply `theme_minimal()` to each one. These should appear in your PDF and do not need to be saved with specific name.
  - Add an appropriate **figure caption** to each chart (use fig.cap="Your text here" in the R chunk settings).
    - An appropriate title should succinctly explain the chart and mention the chosen n value.
  - Add a **caption** *within* the ggplot that says "Created by STUDENT NAME in STA303/1002, Winter 2022."
  - Give the x and y labels appropriate **labels**.

5.5.2.1 Objects that must be carefully assigned to pass autograding

- `n1` and `n2` (both should be integer vectors of length 1)
- `props`, a numeric vector
- `for_plot`, a tibble with three correctly named columns

5.5.3 Task 3: Demonstrating frequentist confidence intervals as long-run probabilities of capturing a population parameter

Goal: Simulate a population of size 1000, using $N(10, 2)$, and take 100 independent, random samples of size 30 observations each from it. Calculate a Wald confidence interval (using an appropriate t-multiplier) for the population mean from each sample. Calculate what proportion of intervals contain the population mean and plot all these intervals, coloured by successful population mean capture or not.

5.5.3.1 Specific steps

- Set the seed to the last three digits of your student ID (this is your numeric student identifier number, NOT your UTORID).
- Set up the following objects with the appropriate simulation parameters, sample size and number of samples.
  - `sim_mean` and `sim_sd`
  - `sample_size`
  - `number_of_samples`
- Calculate the appropriate t-multiplier (function: `qt(...)`) for constructing a 95% confidence interval in this context. Make sure your degrees of freedom are appropriate. Save it as `tmult` for later use.

- Create a vector called `population`, a simulated population using `sim_mean` and `sim_sd` and with 1000 values (function to generate random numbers from a normal distribution: `rnorm(...)`).
- Find the *actual* true mean for your population and save it as `pop_param`. This should be a numeric vector of length 1.
- Get 100 samples of size 30 from your population and save them in a vector called `sample_set`. (This might be a little tricky/unfamiliar, so here is one way to do it. You can just copy and paste this code.)

```
sample_set <- unlist(lapply(1:number_of_samples,
    function (x) sample(population, size = sample_size)))
```

- Create a new vector called `group_id` that will allow you to label the values from the 100 different samples above. Hint: `rep(…)` will be useful here, and has a great little argument called `each`. Take a look at the documentation (`?rep`) to compare the behaviour of `times` and `each`.
- Create a new `tibble` (a data type in R, like a dataframe), `my_sim`, that has two columns:`group_id` and `sample_set.`
- Create a new tibble, `ci_vals`, that starts with the dataset `my_sim` and then groups by `group_id`, and summarizes appropriately to create two new columns: `mean` and `sd`, that hold the means and standard deviations. There should be one row per group.
- Continue to change the tibble `ci_vals` by adding the following variables:
  - `lower` and `upper`, two columns that hold the lower and upper bound of a 95% confidence interval for the group. You will need to calculate this. Consider the equations for a confidence interval and remember that we are using a t-multiplier that you have already calculated.
  - `capture` which takes the values `TRUE` if the population parameter is in the 95% CI, and `FALSE` if not. These should be logical NOT character types.
- Create an object called `proportion_capture` that uses `ci_vals` and stores the proportion of intervals you created that capture the population parameter. This should be done using the object names, not 'hard coded,' so that if you changed your set.seed or your sample size, etc., and run all the code again, it would update this value. It should be a vector of length 1.
- Plot these 100 confidence intervals in one plot, with the means indicated as points, as well as and a dotted line for the population parameter. `geom_errorbar()` will be very helpful.
  - **Colour** the confidence intervals by whether or not they contain the population parameter.
    - If the interval include the population parameter, colour it #122451 if it DOES contain the parameter (TRUE) and #B80000 if it DOES NOT (FALSE).
  - Set the **figure caption** to "Exploring our long-run 'confidence' in confidence intervals. This figure shows how often 95% confidence intervals from 100 simple random samples capture the population mean. The population was simulated from N(10, 2)."
  - Set the caption to "Created by STUDENT NAME in STA303/1002, Winter 2022." Replace STUDENT NAME with your name.
  - Set the **legend title** to "CI captures population parameter."
  - **Flip the coordinates** using `coord_flip()` so the intervals are horizontal across your chart.
- Add the following sentence to your Markdown (not in a code chunk):
  - `` `r proportion_capture*100` `` % of my intervals capture the the population parameter.
  - Important! Use BACKTICKS around r proportion_capture*100. Things you put in backticks with an r at the beginning will be processed as inline R code. This key is located under the escape button on many keyboards.You may need to remove some spaces compared to the above.
- Briefly (~1 paragraph) describe why we *can* include the population parameter in this plot AND why we cannot usually compare the population parameter to our confidence interval in practice (e.g., when working with data that has not been simulated). Write this for a non-statistical audience.

5.5.3.2 Objects that must be carefully assigned to pass autograding

- `sim_mean` and `sim_sd`
- `sample_size`
- `number_of_samples`
- `tmult`
- `population`
- `pop_param`
- `group_id`
- `sample_set`
- `my_sim` , tibble with two columns
- `ci_vals`, tibble with 6 columns (after all steps completed) (See Errata for note)
- `proportion_capture`

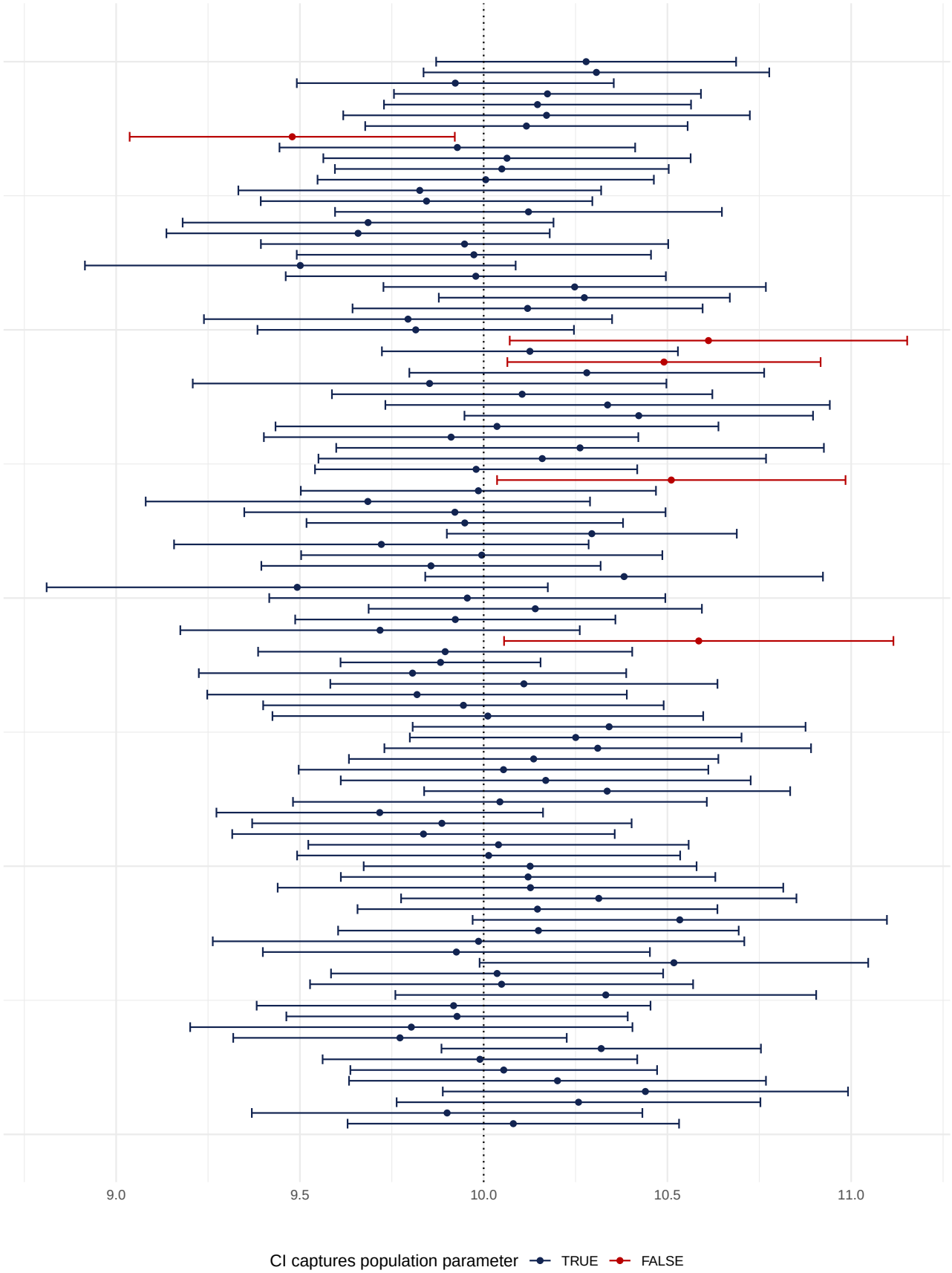


Figure 5.1: This is generally what your output should look like for the confidence interval task. Note: It won't be exactly the same.

5.5.4 Task 4: Investigating whether there is an association between cGPA and STA303/1002 students correctly answering a question on global poverty rates

In the 'getting to know you' survey at the beginning of STA303, students who participated in the survey were asked:

- What their current cumulative grade point average (CGPA) was at U of T.
- Whether the proportion of people living below the global poverty line had halved, doubled or stayed about the same in the last 20 years.

So, what was the correct answer to the question about global poverty? The proportion of the global population living below the poverty line has HALVED!

In the .xlsx file called `sta303-mini-portfolio-poverty.xlsx` there are 200 observations that represent the patterns in our class but from which is is not possible to identify individual students. The final goal of this task is to test whether there is a difference in cGPA between students who correctly answered this question and those who did not.

5.5.4.1 Goal

- Briefly describe the goal of this task to someone who has not read these instructions and only has a general awareness of hypothesis testing. **Put it in your own words.**

5.5.4.2 Wrangling the data

- Load the data, `sta303-mini-portfolio-poverty.xlsx`, into an object called `cgpa_data`, and apply the `clean_names()` function from the `janitor` package. Pay attention to the path to the file when providing it to R. It won't know to look in the data folder if you don't tell it.
- Rename the cGPA variable to `cgpa` and the poverty question answer to `global_poverty_ans`.
- Clean the data so only appropriate cGPA variables are included.
- Create a new variable called `correct` that takes the value TRUE (logical, not character) if the respondent answered 'Halved' and FALSE if they answered 'Doubled' or "stayed about the same'.

All of the above changes should be saved into the dataset called `cgpa_data`. I recommend using pipes `%>%`. There is a keyboard shortcut too: Cmd+Shift+M (Mac) or Ctrl+Shift+M (Windows).

5.5.4.3 Visualizing the data

- Create a set of histograms, in one figure, positioned on top of each other, that will allow you examine the data in a useful way.

5.5.4.4 Testing

- Choose an appropriate test to test whether there is an association between cGPA and if a student in STA303/1002 answered this question correctly. JUSTIFY your choice appropriately.
- Conduct the test AND the equivalent version using `lm()` **interpret the result of the test appropriately** in a sentence or two.

Recall the types of tests you have now encountered, in this and previous courses:

- T-test `t.test(x, data = my_data)` or `t.test(x~y, data = my_data)`

- ANOVA 'summary(aov(x~y, data = my_data))
- Wilcoxon test `wilcox.test(x, data = my_data)`
- Mann-Whitney U `wilcox.test(x~y, data = my_data)` (note different use of wilcox function)
- Kruskal-Wallis Rank Sum test `kruskal.test(x~y, data = my_data)`

Make sure you're considering the assumptions for the tests as well as some of the other criteria about which tests we choose to reach for. You may need ask some questions in office hours or find relevant readings to help you consider the non-parametric tests. Your strategies here might be something to consider in your reflection.

5.6 Writing sample

5.6.0.1 Prompt

Read the below job ad and write about the skills you would need to apply. Write on at least 2 soft skills and 2 analytic skills. The ad targets those with a MSc/PhD or those with 2+ years of experience, but imagine this requirement is not there.

1. **Soft skills**. What soft skills relating to communicating and working with others does the company seek?
    1. In what way do you already possess two of these skills?
    2. What evidence do you have of possessing one or two of these skills?
2. **Analytic skills.** What analytic skills relating to software use and performing data analysis does the company seek?
    1. In what way do you already possess two of these skills?
    2. What evidence do you have of possessing one or two of these skills?
3. **Connection to studies.** What other skills can you develop and what evidence can you accumulate during the remainder of your education to be ready for a similar job ad?

Structure your answer under five headings: 'Introduction,' 'Soft skills,' 'Analytics skills,' 'Connection to studies' and 'Conclusion.' Write your answers in full sentences, with appropriate paragraphing. There should be a brief introduction and conclusion. Imagine a future employer or graduate school admission officer was reading this, you should explain what you are going to do, do it (the three headings) and then sum up what you did.

Word count: 300–500 words. **Please add a statement of your word count at the end of the passage.**

Figure 5.2: Yelp logo

5.6.0.2 Job add

5.6.0.2.1 Data Scientist (Remote) **Category** Data Science & Analytics

Location Toronto, Ontario, Canada

Department Engineering and Product

At Yelp, it's our mission to connect people with great local businesses. Yelp's unique dataset contains billions of interactions between users and local business around the globe, from a user reviewing a neighborhood coffee shop to requesting a repair quote with a photo of a leaky faucet. Data Scientists at Yelp work to make sense of these interactions to deliver impactful analyses and products to our users, business partners and the general public.

The Data Science team performs analyses, builds models, and designs experiments that directly impact Yelp's business and users. Our centralized team is the most wide-ranging consumer of data at Yelp, adept at tasks from modeling content growth and user behavior to sharing insights about the health of local economies. With varied backgrounds and expertise, we strive for learning and growth in a collaborative environment.

We'd love to have you apply, even if you don't feel you meet every single requirement in this posting. At Yelp, we're looking for great people, not just those who simply check off all the boxes.

This opportunity is fully remote and does not require you to be located in any particular region. We welcome applicants from throughout Canada.

We Are Looking For:

- 3+ years of experience as a data scientist or MS/PhD and 2+ years of industry experience in a quantitative role.
- Fluency with SQL and Python or R for data analysis.
- Solid understanding of statistical inference, experimental design and analysis.
- Enthusiasm for clean code and sharing reproducible results.
- Communication skills to work with partners on engineering, product and business teams.
- An eye for great data visualization with Matplotlib, Plotly, ggplot, or Tableau.
- If you don't have 2+ years of industry experience in a quantitative role, please take a look at our College Data Scientist roles instead!

Where You Come In:

- Define key metrics to track Yelp's performance and inform product decisions.
- Assess and frame questions from partners into actionable deliverables.
- Design, execute, and analyze complex experiments impacting millions of users.
- Devise and evaluate models for diverse business needs, such as identifying growth opportunities, personalizing user experience, and matching consumers to businesses.
- Own analyses start-to-finish and communicate key insights to stakeholders.
- Share your technical skills to develop and maintain high-quality, reusable analysis tools.

5.6.0.3 LI-Remote

At Yelp, we believe that diversity is an expression of all the unique characteristics that make us human: race, age, sexual orientation, gender identity, religion, disability, and education — and those are just a few. We recognize that diverse backgrounds and perspectives strengthen our teams and our product. The foundation of our diversity efforts are closely tied to our core values, which include "Playing Well With Others" and "Authenticity."

We're proud to be an equal opportunity employer and consider qualified applicants without regard to race, color, religion, sex, national origin, ancestry, age, genetic information, sexual orientation, gender identity, marital or family status, veteran status, medical condition or disability.

We are committed to providing reasonable accommodations for individuals with disabilities in our job application process. If you need assistance or an accommodation due to a disability, you may contact us at accommodations-recruiting@yelp.com or 415-969-8488.

Note: Yelp does not accept agency resumes. Please do not forward resumes to any recruiting alias or employee. Yelp is not responsible for any fees related to unsolicited resumes.

Accessed from **Yelp®** on Feb 28, 2021, 15:15 PM

Large logo from Wikipedia https://en.wikipedia.org/wiki/Yelp#/media/File:Yelp_Logo.svg

5.7 Reflection

Briefly, 100 to 200 words each, answer the following questions:

- What is something specific I'm proud of in this mini-portfolio?
- How might I apply what I've learned and demonstrated in this mini-portfolio in future work and study, after STA303/1002?
- What is something I'd do differently next time?

5.8 Rubric

There is a breakdown of the overall contribution of each component to your grade below, followed by a more detailed rubric.

Component	Percentage for component
Cover page	1%
Introduction	5%
Statistical skills sample	60%
Commenting of code	1%
Readability of R code	1%
Reproducibility	2%
Writing sample	20%
Reflection	5%
General (multiple components)	5%

Component	Subcomponent	Poor or missing (0%)	Adequate (50%)	Good (75%)	Excellent (100%)	Points
Cover page	Cover page	Not all required cover page components are present	N/A	N/A	All required cover page components are present	1
Introduction	Introduction	Missing OR significantly off topic OR totally inappropriate for audience OR longer than one page.	Some evidence of understanding of the course and the skills demonstrated in this assessment shown, but lacks clarity/specificity. May be someone inappropriate for specified audience. No longer than one page.	Introduction provides a mostly specific introduction to the mini-portfolio, explaining the context of the work and describing the skills demonstrated with mostly reasonble understanding. It is no longer than one page.	Introduction provides a succint and specific introduction to the mini-portfolio, explaining the context of the work and comprehensively describing the skills demonstrated. It is no longer than one page.	5
Statistical skills sample	Task 1	Missing or significant issues with completion of instructions for this task.	Correct libraries loaded but several instructions not followed.	Correct libraries loaded, but chunk may not be named or printed in the PDF appropriately.	Correct libraries loaded without package messages being included in the PDF, and the chunk is named 'setup'. This code chunk appears in the PDF, under the appropriate heading.	1
Statistical skills sample	Task 2 code and chart	Missing or goal not achieved.	A reasonable attempt made to complete the instructions, but with so many omissions or error that the goal is not fully achieved.	Most instructions correctly followed and mostly appropriate charts produced, but some major omissions or errors or a large number of minor errors.	All instructions correctly followed and appropriate charts produced. A small number of minor errors allowable.	5
Statistical skills sample	Task 3 code and chart	Missing or goal not achieved.	An attempt made to complete the instructions, but appropriate plot not created generated.	Most instructions correctly followed and a mostly appropriate chart produced, but some major omissions or errors or a large number of minor errors.	All instructions correctly followed and appropriate chart produced. A small number of minor errors allowable.	10
Statistical skills sample	Task 3 writing	Missing or more than one of the issues in the Adequate column.	Statement of % of intervals that capture the population parameter is incorrect OR hardcoded OR major incorrect claims made OR part of prompt not addressed.	Statement of % of intervals that capture the population parameter is correct and set programmatically (not hardcoded). A mostly clear and correct explanation of why we can show the pop parameter on this chart AND appropriate discussion of why this is not usually the case in practice. Minor omissions allowable, but should not make any incorrect claims.	Statement of % of intervals that capture the population parameter is correct and set programmatically (not hardcoded). Clear and correct explanation of why we can show the pop parameter on this chart AND appropriate discussion of why this is not usually the case in practice.	12
Statistical skills sample	Task 4 goal	Missing or more than one of the issues in the Adequate column.	Attempt made to put goal in own words, but paraphrasing insufficient OR description of goal includes many errors OR description inappropriate for the audience.	Goal correctly stated and put into own words, but may lack some clarity OR not be somewhat inappropriate for the audience.	Goal clearly, correctly and succinctly stated in the writers own words, in a way that is appropriate for the specified audience.	5

(continued)

Component	Subcomponent	Poor or missing (0%)	Adequate (50%)	Good (75%)	Excellent (100%)	Points
Statistical skills sample	Task 4 wrangling	Missing or goal not achieved.	An attempt made to complete the instructions, but final dataset	Most instructions correctly followed.	All instructions correctly followed.	5
Statistical skills sample	Task 4 viz	Missing or significant issues with completion of instructions for this task.	An attempt at a plot that would fulfil the instructions, but many errors or omissions. (Carryover errors from wrangling ignored.)	Mostly correct plot produced. (Carryover errors from wrangling ignored.)	Fully correct plot produced. (Carryover errors from wrangling ignored.)	7
Statistical skills sample	Task 4 test	Incorrect test chosen OR major errors in interpretation of p-value.	Correct test chosen, but without sufficient justification OR p-value description incorrect/missing OR incorrect test chosen but everything else done to good level.	Correct test chosen but justification somewhat lacking OR lm() implementation missing OR p-value interpretation had minor errors OR incorrect test chosen but everything else done to excellent level.	Correct test chosen and clearly justified. Both versions (test function + lm()) correctly implemented and interpretation of p-value correctly put into plain language.	15
Commenting of code	Commenting of code	No relevant comments or some comments, but limited or inappropriate.	Some reasonable comments, though they would be only limited use to a reader.	Reasonable comments, though may be unhelpful to a reader in some sections.	Code is well-commented and so that a new reader could pick up and get a strong idea of what was being done, where—even if they are not an R coder.	1
Readability of R code	Readability of R code	No R code in document OR major issues in clarity of code as to be mostly unreadable.	Code is mostly well formatted, but some inappropriate choices make some sections difficult to read.	N/A	Code is well formatted and easy to read.	1
Reproducibi	Reproducibi	Rmd does not successfully knit OR was not included in the upload.	Minor changes need to be made to the Rmd to successfully knit to PDF AND/OR PDF not included in submission.	N/A	Rmd can be run to recreate the submitted PDF exactly AND both Rmd and PDF were submitted. E.g., Seeds set correctly and no errors that prevent knitting.	2
Writing sample	Addresses 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.	8
Writing sample	Structure	No response OR there is no structure, very difficult to follow.	Some structure but difficult to follow.	The organization follows a mostly logical structure, with some issues in paragraphing OR meets criteria for excellent but word count not included at the end.	Well organized, follows a logical structure. Word count included at the end.	2
Writing sample	Introduction	No response OR there is no concluding sentence(s).	An introduction is present but is only weakly connected to the topic/body of the writing sample or very somewhat generic.	An introduction is present but may lack some connection to the topic/body of the writing sample or is somewhat generic.	A clear introduction succinctly introduces the central ideas and purpose of this writing sample.	5

(continued)

Component	Subcomponent	Poor or missing (0%)	Adequate (50%)	Good (75%)	Excellent (100%)	Points
Writing sample	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 ideas.	A clear conclusion summarizes the central ideas.	5
Reflection	Reflection	No reflection OR prompts not addressed appropriately/answers off topic OR inappropriate tone for reflection in a professional context.	Majority of prompts addressed appropriately, but may be overly general AND/OR lack evidence of thoughtful reflection AND/OR the tone may be somewhat inappropriate for a professional context.	All prompts are addressed appropriately. Reasonably clear and mostly specific, though perhaps with some overly general comments that are not more deeply developed. Evidence of some good reflective though. Tone is appropriately professional for academic/work context.	All prompts are answered clearly and specifically. Strong evidence of reflective thought. Tone is appropriately professional for academic/work context.	5
General (multiple components)	Writing mechanics	Several missing written components 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.	5

6

Portfolio

Information	Note
Name	Portfolio
Type (Main, Mini or Basket)	Main
Value	20% (Path A) 25% (Path B)
Due	Thursday, February 17, 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.

Portfolio assessments aim to help you demonstrate your technical coding, statistical thinking, communication and reflection skills. This portfolio also aims to recap and refresh knowledge from your previous statistics courses as well as building your ability to create quality data visualizations.

6.1 General instructions

- Be very careful to follow instructions on variable naming. If you do not, your code won't pass auto-grading and you will not receive the grades. This will not be eligible for regrading requests.
- Comment your code! In an R code chunk comments start with a # (pound sign or hashtag symbol). Don't confuse this with the use of # to denote different levels of headings in the text parts (governed by Markdown syntax) of an R Markdown document.
- **All figures and tables should have appropriate (and appropriately positioned) captions.**
- You should neatly format your code. No strict style is required, but make sure it is easy to read your work. (Optional) You may find Google's R Style Guide or the Tidyverse Style Guide interesting.
- Include your code in the body of the PDF itself (don't set echo=FALSE, don't hide/suppress etc.). Note that this is different to what you will be asked to do in the final project or in professional reporting. This is a demonstration of your skills.
- If asked to write and answer in words, write in full sentences.
- KNIT EARLY AND OFTEN! Don't leave things till the last minute, your Rmd not knitting is not an emergency for which an extension will be granted.

6.1.1 Template

You can access the template for this assessment here.

There is currently a lot of 'filler text' and 'filler code' in the template that you will want to **delete.** Fun fact: All filler text sourced from Hipster Ipsum, which Katy Wang in the UG Stats office introduced me to.

6.2 Submission instructions

- Submit both your Rmd (must be called: sta303-w22-portfolio.Rmd) and PDF (must be called: sta303-w22-portfolio.pdf) on MarkUs.
- You do not need to submit any data or tex files.
- The PDF must be knit directly from the Rmd (not Word or HTML first).

6.3 Cover page

You don't have to use the provided template, but you DO need to write your portfolio in RMarkdown and include a standalone cover page. The cover page must have:

- A title and subtitle (you can use my examples in the template or update them as you see fit, no points for changing, but it might make it a better part of your own website or GitHub portfolio, in the future)
- Your name
- Date (assessment submission date is fine)

It should not have your table of contents or any analyses on it.

In the template, you can change the colour of this cover to any colour you would like by replacing 6C3082 in the YAML (line 13) to another hex code. You could use this tool to help you: https://htmlcolorcodes.com/color-picker/

6.4 Introduction

Write this section last or second to last (before the refection).

In the introduction section, write a brief summary of the skills you have demonstrated in this portfolio, **across the statistical skills sample, writing sample and reflection sections**. Think of it like a **cover letter** for this document. It should be appropriate for a fairly general audience—imagine a future employer reading this. You may want to briefly explain the course context, as you understand it. What is STA303/1002 about? (Consider the learning objectives in the syllabus)

Your introduction should be **no longer than 300 words** and must **fit on one page**. If you did the mini-portfolio, you may *reuse* your own phrasing from that introduction (e.g., in introducing the course context), but this is also an opportunity to *improve* upon it.

6.5 Statistical skills sample

6.5.1 A note on time management.

Tasks 3 and 5 are essentially revision tasks for previous courses as I saw that there were still quite a few misunderstandings about these core concepts. If you find yourself unable to completely finish this portfolio, focusing your attention on the new material in Task 2 is a good strategy.

I have tried to make the revision a useful practice opportunity for other skills. You'll see that Task 3 asks you to build a function. You may not have seen this in other classes before, but use the templates provided to help you. I expect your focus to be on modifying these templates to display appropriate text, I don't expect you to be an expert on error handling in R functions.

6.5.2 Task 1: Setting up libraries and seed value

- Set up a chunk called `setup_portfolio` where you load the `tidyverse` library.
- Set your chunk options to `message=FALSE` so all the package loading information isn't included in your our output. You will need to make sure you run this chunk each time you start a new session so you can use many of the functions required.
- Create an object called `last3digplus` that takes the value 100 + the last three digits of your student ID number.
- Delete any placeholder library chunks in the template (if using).

6.5.3 Task 2a: Return to Statdew Valley: exploring sources of variance in a balanced experimental design (teaching and learning world)

In a class interactive, we explored some tomato and honey data in a statistical farming simulator.

Suppose you also have some data about strawberry yields (in kgs) across 18 different strawberry patches on your farm. There had been a problem with birds eating more than their fair share of your grandmother's strawberries (there wouldn't be enough for her famous jam!) so she had tried out three 'treatment' types: a scarecrow, covering the strawberries with netting and doing nothing (as a control).

Each patch was harvested 6 times, twice under each treatment. Your grandmother says weather and soil conditions were remarkably consistent across the entire time.

6.5.3.1 Growing your (grandmother's) strawberry patch

```
# Don't edit this file
# Sourcing it makes a function available
# You CAN edit these comments
# Load grow_my_strawberries() function
source("grow_my_strawberries.R")
```

- Run the function `grow_my_strawberries()` (the chunk to source the function must be run first for this to work) with `seed = last3digplus` as the input, saving the output as the object `my_patch`.
- Alter the `my_patch` data so that `treatment` is a factor variable with the levels ordered as follows: "No netting," "Netting," "Scarecrow."

6.5.3.2 Plotting the strawberry patch

Create a plot, using `ggplot2`, where the x-xis has each plot (alphabetical) and the y-axis represents yield.

- Just for fun, make each point an upside-down triangle (strawberry-shaped).
  - Within `geom_point()`, set `pch = 25`.
- Colour the points with the following colours:

  - No netting (#78BC61), Netting (#E03400), Scarecrow (#520048).

  - You will need to point both `fill` and `color` to the appropriate variable in your aesthetic mapping and use `scale_fill_manual` and `scale_color_manual` to set the colours.

- Apply `theme_minimal()` to the plot
- Add a caption "Created by YOUR NAME in STA303/1002, Winter 2022," with your name included
- Add an appropriate figure caption

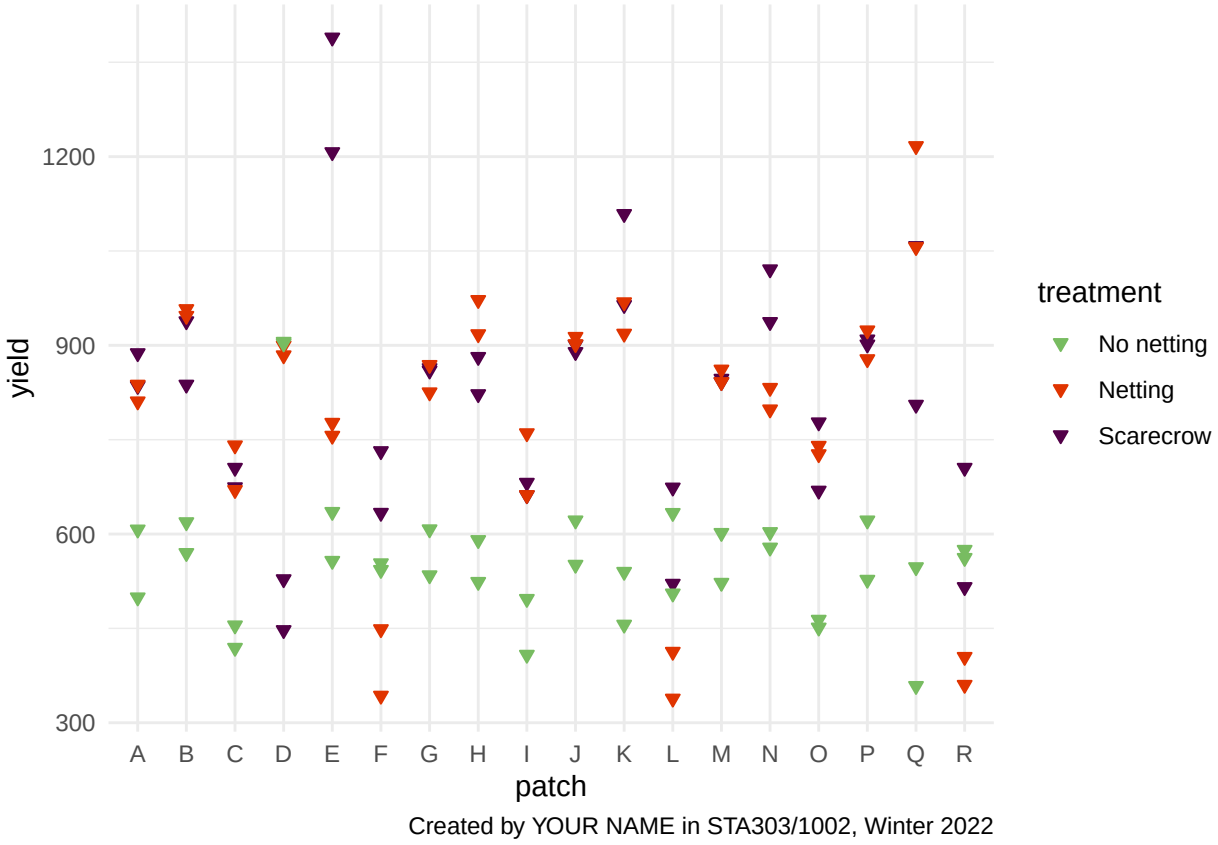


Figure 6.1: Target figure for task 2a (your patch data will be different, though!)

6.5.3.3 Demonstrating calculation of sources of variance in a least-squares modelling context

6.5.3.3.1 Model formula Write out the model formulation for the model that fits a fixed effect for treatment, a random effect for patch, and a random effect for the interaction of treatment and patch.

- Make sure you also EXPLAIN each term. All random effects should have a statement of their distribution.
- While there are multiple ways to write out this model, I strongly suggest you mimic the structure written out for the honey model in the Statdew Valley interactive.
- Create a table with the following columns: `Source of variation`, the name for this source of variation `Variance`, the values you've calculated (`var_patch`, `var_int`, `var_ab`), `Proportion`, which is the proportion of variance in yield (not explained by treatment) explained by this source (round to 2 decimal places). Use `kable` to display the table nicely and add an appropriate caption.

6.5.3.3.2 Objects to create (proof-read carefully for autograding)

Table 6.2: An example table, where you can replace the appropriate values, including removing the quote marks around the variance values and calculating the proportions values

Source of variation	Variance	Proportion
Source 1	var_b	0.33
Source 2	var_ab	0.33
Source 4	var_int	0.33

6.5.3.3.2.1 Tibbles

- `my_patch`, your 'patch' of strawberries, grown based on using your `last3digplus` as the seed (created in the first part of the task).
- `agg_patch`, starting with `my_patch`, group by `patch` and then summarize to create a new variable, `yield_avg_patch`.
- `agg_int`, starting with `my_patch`, group by both `patch` and `treatment` and then summarize to create a new variable, `yield_avg_int` that is the average strawberry yield for each `patch` and `treatment` combination.

6.5.3.3.2.2 Models

- `int_mod`, an interaction model including main effects. `yield` is the response and `patch` and `treatment` should be the predictors.
- `patch_mod`, an intercept only model, where `yield_avg_patch` is the response. Use appropriate data aggregated to find the average yield per patch.
- `agg_mod`, a main effects model where `yield_avg_int` is the response and `patch` and `treatment` are the predictors. Use appropriate data aggregated across both `patch` and `treatment`.

6.5.3.3.2.3 Numeric values Note: These should not be hardcoded (i.e., not `var_patch = 3`), they should make use of your models.

- `var_patch`, variance in average yield explained by patch-to-patch variance, after accounting for fixed effects and other sources.
- `var_int`, residual variance after fitting the version of this linear model with the most parameters
- `var_ab`, variance in yield explained by the interaction between patch and treatment, after accounting for the fixed effects and other sources.

```
# Example tibble
tibble(`Source of variation` = c("Source 1",
                                 "Source 2",
                                 "Source 4"),
       Variance = c("var_b", "var_ab", "var_int"),
       Proportion = c(round(0.3333333, 2),
                      round(0.3333333, 2),
                      round(0.3333333,2) )) %>%
  # It is okay if that caption goes off the page when knitting
  knitr::kable(caption = "An example table, where you can replace the appropriate values, including removi
```

6.5.4 Task 2b: Applying linear mixed models for the strawberry data (practical world)

- Fit 3 models (yield will be the response in all of them)
  - `mod0`, a linear model with only `treatment`,
  - `mod1`, a linear mixed model with `treatment` and `patch` (appropriate choices about what is a fixed vs random effect should be made)
  - `mod2`, a linear mixed model with `treatment`, `patch` and the interaction of `treatment` and `patch` (appropriate choices about what is a fixed vs random effect should be made)
- Set up appropriate comparisons between these models, using a likelihood ratio test (`lmtest::lrtest`)
  - **Writing**: make explicit comment on whether/where you are using REML / ML and WHY)

6.5.4.1 Justification and interpreation

- Chose which model, out of `mod0`, `mod1` and `mod2` is the most appropriate final model (make it clear which one you are choosing and justify why).
  - Interpret the fixed effect coefficients.
  - Discuss the proportion of variance, not explained by the fixed effects, that is explained by each source of variance in the model.
  - Hint: If you have done both Task 2a correctly, it will match your variances from one of these model.

This should be at MAX 1-2 paragraphs.

6.5.5 Task 3a: Building a confidence interval interpreter

Write a function that helps you interpret a frequentist confidence interval with the appropriate language. Remember that we need to be careful about what our confidence level means and what value the interval is for.

Note: For this and the next task, you are largely replacing and reworking what I have set up. I am definitely not claiming these are excellently specified functions, there are lots of other improvement we might try to make if we were really adding a useful function to a package. For now, I just want you to get a bit of a sense of where the errors and warnings you see when using other functions come from. **The main focus should be on practising a sensible template for how to interpret these values.** Honestly, this is not a place where you need to be *creative* in your writing. It will speed up your writing and communication if you have a mental script for each time you have to interpret a p value or confidence interval.

6.5.5.1 Specifications:

- The function should be called `interpret_ci`.
- It should take 4 arguments:
  - `lower`, the lower bound of the confidence interval (numeric)
  - `upper`, the upper bound of the confidence interval (numeric)
  - `ci_level`, the confidence level this interval was calculated at, e.g. `99` or `95` (numeric)
  - `stat` a description of the statistic of interest
- The output must be a full sentence that uses all 4 of the inputs to the function. It must be CLEAR what the confidence interval is making a claim about and the interpretation must be statistically appropriate.
- Warnings must be produced in the following cases:
  - If `lower` or `upper` is not numeric.

  - If `ci_level` is not an numeric value between 0 and 100.
  - If `stat` isn't a character string.

- You must run (and include) the following code after setting up your function:

  - `interpret_ci(10, 20, 99, "mean number of shoes owned by students")`
  - `interpret_ci(10, 20, -1, "mean number of shoes owned by students")`
  - `interpret_ci(10, 20, -1, tibble(stat = 3))`
  - The results of the above must be printed out in the main text of the PDf (see examples in template)

```
interpret_ci <- function(lower, upper, ci_level, stat){
  if(!is.character(stat)) {
    # produce a warning if the statement of the parameter isn't a character string
    # the spacing is a little weird looking so that it prints nicely in your pdf
    warning("
    Warning:
    stat should be a character string that describes the statistics of
    interest. The first part of this would be appropriate error text for this
    issue. You can reuse it as in in your code (not plagairisim concern),
    just make sure you delete this part.")
  } else if(!is.numeric(lower)) {
    # produce a warning if lower isn't numeric
    warning("
            Warning: NOPE! This error message should be improved.")
  } else if(!is.numeric(upper)) {
    # produce a warning if upper isn't numeric
    warning("
            Warning: NOPE! This error message should be improved.")
  } else if(!is.numeric(ci_level) | ci_level < 0 | ci_level > 100) {
    # produce a warning if ci_level isn't appropriate
    warning("
    Warning: Your ci_level is wroooooong.
    This error message should be improved.")
  } else{
    # print interpretation
    # this is the main skill I want to see, writing a good CI interpretation.
  str_c("This is the confidence level: ", ci_level,
        "%. And the text fed to stat is ", stat,
        ". There is also the lower and upper bounds: ", lower, " and ", upper,
        ". This part needs to be fixed!" )
  }
}

# Test 1
ci_test1 <- interpret_ci(10, 20, 99, "mean number of shoes owned by students")

# Test 2
ci_test2 <- interpret_ci(10, 20, -1, "mean number of shoes owned by students")

# Test 3
ci_test3 <- interpret_ci(10, 20, -1, tibble(stat = 3))
```

CI function test 1: This is the confidence level: 99%. And the text fed to stat is mean number of shoes owned by students. There is also the lower and upper bounds: 10 and 20. This part needs to be fixed!

CI function test 2: Warning: Your ci_level is wroooooong. This error message should be improved.

CI function test 3: Warning: stat should be a character string that describes the statistics of interest. The first part of this would be appropriate error text for this issue. You can reuse it as in in your code (not plagairisim concern), just make sure you delete this part.

6.5.6 Task 3b: Building a p value interpreter

Write a function that will interpret p values based on strength of evidence. Use this table to help you. You can decide how to set up the 'greater than' vs 'strictly greater than's, etc. just be consistent (and make sure you understand why I'm a little fuzzy about the borders anyways.)

6.5.6.1 Here is an EXAMPLE. This is not the answer. I DO NOT want to see this submitted as an answer.

This would be an example of a p values interpreter that uses a 0.05 threshold to reject or fail to reject the null hypothesis.

```
# This is just an example! MODIFY THIS CODE
interpret_pval <- function(pval, nullhyp){
  if(!is.character(nullhyp)) {
    warning("
            Warning: Excuse me, why isn't nullyhyp a character string?
            This is not an appropriate error message, FYI.
            Make yours more professional.")
  } else if(!is.numeric(pval)) {
    warning("You p value should be a number.")
  }  else if(pval > 1) {
    warning("
            Warning: What do you mean, a p value greater than 1?!?!
            No way.
            This is also not an appropriately professional error message.")
  } else if(pval < 0){
    warning("
            Warning: Negative? I don't think so!
            This is also not an appropriately professional error message.")
  } else if(pval > 0.05){
    str_c("The p value is ", round(pval, 5),
                ", we fail to reject the hyothesis that ", nullhyp)
  } else if(pval <= 0.05){
    str_c("The p value is ", round(pval, 5),
                ", we reject the hypothesis that ", nullhyp, ".")
  }
}

pval_test1 <- interpret_pval(0.0000000003,
                             "the mean grade for statistics students is the same as for non-stats students

pval_test2 <- interpret_pval(0.0499999,
                             "the mean grade for statistics students is the same as for non-stats students

pval_test3 <- interpret_pval(0.050001,
                             "the mean grade for statistics students is the same as for non-stats students

pval_test4 <- interpret_pval("0.05", 7)
```

p value function test 1: The p value is 0, we reject the hypothesis that the mean grade for statistics students is the same as for non-stats students.

p value function test 2: The p value is 0.05, we reject the hypothesis that the mean grade for statistics students is the same as for non-stats students.

p value function test 3: The p value is 0.05, we fail to reject the hyothesis that the mean grade for statistics students is the same as for non-stats students

p value function test 4: Warning: Excuse me, why isn't nullyhyp a character string? This is not an appropriate error message, FYI. Make yours more professional.

- Specifications for your function:
  - Your function must be called `interpret_pval`.
  - It must take two arguments: `pval` and `nullhyp`.
  - There must be an warning produced, with an *appropriately informative* (and **professional**!) error message, for values that p values cannot take AND non-character string null hypotheses. Use the template above and improve the text as well as adding new `else if`s to the chain.
  - The interpretation messages must be statistically appropriate and should *not* talk about rejecting or failing to reject the null. Instead the strength of evidence should be discussed appropriately.
  - The output should be:
    * a full sentence,
    * that included the p value, rounded to 3 decimal places or replaced with "<.001" if less than 0.001 (based on APA guidelines for reporting statistics), and
    * incorporate the text of the null hypothesis.

6.5.7 Task 3c: User instructions and disclaimer

Neither of these sections should be very long (1 to 3 paragraphs maximum). Please write in full sentences.

6.5.7.1 Instructions

Write **brief** instructions for how to use your two interpreters to someone who is new to statistics, but has been told they need to apply some basic statistical methods in their summer research project.

This should touch on what a **null hypothesis** is and what a **population parameter** is, and should give some examples and tips on wording a null hypothesis appropriately. It should also mention some common pitfalls in interpreting frequentist confidence intervals.

6.5.7.2 Disclaimer

Write a **disclaimer/warning** for your p value interpreter, specifically. What should people know and be cautious of before using it? Note: This should not be "be careful, I might have messed up my code," this should be based on what guidance someone should be aware of if they were going to use this interpreter to interpret the results of a independent samples t-test, for example.

6.5.8 Task 4: Creating a reproducible example (reprex)

In Module 1, we discussed creating reproducible examples in R.

Suppose you had written the below code. You want to summarize the dataset `my_data` into a new dataset called `my_summary` and find the mean value for each group, and just have one row per group, but for some reason it is not working.

Note: DO NOT *FIX* THIS CODE. That is not the *point* of this activity.

Goal: Create a reproducible example that would allow someone ELSE to understand the code and situation.

In your submission there should be two things:

1. A brief **description** IN YOUR OWN WORDS of what a reprex is and what you needed to consider to produce the reprex.
2. The reprex **output** that would allow someone to completely reproduce your problem exactly. Think carefully about what you would need to provide. I.e., what you might post on a help-board as part of a request for help. You do NOT need to write up the rest of the description.

Hint: The workflow will be something like (1) create the code chunk for the reprex, (2) copy the code chunk to clipboard (Ctrl+C or Cmd+C), (3) run `reprex::reprex()` in the console, (4) paste output into your template (it *won't* be a runable code chunk, you should be able to just paste as is and not edit it at all). Delete the original chunk (i.e., there should be now code run in this section of your report when you knit it.)

```
my_data <- tibble(group = rep(1:10, each=10),
                  value = c(16, 18, 19, 15, 15, 23, 16, 8, 18, 18, 16, 17, 17,
                            16, 37, 23, 22, 13, 8, 35, 20, 19, 21, 18, 18, 18,
                            17, 14, 18, 22, 15, 27, 20, 15, 12, 18, 15, 24, 18,
                            21, 28, 22, 15, 18, 21, 18, 24, 21, 12, 20, 15, 21,
                            33, 15, 15, 22, 23, 27, 20, 23, 14, 20, 21, 19, 20,
                            18, 16, 8, 7, 23, 24, 30, 19, 21, 25, 15, 22, 12,
                            18, 18, 24, 23, 32, 22, 11, 24, 11, 23, 22, 26, 5,
                            16, 23, 26, 20, 25, 34, 27, 22, 28))

my_summary <- my_data %>%
  summarize(group_by = group, mean_val = mean(value))

glimpse(my_summary)
```

```
## Rows: 100
## Columns: 2
## $ group_by <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3~
## $ mean_val <dbl> 19.67, 19.67, 19.67, 19.67, 19.67, 19.67, 19.67, 19.67, 19.67~
```

6.5.9 Task 5: Simulating p values

6.5.9.1 Setting up simulated data

This revision task is meant to connect to Question 16 of your pre-knowledge quiz and help you visualize the behaviour of_p_ values.

- Set your seed to `last3digplus` (that you set up in Task 1)
- Create 3 simulated data sets (`sim1` to `sim3`) of length 100000 and with two columns: `group` and `val`.
  - Use `tibble` to create each dataset.
  - For each dataset, set `group = rep(1:1000, each = 100)`
  - For `sim1` set val to be 100000 values drawn from $N(0, 1)$ (see function `rnorm()`‘).
  - For `sim2` set val to be 100000 values drawn from $N(0.2, 1)$.
  - For `sim3` set val to be 100000 values drawn from $N(1, 1)$.
- Stack all 4 datasets into one new dataset called `all_sim`, with an new column at the beginning called `sim` that takes the value 1 if the observation is from `sim1`, 2 from `sim2`, and so on.
  - This is most of the code you will need `bind_rows(sim1, sim2, sim3, .id = "sim")`
- Make any alterations necessary to `all_sim` so that you can join on the dataset `sim_description` to provided better labels for each simulation (but without including any irrelevant labels).
  - You will need to include the code below to create this dataset.
- Plot histograms for the **first three groups** for each simulated dataset in one plot. Your final plot should use `ggplot` and `facet_wrap`. There should be three rows and three columns. The facets should be labelled with `desc` (the more informative description than just 1, 2, 3).
  - Provide an appropriate figure caption, set with your chunk options.
  - Add a within-image caption: “Created by YOUR NAME in STA303/1002, Winter 2022,” with your name included
  - Apply `theme_minimal()`.

```
# Create sim_description
# Dataset to merge with improved simulation names
sim_description <- tibble(sim = 1:4,
                          desc = c("N(0, 1)",
                                   "N(0.2, 1)",
                                   "N(1, 1)",
                                   "Pois(5)"))
```

```
all_sim %>%
  filter(group <= 3) %>%
  ggplot(aes(x = val)) +
  geom_histogram(bins = 40) +
  facet_wrap(desc~group, nrow = 3) +
  theme_minimal() +
  labs(caption = "Created by YOUR NAME in STA303/1002, Winter 2022")
```

6.5.9.2 Calculating p values

- Create a new dataset called `pvals` that starts with `all_sim`, groups by both `desc` AND `group`, and then summarizes to find the p value for each group, based on a **one sample, 2-sided t.test, where the null hypothesis is the the population mean is 0.**

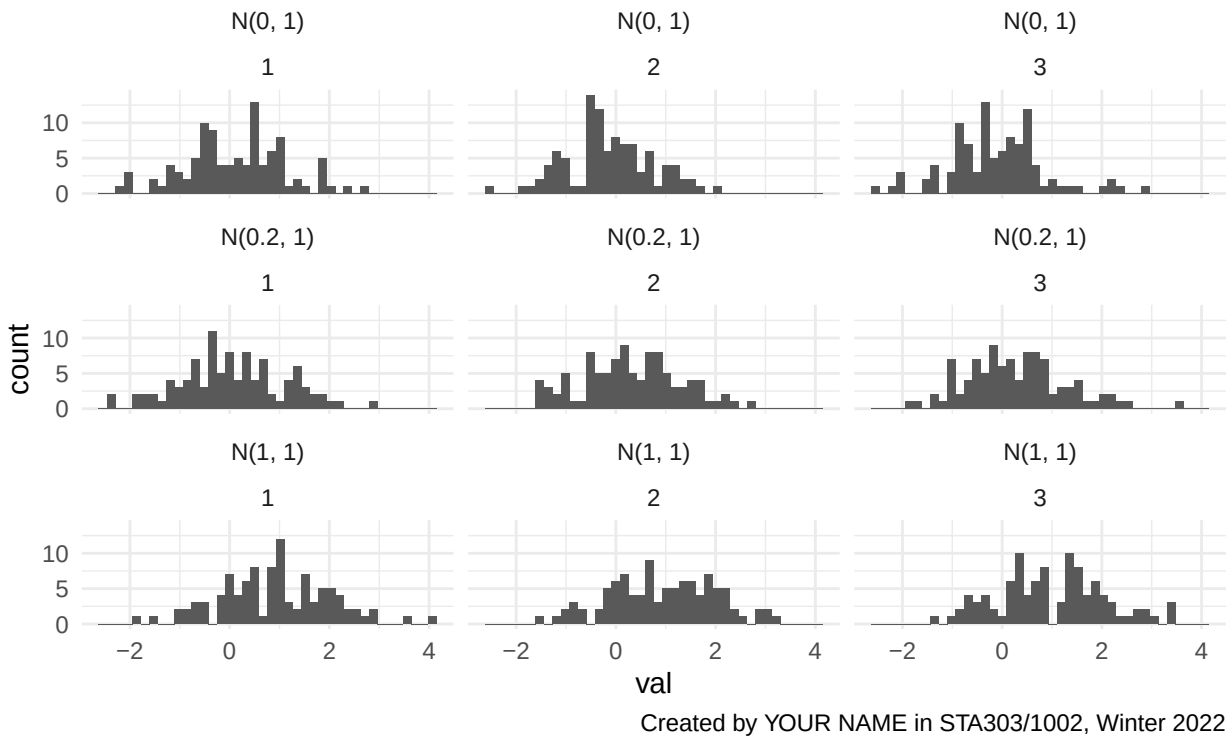


Figure 6.2: Target first visualisation for task 5

- `t.test(val, mu = 0)$p.value` will pull out the p value as a value for you (when included within the appropriate code).

- `.groups = "drop"` may be useful to add to your summarize statement if you are getting a message about groups.

- Plot histograms of the p values, faceted by `desc`. I would recommend adding `scales = "free_y"` to your `facet_wrap()` call as well (not assessed: can you figure out what it does?)

  - Use these statements appropriately in in your code to make the figure:
    * `geom_histogram(boundary = 0, binwidth = 0.05, fill = "grey", color = "black")`
    * `xlim(0,1)`
  - Provide an appropriate figure caption, set with your chunk options.
  - Add a within-image caption: "Created by YOUR NAME in STA303/1002, Winter 2022," with your name included

- Apply `theme_minimal()`.

6.5.9.3 Q-Q plots

- Create one final plot that creates a 3x1 figure with Q-Q plots for each simulation. You must determine what the distribution of a p value should be, WHEN THE NULL HYPOTHESIS IS TRUE, and compare it it to the distributions of p values we have generated.

I have provided a template for the code, you just need to decide which quantile function to put where the XXXXXXX is now. (e.g., `qpois` is the quantile function for the Poisson distribution). Also adjust the caption with your authorship and add an appropriate figure caption.

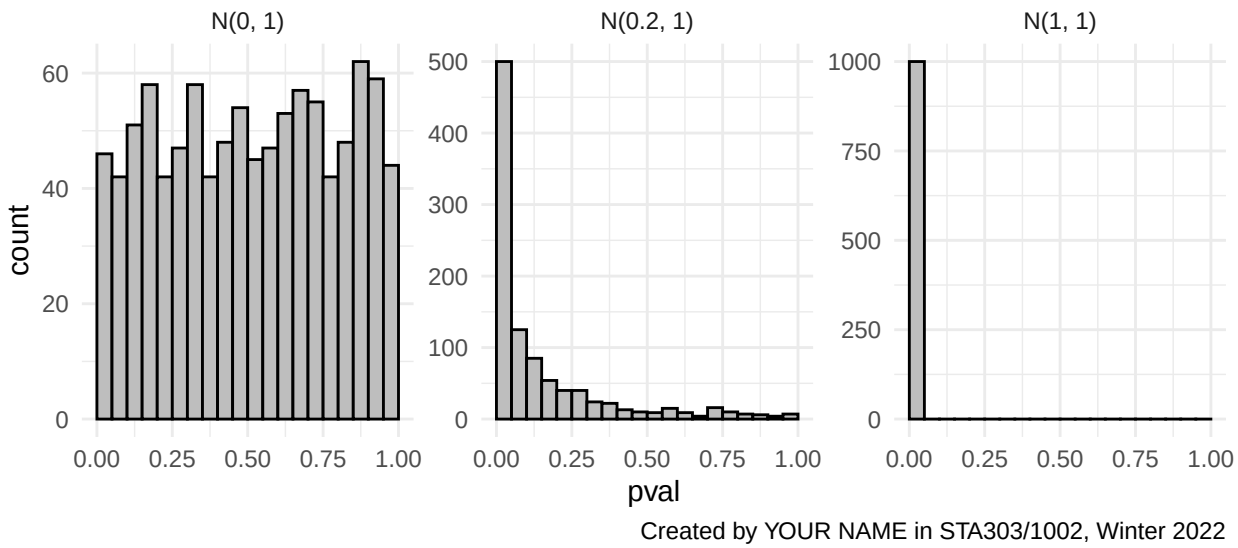


Figure 6.3: Target second visualisation for task 5

```
pvals %>%
  ggplot(aes(sample = pval)) +
  geom_qq(distribution = XXXXXXX) +
  geom_abline(intercept = 0, slope = 1) +
  facet_wrap(~desc) +
  theme_minimal() +
  labs(caption = "Created by YOUR NAME in STA303/1002, Winter 2022")
```

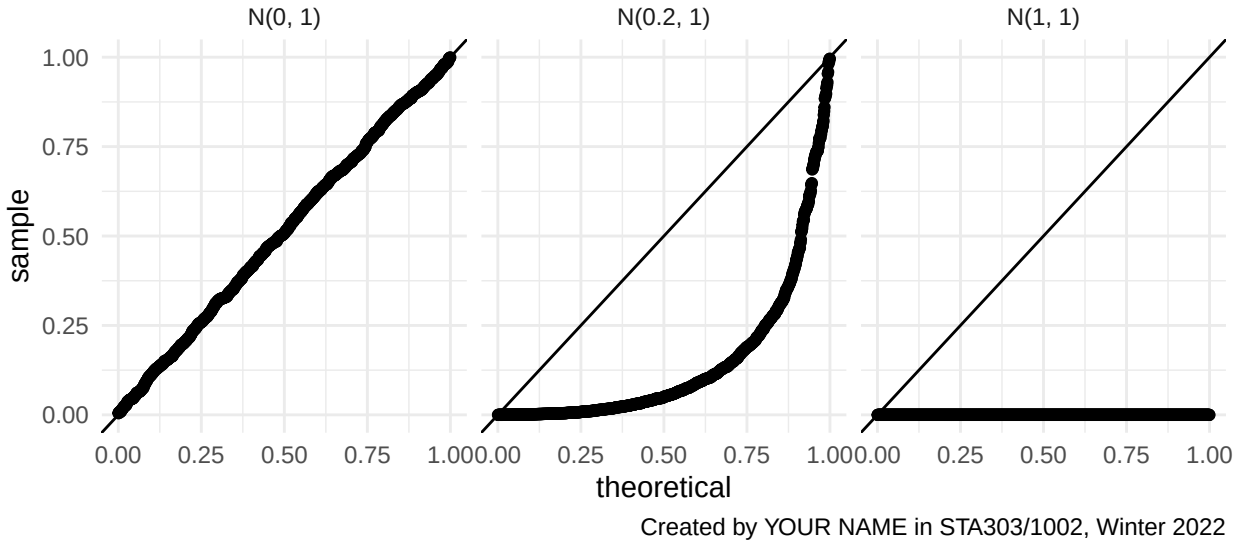


Figure 6.4: Target final visualisation for task 5

6.5.9.4 Conclusion and summary

Remark briefly on:

- how what you have done in this task connects to the **definition** of a p value, and
- specifically how it connects to answering the below question from the pre-knowledge check.

(Note: so that you didn't have to wait to long for your PDF to knit, I made the number of simulations and simulation size a little smaller, but the idea is the same.)

Question 16 1 pts

Suppose you were to **simulate** 10,000 sets of 10,000 **normally distributed data points**, $\mathcal{N}(0, 1)$, and perform a one-sample t-test on each dataset, where $H_0 : \mu = 0$.

Which **ONE** of the following statements is **TRUE**?

- ◯ Plotting a histogram of the p-values will show that the p-values will also have a normal distribution.
- ◯ Most of the p-values will be close to 1.
- ◯ Most of the p-values will be close to 0.
- ◯ Approximately 10% of the p-values will be between 0.9 and 1.
- ◯ It is not possible to make any claims about the distribution of these p-values.

6.6 Writing sample

6.6.1 Prompt

Read Motulsky (2014), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4203998/pdf/210_2014_Article_1037.pdf, and **paraphrase** it with the goal of giving advice to your future self. I.e., YOU in the future is your audience, but keep the tone professional enough that you could share it as a LinkedIn post, or provide a copy to an acquaintance as advice. This *shouldn't* be a summary of the full article. Instead, focus on what surprised you, what you didn't know before, what you don't want to forget. You can also disagree with the claims in the article if you see fit.

As your task is to explicitly comment on this article, you must **reference** the article! The in-text APA style reference for this article is: (Motulsky, 2014) with the below the full reference to include at the end:

> Motulsky, H. J. (2014). Common misconceptions about data analysis and statistics. *Naunyn-Schmiedeberg's Archives of Pharmacology*, *387*(11), 1017–1023. https://doi.org/10.1007/s00210-014-1037-6

You might also incorporate it as *"Motulsky (2014) states that…"*. You don't need the reference at the end of every sentence, but **it should be clear to a reader what ideas have come from the article and what has come from you, throughout the piece.**

You can use another referencing style if you wish, or you can just copy this (that is not an academic offence!).

Please be very mindful of proper paraphrasing! The main point of this activity is to put things in YOUR OWN WORDS. Read the provided advice in Module 2.

There should not be too many direct **quotes** from the article, and better if there are none at all. **IMPORTANT NOTE: A zero grade on paraphrasing will result in a zero on the ENTIRE writing component of your portfolio.**

Make sure your writing sample has an **introduction** and comes to a **conclusion.**

Word count: 300–500 words (not including reference at the end). Please add a statement of your word count at the end of the passage.

6.7 Reflection

Briefly, 100 to 200 words each, answer the following questions:

- What is something specific I'm proud of in this portfolio?
- How might I apply what I've learned and demonstrated in this portfolio in future work and study, after STA303/1002?
- What is something I'd do differently next time?

6.8 Rubric

There is a breakdown of the overall contribution of each component to your grade below, followed by a more detailed rubric.

Component	Percentage for component
Cover page	1%
Introduction	5%
Statistical skills sample	60%
Commenting of code	1%
Readability of R code	1%
Reproducibility	2%
Writing sample	20%
Reflection	5%
General (multiple components)	5%

Detailed rubric TBA.

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 to 8:00 p.m. ET
Submission instructions	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

7.0.1.1 How your grade is calculated

- The 98% of your mini-mixed grade is based on your performance on the **timed** component.
- 2% of your grade is based on the correctness of your Rmd. There will be a student facing autotest you can run on your submissions to check if the objects required are there and appear to be mostly sensible (note, this doesn't guarantee in all cases that they are fully RIGHT, just that they passes the checks set up for this component).
- If you do the timed component, but DON'T submit the appropriate Rmd and PDF to MarkUs by the end of the window, there is an **additional** 10 percentage point penalty.
  - Note the file name requirements: `sta303-w22-mini-mixed.Rmd` and `sta303-w22-mini-mixed.pdf`.
  - You can upload as many times as you like BEFORE the end of the window on Wednesday, so make sure you upload a 'safety' copy of your Rmd and PDF once you have started working on it.

7.0.1.2 Getting help

Piazza will be available until the start of the window. You may ask questions about the untimed component of the mini-mixed assessment, using the mini-mixed assessment tag.

Tech office hours will take the place of our usual class time on Wednesday. Use the usual link to join. Content questions will not be answered, but I recommend you aim to complete the timed component before/during these times as then I can provide timely assistance if there are any issues.

Review materials:

- Check out the LMM study guides your classmates have been making.
- Asynchronous review versions of Team Up! activities
  - Module 1 (Intro, linear regression as common tests)
  - Module 2 (Data wrangling, ethics, communication)
  - Module 3 (Linear Mixed Models)

7.0.1.3 Academic integrity

Please respect all of the hard work you've done so far this year as you complete the following assessment of your learning by making sure that the work you do here is your own. Don't put your academic career at risk.

There will be a question at the beginning of the timed component that you should read carefully.

7.0.1.4 Untimed component

JupyterHub pull link (look for `sta303-w22-mini-mixed`)

7.0.1.5 Timed component

The timed component will become available at 8:00 a.m. ET and remain available until 8:00 p.m. ET. You will have 50 minutes to complete it.

Before attempting the timed component you should work through the guided analysis in the untimed component. There will be questions related to these tasks.

You must also submit your .Rmd and.pdf of the guided analysis before the end of the assessment window (8:00 p.m. ET on Wednesday, March 9).

Must read before starting assessment:

- I suggest you use Chrome or Firefox to access this quiz. I use Chrome to write and check the quizzes. Additionally, make sure your browser is up to date as out-of-date versions also cause issues.
  - Possible 'hack' if images don't load: Sometimes, right-clicking the image that is failing to load, and opening it in a new tab or downloading it, will let you see the image.
- This timed assessment is open book, so you are welcome to review the notes while completing it.
- From the time you start, you will have either 50 minutes to complete the quiz or until 8:00 p.m. ET on Wednesday, March 9, whichever comes first.
- You cannot pause the quiz once you start it.
- Ensure you will have uninterrupted access to the internet and a quiet space to work for the duration of the test.
- A desktop computer or laptop is preferable to a tablet or phone.
- Make sure your device is plugged in or will definitely have sufficient charge for the period of the test.
- The quiz will auto-submit any answers at the deadline.

- You have one attempt.
- You do not have to run code while doing this quiz, but you are allowed to if you would like to.
- BE PREPARED. Have back-ups planned as we will not be accepting technical issues as a reason for not submitting.
- Make sure you submit the .pdf and .Rmd of your untimed component before the end of the timed window, also.
- Academic integrity: When you take this quiz, you will be asked to confirm that your behaviour during the assessment adheres to the Code of Behaviour on Academic Matters. READ THIS QUESTION CAREFULLY. You would be surprised how often students don't do this. An incorrect answer may result in a 0 on the mini-mixed. Show me you are really reading and paying attention.

8

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 to 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.

8.0.1 Instructions

The mixed assessment has THREE components:

- Untimed guided analysis (this), which must be submitted (both Rmd AND PDF) on MarkUs.
- There is only ONE untimed component, it is relevant to both timed parts.
- Timed assessment (PART A) (50 minutes; 12-hour assessment window is 8:00 a.m. to 8:00 p.m. ET Wednesday, March 16).
- Timed assessment (PART B) (50 minutes; 12-hour assessment window is 8:00 a.m. to 8:00 p.m. ET Wednesday, March 16).

The timed components are split across two timed quizzes so you can take a break in the middle if you wish. I.e., you could do them one after the other, or plan to do one, take an hour break and do the next one. Whatever suits you is good, as long as you submit everything before 8:00 p.m. ET.

8.1 How your grade is calculated

- The 98% of your mixed grade is based on your performance on the **timed** component.

- 2% of your grade is based on the correctness of your Rmd. There will be a student facing autotest you can run on your submissions to check if the objects required are there and appear to be mostly sensible (note, this doesn't guarantee in all cases that they are fully RIGHT, just that they passes the checks set up for this component).

- If you do the timed components, but DON'T submit the appropriate Rmd and PDF to MarkUs by the end of the window, there is an **additional** 10 percentage point penalty.
  - Note the file name requirements: `sta303-w22-mixed.Rmd` and `sta303-w22-mixed.pdf`.
  - You can upload as many times as you like before the end of the window, so make sure you upload a 'safety' copy of your Rmd and PDF once you have started working on it.

8.2 Getting help

Piazza will be available until the start of the window. You may ask questions about the untimed component of the mini-mixed assessment, using the mini-mixed assessment tag.

Tech office hours will take the place of our usual class time on Wednesday. Use the usual link to join. Content questions will not be answered, but I recommend you aim to complete the timed component before/during these times as then I can provide timely assistance if there are any issues.

Review materials:

- Check out the LMM study guides your classmates have been making.
- GLM study guides
- Asynchronous review versions of Team Up! activities
  - Module 1 (Intro, linear regression as common tests)
  - Module 2 (Data wrangling, ethics, communication)
  - Module 3 (Linear Mixed Models)
  - Module 4 (Generalized Linear Models)

8.2.1 Academic integrity

Please respect all of the hard work you've done so far this year as you complete the following assessment of your learning by making sure that the work you do here is your own. Don't put your academic career at risk.

There will be a question at the beginning of each timed component that you should read carefully.

8.3 Untimed component

JupyterHub pull link (look for `sta303-w22-mixed`)

8.4 Timed components

The timed components will become available at 8:00 a.m. ET and remain available until 8:00 p.m. ET. You will have 50 minutes to complete each of them (i.e., 100 minutes total).

Before attempting the timed component you should work through the guided analysis in the untimed component. There will be questions related to these tasks.

You must also submit your .Rmd and.pdf of the guided analysis before the end of the assessment window (8:00 p.m. ET on Wednesday, March 16).

Must read before starting either timed assessment:

- The order you do the two timed quizzes in shouldn't matter.

- I suggest you use Chrome or Firefox to access this quiz. I use Chrome to write and check the quizzes. Additionally, make sure your browser is up to date as out-of-date versions also cause issues.
  - Possible 'hack' if images don't load: Sometimes, right-clicking the image that is failing to load, and opening it in a new tab or downloading it, will let you see the image.
- This timed components is open book, so you are welcome to review the notes while completing them.
- From the time you start each quiz, you will have either 50 minutes to complete the quiz or until 8:00 p.m. ET on Wednesday, March 16, whichever comes first.
- You cannot pause a quiz once you start it.
- Ensure you will have uninterrupted access to the internet and a quiet space to work for the duration of the tests.
- A desktop computer or laptop is preferable to a tablet or phone.
- Make sure your device is plugged in or will definitely have sufficient charge for the period of the tests.
- The quizzes will auto-submit any answers at the deadline.
- You have one attempt per quiz.
- You do not have to run code while doing these quizzes, but you are allowed to if you would like to.
- BE PREPARED. Have back-ups planned as we will not be accepting technical issues as a reason for not submitting.
- Make sure you submit the .pdf and .Rmd of your untimed component before the end of the timed window, also.
- Academic integrity: When you take this quiz, you will be asked to confirm that your behaviour during the assessment adheres to the Code of Behaviour on Academic Matters. READ THIS QUESTION CAREFULLY. You would be surprised how often students don't do this. An incorrect answer may result in a 0 on the mini-mixed. Show me you are really reading and paying attention.

9

Final project

Information	Note
Name	Final project
Type (Main, Mini or Basket)	Main
Value	45%
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.

9.0.1 Instructions

This assessment has its own website: https://sta303-bolton.github.io/sta303-w22-final-project/

9.0.1.1 Project groups

You can work on your project alone or with up to **three** other students (regardless of section) for a group of **four** total students.

Groups will need to register a few weeks in advance of the deadline (date TBC).

There will be opportunities to meet others in the class through some of the activities planned, and closer to the deadline to register groups, I will make the Piazza "Search for Teammates!" area available.

10

Knowledge basket overview

You can max your knowledge basket out to the full 5% of your final grade.

Assessment	Possible percentage	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)
'Getting to know you' survey	0.1%	Jan 13
Module 1 graduate school info session and panel	0.1%	Jan 19 (session) | Jan 21 (survey)
Module 3 writing workshop attendance and survey	0.25%	
Hack your class workshop attendance and reflection	0.25%	
Sports analytics workshop attendance and reflection	0.25%	
5 Ways to Well-being attendance and reflection	0.5% |	
Academic resilience workshop attendance and reflection	0.5% |	
Professional development task proposal	1% (graded on quality)	Feb 3
Professional development task evidence & reflection	3% (graded on quality)	Mar 31
Writing & peer review (Create-Assess-Reflect) x 5 • Module 1 • Module 2 • Module 3 • Module 4 • Module 5	0.5% x 5	Friday-Tuesday-Friday each Module
Module check-in x 5 • Module 1 • Module 2 • Module 3 • Module 4 • Module 5	0.1% X 5	Last Friday of each Module

Assessment	Possible percentage	Due date
Team Up! Activities • Module 1 (get links on course overview page) • Module 2 (get links on course overview page) • Module 3 (get links on course overview page) • Module 4 (get links on course overview page) • No Team Up! for Module 5	0.5% (graded on quality) X ~5	First Thursdays of modules (not guaranteed for all)
Study guide: LMM	1% (graded on quality)	March 15
Study guide: GLM	1% (graded on quality)	March 15
Punctuation art	1% (graded on quality)	April 8
Breathing exercise	1% (graded on quality)	April 8

10.0.1 Planning and tips

There are LOTS of ways to fill up your knowledge basket and I really encourage you to spend some time thinking about both your own working style, what the rhythm of your semester is going to look like and, most importantly, **what your personal learning and growth goals are**. Past students have told me they hated doing the writing tasks, but then embarked on a summer internship and found the practice had been really helpful to them and allowed them to stand out as both data cruncher AND communicator.

DO NOT LEAVE THIS TILL THE END! There will not be make up tasks offered if you get to the end of the semester and realize you're missing a few percentage points.

The knowledge basket activities listed in the (syllabus)(#syllabus) are guaranteed, and there will also be some additional opportunities offered (for example the punctuation art and Team Up! options also included in the examples below).

10.0.2 Examples

All the student below would earn their full 5%, but in very different ways.

10.0.2.1 Student A

Task	Example grade/note	Contribution to basket
Module 1 Team Up!	8% score	0.4%
Pre-knowledge check	60% score	0.5%
Module 5 writing task	Completed all phases	0.5%
PD Proposal	70% score	0.7%
PD Evidence & reflection	70% score	2.1%
Punctuation art	100% score	1%

10.0.2.2 Student B

Task	Example grade/note	Contribution to basket
Pre-knowledge check	88% score	1%

Task	Example grade/note	Contribution to basket
PD Proposal	100% score	1%
PD Evidence & reflection	99% score	2.7%
Punctuation art	100%	1%

10.0.2.3 Student C

Task	Example grade/note	Contribution to basket
Pre-knowledge check	67% score	0.5%
Pre-knowledge check	Workshop	0.5%
Module 1 check-in	Complete	0.1%
Module 2 check-in	Complete	0.1%
Module 3 check-in	Complete	0.1%
Module 4 check-in	Complete	0.1%
Module 5 check-in	Complete	0.1%
Module 1 Team Up!	80% score	0.4%
Module 2 Team Up!	70% score	0.35%
Module 3 Team Up!	85% score	0.425%
Module 4 Team Up!	50% score	0.25%
Module 5 Team Up!	10% score	0.5%
Module 1 writing task	Completed all phases	0.5%
Module 2 writing task	Completed Create phase only	0.245%
Module 3 writing task	Completed all phases	0.5%
Module 5 writing task	Completed all phases	0.5%

11

Knowledge basket: Writing and peer feedback

11.1 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

11.1.1 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).

11.1.1.1 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

11.1.2 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.

11.1.3 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.

11.1.4 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.

11.2 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.

11.2.1 M1 Instructions

Make sure you are familiar with the general instructions for these types of tasks.

Component	Poor or Missing	Adequate	Good	Excellent
Addresses 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).

11.2.1.1 M1 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?

11.2.1.2 M1 Rubric

11.3 Module 2 writing task

Information	**Note**
Name	Module 2 writing task
Type (Main, Mini or Basket)	Basket
Value	0.5% (0.245%, 0.245%, 0.01%) Completion
Due	**Create phase**: Friday, January 28, 2022 at 3:03 p.m ET; **Assess phase**: Tuesday, February 1, 2022 at 3:03 p.m ET; **Reflect phase**: Friday, February 4, 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.

11.3.1 M2 Instructions

Make sure you are familiar with the general instructions for these types of tasks.

11.3.1.1 M2 Prompt

Read *Science isn't broken* (Aschwanden, 2015) using the *previewing* and *skimming* strategy outlined below. Discuss what you learn from the article, with a focus on what you **notice about your own thinking** when playing with the *"Hack Your Way To Scientific Glory"* interactive. Make sure you paraphrase appropriately (see Module 2 paraphrasing section) and add the reference for the article. (It is okay to put the reference by itself on a second page, if required.)

Instead of reading from beginning to end try ***previewing*** and then ***skimming*** to be faster:

1. Read the **title**. Pay attention to the "Filed Under" and date of publication.

Why do you think we have assigned this reading to you?

2. Skip the text until the **interactive figure** called "*Hack Your Way To Scientific Glory*."

When you change political party in section 1, or other items in section 2, how does the p-value change in section 4 of the interactive figure?

What do you understand from this figure? What do you notice about your own thinking as you play with it?

Tip: Reading one paragraph before and after a figure helps clarify the figure.

3. Skip the text until the **figure** of the football field.

What do you understand from this figure?

4. Skip the text until the **figure** called "*Same Data, Different Conclusions*."

What do you understand from this figure?

6. Read the two **bolded quotes**: "*Science is great, but…*" and "*Human fallibilities….*"
7. Read the **last paragraph** (often the concluding paragraph).
8. Now write your **200–500-word reflection** on what you have learned.
    - **Paraphrase!** That is, use your own words, don't copy/paste. Use the paraphrasing how-to from the week 2 module to help you.
    - Discuss what you noticed about your thinking while playing with the *Hack Your Way To Scientific Glory*" interactive.
    - Include a **conclusion** at the end.
    - Use in-text **citation** as follows (Aschwanden, 2015) and put the reference at the end.
9. Go back and read the entire text to see if your understanding was correct.

11.3.1.1.1 M2 Reference Aschwanden, C. (2015). *Science Isn't Broken: It's just a hell of a lot harder than we give it credit for*. Retrieved from https://fivethirtyeight.com/features/science-isnt-broken

11.3.1.2 M2 Rubric

11.4 Module 3 writing task

Information	Note
Name	Module 2 writing task
Type (Main, Mini or Basket)	Basket
Value	0.5% (0.245%, 0.245%, 0.01%) Completion

Information	Note
Due	**Create phase**: Friday, February 11, 2022 at 3:03 p.m ET; **Assess phase**: Tuesday, February 15, 2022 at 3:03 p.m ET; **Reflect phase**: Friday, February 18, 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.

11.4.1 M3 instructions

Make sure you are familiar with the general instructions for these types of tasks.

11.4.1.1 M3 prompt

Write the methods section of parts of the case study from Chapter 8 Case Study: Music Performance Anxiety (Roback & Legler, 2021).

I strongly suggest that you use **RMarkdown** for your writing in preparation for your final project. Knit and submit as a **PDF**. (You won't lose marks for doing it in Word, but if you're not comfortable in RMarkdown yet, I'd really recommend working on that! Guide to writing LaTex equations in Word.)

Write a short methods section that covers the final model from section 8.10.

- Add a reference to the version of R this analysis was conducted with AND a reference to the package used.
  - Recommend including a statement like the following: "*We fit a linear mixed model in R version 4.0.5 (R Core Team, 2021) using the* ***lme4*** *package (Bates, Maechker, Bolker, & Walker, 2015).*" (For this writing, it is okay to copy this exactly or modify is slightly.
    - You can find the version of R you are using by running `version` into your console.
    - You can get the appropriate R citation running `citation()` and `citation("lme4")` in your browser as well.
- Include the the **model,** written out in LaTeX.
  - In the current reading it is written as a series of equations. You should rewrite it to be a single **composite** equation, like the way we have been writing them in class.
  - **Explain each model term below the model**; paraphrase (do not just copy/paste) from the book.
- DO NOT plagiarize Roback and Legler (2021) or the original article (Sadler & Christopehr, 2010). Use your own words. You will be writing something shorter, anyway.
- Write it as if *you were* one of the authors of the study (you can use 'we').
- Include all of the methods components you can from the list below ('Methods components'). Section 8.2 will be especially helpful with this.
  - You also can access the original research article through the U of T library, if you wish.
    - Need help navigating library resources? Reach out to the Head Librarian at the Mathematical Sciences Library, Bruce Garrod. You can find his contact details in the Library Resources tab of the Course Navigation menu.
- Place these **references** at the end (can be on a ***second*** page IF required):

Component	Poor or Missing	Adequate	Good	Excellent
Addresses 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).
Paraphrasing	Text is too similar to the article, with only minimal changes made and/or quotes overused and/or sections copied verbatim without quotes. No referencing.	Some changes in wording and structure but may be vague and does not help the reader understand the main points. And/or use of quotes somewhat overdone.There is limited clarity about what is from Motulsky's article and what are the writer's own opinions.	Main ideas largely clear but may still be vague in some places and/or quotes are used where paraphrasing would have been more appropriate in some cases. It is mostly clear which ideas are coming from Motulsky's article and what is the opinion of the writer through referencing and use of language.	Clearly communicates ideas from the article using the student's own words. Structure and wording sufficiently different. It is clear which ideas are coming from Motulsky's article and what is the opinion of the student through referencing and use of language.

  - Roback, P., & Legler, J. (2021). 8 Introduction to Multilevel Models. In Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R. Retreived from https://bookdown.org/roback/bookdown-BeyondMLR/ch-multilevelintro.html. Boca Raton: CRC Press.
  - Sadler, Michael E., and Christopher J. Miller. 2010. "Performance Anxiety: A Longitudinal Study of the Roles of Personality and Experience in Musicians." *Social Psychological and Personality Science* 1 (3): 280–87. http://dx.doi.org/10.1177/1948550610370492.
  - Get the citation for R with `citation()` and include it.
  - Get the citation for **lme4** with `citation("lme4")` and include it.

- You may wish to use phrases from the Methods section of the Manchester Academic Phrase bank to help you describe your methods.

The methods section is usually written in **past tense** and often in the **passive voice** (Swales & Feak, 2012), though there is LOTS of discussion in academia about the use of passive vs active voice. We are not going to require you to write one way or the other for this—consistency is the key thing to aim for.

Methods components
Study design
Setting
Participants
Data sources/ measurement
Statistical methods

(b) Write the formula of the model

(c) Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Explain how quantitative variables were handled in the analyses.

Adapted from STROBE checklist (von Elm et al., 2007):
https://www.strobe-statement.org/index.php?id=available-checklists

11.4.1.1.1 Example from a *different* article: Modified from (Berlowitz et al., 2017):

(1) We conducted a study to compare the effect of two strategies to control blood pressure on physical health: an experimental intensive treatment and the standard treatment currently in use. After recruitment, participants were randomly allocated (randomization) to either the intensive treatment or the standard treatment. (2) The study involved recruitment from 102 clinical sites from November 2010 to March 2013. (3) A total of 9,361 participants 50 years of age or older who had evidence of hypertension were included, but those with diabetes or prior stroke were excluded. (4) The source of the data is the SPRINT study. Physical health was measured using a questionnaire with scores ranging from 0 to 100, with greater scores meaning better physical health. Data were collected at baseline and follow up.

(5a) To compare longitudinal change for physical health between the treatment groups, "we estimated a linear mixed model assuming linear mean change over time within each treatment group." (5b) "The model for (j^{th}physical health measurement for i^{th} participant measured at time t_{ij} (in years)) is as follows:

$= 0 + _1 + _2 + _3 + _1 + _2 +$ (5c) Where is the randomization assignment for the i^{th} participant (1=Intensive-treatment; 0=Standard-treatment), denotes the clinic site for the i^{th} participant (participants were randomized at 102 clinic sites), $_1$ is a subject-specific random effect, $_{2,}$ is a clinic-specific random effect, and denotes the residual error component."

"The models included participant-specific and clinic-specific random effects to address within-participant correlations as a result of repeated assessments and correlations among participants at the same [study] site. Fixed effects in the model included treatment group, follow-up time, and the interaction between the treatment group and follow-up time."

11.4.1.1.2 Reference: Berlowitz, D. R., Foy, C. G., Kazis, L. E., Bolin, L. P., Conroy, M. B., Fitzpatrick, P., . . . Group, S. R. (2017). Effect of Intensive Blood-Pressure Treatment on Patient-Reported Outcomes. *N Engl J Med, 377*(8), 733-744. doi:10.1056/NEJMoa1611179

11.4.1.2 References for this prompt

Berlowitz, D. R., Foy, C. G., Kazis, L. E., Bolin, L. P., Conroy, M. B., Fitzpatrick, P., . . . Group, S. R. (2017). Effect of Intensive Blood-Pressure Treatment on Patient-Reported Outcomes. *N Engl J Med, 377*(8), 733-744. doi:10.1056/NEJMoa1611179

Roback, P., & Legler, J. (2021). 8 Introduction to Multilevel Models. In Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R. Retreived from https://bookdown.org/roback/bookdown-BeyondMLR/ch-multilevelintro.html. Boca Raton: CRC Press.

Sadler, Michael E., and Christopher J. Miller. 2010. "Performance Anxiety: A Longitudinal Study of the Roles of Personality and Experience in Musicians." *Social Psychological and Personality Science* 1 (3): 280–87. http://dx.doi.org/10.1177/1948550610370492.

Swales, J., & Feak, C. B. (2012). *Academic writing for graduate students: essential tasks and skills* (3rd ed.). Ann Arbor: University of Michigan Press.

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gotzsche, P. C., Vandenbroucke, J. P., & Initiative, S. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet, 370*(9596), 1453-1457. doi:10.1016/S0140-6736(07)61602-X

11.4.1.3 M3 Rubric

	Poor or Missing	Adequate	Good	Excellent
Study information	0-1 items of study design, setting, participants, data sources and measures defined.	Only 2 items of study design, setting, participants, data sources and measures defined.	Only 3 items of study design, setting, participants, data sources and measures defined.	All 4 items of study design, setting, participants, data sources and measures defined.
Statistical methods	Neither statistical methods used named nor reason for their use given. Not all variables used in each statistical method described.	Not all statistical methods used named but reason for use given for those named. Variables used in each statistical method described.	All statistical methods used named but no reason for use given. Not all variables used in each statistical method described.	All statistical methods used named and reason for use given. Variables used in each statistical method described.
Model formula	No model formula included. Model terms not explained.	Model formula written but may not be correct. Model terms not explained clearly.	Model formula written but may not be correct. Each model term explained clearly.	Model formula written correctly. Each model term explained clearly.
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.

	Poor or Missing	Adequate	Good	Excellent
Paraphrasing	Methods and model terms described is too similar to the source material, with only minimal changes made and/or quotes overused and/or sections copied verbatim without quotes. In-text citation may or may not be used. References NOT provided at the end.			Methods and model terms described in the writer's own word and are sufficiently different from the source material provided. In-text citation may or may not be used. References provided at the end.

11.5 Module 4 writing task

Information	Note
Name	Module 4 writing task
Type (Main, Mini or Basket)	Basket
Value	0.5% (0.245%, 0.245%, 0.01%) Completion
Due	**Create phase**: Friday, March 04, 2022 at 3:03 p.m ET; **Assess phase**: Tuesday, March 08, 2022 at 3:03 p.m ET; **Reflect phase**: Friday, March 11, 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.

11.5.1 M4 instructions

Make sure you are familiar with the general instructions for these types of tasks.

11.5.1.1 M4 prompt

The rubric for this assessment (below) has some specific elements for this prompt, so make sure you take a look at it.

Write an **executive summary** for the case study in your reading this week on Poisson regression, section 4.10 of (Roback & Legler, 2021): https://bookdown.org/roback/bookdown-BeyondMLR/ch-poissonreg.html#cs:drinking.

An executive summary is a summary of a larger text mostly used in a business setting prepared for executives, managers and policymakers.

You will need to write one for your project.

	Executive Summary	Abstract
Summary of a larger text	Yes	Yes
Audience	Not technical: executives, managers, policymakers	Technical: researchers, experts, …
Background & Aim	Yes	Yes
Methods	No or brief: non-technical	Yes: mostly technical
Key findings / Limitations	Yes, as bullet points	Yes
Tables or figures	Sometimes	Almost never

	Executive Summary	Abstract
Length	1-2 pages	150–300 words

11.5.1.2 *Example* executive summary

Here is an executive summary from a consulting company whose report was presented to the Canadian Government on the economic impact of international education in Canada: https://www.international.gc.ca/education/report-rapport/impact-2017/index.aspx?lang=eng.

Notice the brief background and aims, in this case, without any headings and all in one paragraph. This is followed by bullet point findings and a final concluding statement which can be easily remembered by anyone: *"A 10% increase or decrease in spending by international students results in an approximately 10-11% change in economic impacts."*

Assignment format: This module writing has TWO pages, see the instructions below. (Still a similar amount of writing.)

Audience: University health and well-being team (managers, non-statisticians, non-technical).

11.5.1.3 Page 1: Text

Write the executive summary text. Use the following headings and structure:

Heading	Implied question	Writing style	How many?
Background & Aim	What do we know about the topic? What is this study about?	Paragraphs	1-2
Key findings	What was discovered?	Bullet points, one per finding	6
Limitations	What are the limitations?	Bullet points, one per limitation	1-3

- Describe **6 'Key findings'** and **1-3 'Limitations'** as bullet points.
- Refer to both Table 1 and Figure 1 in the bullet points at least once.
- Do NOT use p values, name of statistical tests, 95% CI, or other technical details.
- Median and interquartile range are okay but need explaining when first used.

11.5.1.4 Page 2: Table and Figure

- Copy/paste Table 1 (after modification, see "**Include Table 1"** bullet below) and Figure 1 to page 2 of your executive summary.
- **Do not include Table 2**.
- Refer to Table 1 and Figure 1 at least once in the main text, i.e., on page 1.

Heading	Implied question	Writing style	How many?
Table	-	Table with a caption on top	1
Figure	-	Figure with a caption below	1

- **Include Table 1** on page 2 of the executive summary. **Remove technical details from Table1** (p-value column and second footnote of names of statistical tests highlighted) before copy/pasting to page 2. Notice that the table caption is on top, as discussed in Module 2 and the writing workshop.

Table 1. Characteristics of students overall and based on whether they are a first-year or not.

Characteristic	Overall	First-year student	Not a first-year student	*p* value^b
Counts	$N = 77$	$n = 18$	$n = 59$	
Residence				0.2
Dorm	70 (91%)	18 (100%)^a	52 (88%)^a	
Off campus	7 (9.1%)	0 (0%)	7 (12%)	
Sex				0.13
Female	56 (73%)	16 (89%)	40 (68%)	
Male	21 (27%)	2 (11%)	19 (32%)	
Drinks last weekend	1 (0, 3)	0 (0, 1)	1 (0, 4)	0.023

^a n (%); Median (IQR)

Median separates the highest 50% of values from lowest 50% of values when values are arranged from lowest to highest (i.e. at least half of first year students had no alcoholic drinks the prior weekend).

IQR is interquartile range, representing the range of values from the bottom 25% to top 75% of values (i.e. 25% of first year students had at least 1 alcoholic drink the prior weekend).

[b] Fisher's exact test; Wilcoxon rank sum test.

- **DO NOT include Table 2** on page 2 of your executive summary. State findings in simple words in 'Key findings' (no p-values, 95% CI). If you decide to describe log(IRR), do so with simple words.

Table 2. Poisson model to predict the number of alcoholic drinks consumed during the prior weekend.

Characteristic	**log(IRR)^a**	**95% CI^a**	***p* value**
Residence			
Dorm	—	—	
Off campus	0.90	0.49, 1.3	<.001
Sex			
Female	—	—	
Male	1.1	0.80, 1.4	<.001

^a IRR = Incidence Rate Ratio, CI = Confidence Interval
Model: Number of alcoholic drinks ~ Off-campus status + Sex

- **Zero-inflated Poisson model:** State the estimated chance that a first-year student is a non-drinker in 'Key findings' as a bullet point (no p-values, no statistical test names in 'Key findings').

- **Include Figure 1** on page 2 of the executive summary. Include figure title/legend below figure.

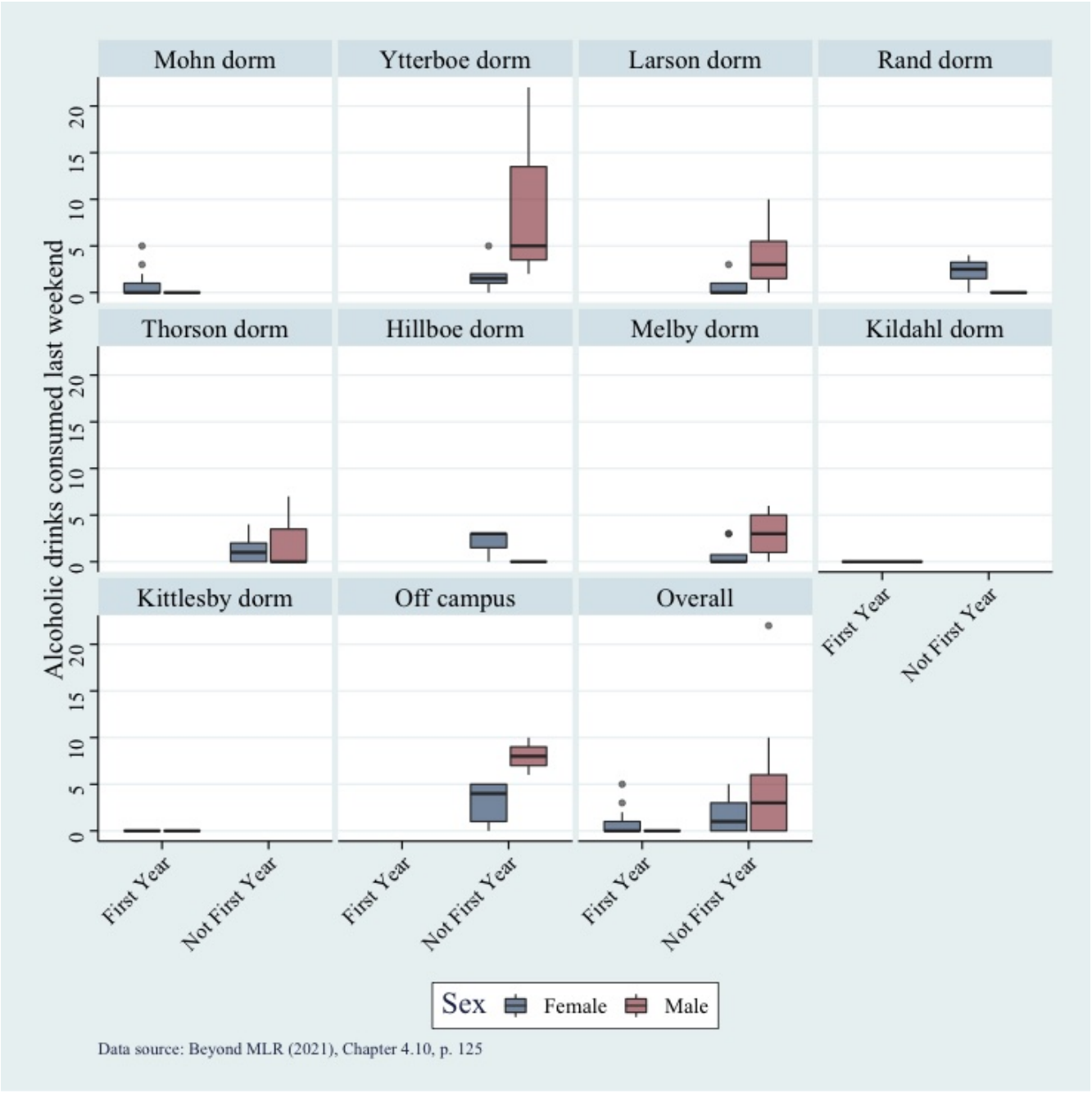


Figure 1. Distribution of drinking habits across 9 dorms, off campus residences, and overall, by sex and first year status of student. The horizontal *median line* in the middle of the boxes represents the point separating the highest 50% number of alcoholic drinks from the lowest 50% (the middle value). The top of the box represents 75% and the bottom of the box 25% of number of alcoholic drinks. The lines and dots extending beyond the box give the full range of values, with more extreme values further away from the box.

Note 1: The code/data is available in the Weekend Drinking case study but you can just copy objects from this prompt directly. Amin has used the `ggthemes` package to apply the 'Stata' theme to this particular plot.

Note 2: In your project executive summary you might want to place images and tables more strategically throughout the text, but for this short writing activity we didn't want you to spend time on formatting.

Note 3: It is unlikely that you will need to cite anything, as you are writing as if ***you*** are the one who conducted this study. If you were ever to post this writing outside of class, then you should carefully explain the assessment context AND reference the textbook from which the case study comes BUT for in-class submission purposes, don't worry about it.

Aim to spend about 30 minutes on this task. Write about 200–400 words responding to the prompt. 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 or bullets 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 introduced in the introduction and well summarized in a conclusion.
- There should be 2 pages, structured as instructed.

11.5.2 M4 Rubric

You will be asked to assess the writing of 2 peers on the following rubric. Peers will also assess you on this rubric, so it is important to keep in mind when you are writing.

	Poor or Missing	Adequate	Good	Excellent
Addresses prompt	Does not address the prompt for this week.	Only one item of the prompt is addressed or much of the response is not relevant/off-topic. Inappropriate for the audience. More than 1 mention of p-value or statistical test.	Prompt is addressed, though may go somewhat off-topic at points, or does not mention at least one of main points. Either Table 1 or Figure 1 is not used. Mostly appropriate for the audience. One mention of a p-value or a statistical test.	Prompt is answered with regard to all three main points: association of sex, residence and first-year status on alcoholic drinking. Table 1 and Figure 1 are used. Appropriate for the audience. No p-values or names of statistical tests.
Structure	There is no structure, very difficult to follow. No paragraphs, single block of text.	1–2 headings used. Neither 'Key findings' or 'Limitations' are in bullet points. More than 1 paragraph/bullet point contains unrelated ideas, or more than 1 paragraph does not introduce the topic of the paragraph at the beginning. Some structure but difficult to follow.	3–4 headings used. Either 'Key findings' or 'Limitations' are not in bullet points. At least 1 paragraph/bullet point contains an unrelated idea and/or at least 1 paragraph does not introduce the topic of the paragraph at the beginning. Well organized, follows a logical structure.	Five headings used. 'Key findings' and 'Limitations' are in bullet points. Each paragraph/bullet point contains a clear main point and related ideas. Paragraphs introduce the topic of the paragraph at the beginning. Well organized, follows a logical structure.

	Poor or Missing	Adequate	Good	Excellent
Writing mechanics	Considerable writing and grammatical issues that obscure the meaning OR lots of slang and inappropriate word choice.	Multiple sentences/phrases are difficult to read, have slang or inappropriate word choice, or grammar problems obstructing meaning.	Slight difficulty in understanding one or two sentences/phrases but otherwise understandable.	Can read and follow along with minimal effort. Some grammatical or word choice errors are allowable, but they must not obstruct meaning.

11.5.3 M4 Reference

Roback, P., & Legler, J. (2021). 4.10 Case Study: Weekend Drinking. In *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R* (1. ed., pp. 436). Boca Raton: CRC Press.

11.6 Module 5 writing task

Information	Note
Name	Module 5 writing task
Type (Main, Mini or Basket)	Basket
Value	0.5% (0.245%, 0.245%, 0.01%) Completion
Due	**Create phase**: Friday, March 25, 2022 at 3:03 p.m ET; **Assess phase**: Tuesday, March 29, 2022 at 3:03 p.m ET; **Reflect phase**: Friday, April 01, 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.

11.6.1 M5 instructions

Make sure you are familiar with the general instructions for these types of tasks.

11.6.1.1 M5 prompt

Summarize and paraphrase section **6.7.5 Model Discussion and Summary** of (Roback & Legler, 2021a). The first paragraph is key results, and the next 3 paragraphs are limitations. Paraphrase and summarize key results and limitations into two paragraphs, and then write your own conclusion in a few sentences.

11.6.1.2 Address the prompt using the following material:

Section	Activity	Location	Source
Study aim	Paraphrase	Paragraph 6.7 Case Study: Trying to Lose Weight: We are interested in… 4	
Key results	Paraphrase	Paragraph 6.7.5 Model Discussion and Summary: We found that… 1	

Section	Activity	Location	Source
Limitations	Summarise & Paraphrase	Paragraphs 2,3,4	6.7.5 Model Discussion and Summary: A limitation of …
Conclusion	Write		Your own thoughts

11.6.1.3 Structure of writing

Discussion

Part	To do	Length
Study aim(s) & Key results	(a) Restate study aim(s) from introduction in condense format (b) Summarize key results	1 paragraph
Limitations	Discuss limitations of the study, taking into account sources of potential bias or imprecision, such as: (a) data missingness (b) definitions and selection or survivorship bias (c) data collection (survey) methods (d) confounders, if any	1 paragraph
Conclusion	Give a cautious overall interpretation of results considering aim(s), key results, and limitations	1-3 sentences
Reference	Place reference at the end of your writing. In-text citation is not required, aim to write as if you were part of the study. Roback, P., & Legler, J. (2021). 6.7.5 Model Discussion and Summary. In *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R* (1. ed., pp. 180-181). Boca Raton: CRC Press.	1 reference

Adapted from STROBE checklist (von Elm et al., 2007).

11.6.1.4 Useful phrases to consider

Examples and resources to use with or without modification. You can also write your own.

Part	Example phrases
Study aim & Key results	This study set out to … The aim of the present research was to examine … The current study found that … This experiment did not detect any evidence for … Surprisingly, no differences were found in ….
Limitations	There are three limitations to this study. The first limitation is … These findings may be somewhat limited by … These data must be interpreted with caution because …
Conclusion	In conclusion … The results of this study indicate that … These findings may help us to understand … Notwithstanding these limitations, the study suggests that …

Adapted from Manchester Academic Phrase Bank sections: Discussing Findings and Writing Conclusions.

11.6.1.5 How to restate study aims in the discussion

There are four aims in the *introduction* of the current study (paragraph 4 of 6.7 Case Study: Trying to Lose Weight: We are interested in…), but when restating the aims in the *discussion* we can summarize and paraphrase them by condensing them to a sentence or two instead of repeating all the questions or aims. One good example is sentence 3 of paragraph 1 of 6.7 Case Study: Trying to Lose Weight: "Here we examine characteristics of young people who are trying to lose weight." We want a little more specificity for our discussion, try to condense the four aims into a sentence (or two) of your own.

11.6.1.5.1 Example 1 (Roback & Legler 11.2 Case Study: College Basketball Referees) Aims in the introduction: "Data have been gathered from the 2009–2010 college basketball season for three major conferences to investigate the following questions (Noecker & Roback, 2012):

- Does evidence that college basketball referees tend to "even out" calls still exist in 2010 as it did in 2005?
- How do results change if our analysis accounts for the correlation in calls from the same game and the same teams?
- Is the tendency to even out calls stronger for fouls over which the referee generally has greater control? Fouls are divided into offensive, personal, and shooting fouls, and one could argue that referees have the most control over offensive fouls (for example, where the player with the ball knocks over a stationary defender) and the least control over shooting fouls (where an offensive player is fouled in the act of shooting).
- Are the actions of referees associated with the score of the game?" (Roback & Legler, 2021b)

Condensed aims for discussion: This study set out to investigates the existence of "even out" calls in basketball games, and their correlation with four items: player teams, game being played, foul type and score of game. The current study found that …

11.6.1.5.2 Example 2 Aims in the introduction: "An important question related to pandemic preparedness remains unanswered: what killed people during the 1918–1919 pandemic and subsequent influenza pandemics? In the present study, we have examined recut tissue specimens obtained during autopsy from 58 influenza victims in 1918–1919, and have reviewed epidemiologic, pathologic, and microbiologic data from published reports for 8,398 post-mortem examinations bearing on this question. We have also reviewed relevant information, accumulated over 9 decades, related to the circulation of descendants of the 1918 virus." (Morens, Taubenberger, & Fauci, 2008)

Condensed aims for discussion: We examined epidemiologic, pathologic, and microbiologic evidence from the 1918–1919 pandemic continuing to subsequent pandemics from similar viruses to determine the likely cause of death during such pandemics. The results of this investigation show that …

11.6.2 M5 Rubric

You will be asked to assess the writing of 2 peers on the following rubric. Peers will also assess you on this rubric, so it is important to keep in mind when you are writing.

	Poor or Missing	Adequate	Good	Excellent
Addresses prompt	Does not address the prompt for this week.	Only one item of the prompt is addressed or much of the response is not relevant/off-topic. Inappropriate for the audience. More than 1 mention of p-value or statistical test.	Prompt is addressed, though may go somewhat off-topic at points, or does not mention at least one of main points. Either Table 1 or Figure 1 is not used. Mostly appropriate for the audience. One mention of a p-value or a statistical test.	Prompt is answered with regard to all three main points: association of sex, residence and first-year status on alcoholic drinking. Table 1 and Figure 1 are used. Appropriate for the audience. No p-values or names of statistical tests.
Structure	There is no structure, very difficult to follow. No paragraphs, single block of text.	1–2 headings used. Neither 'Key findings' or 'Limitations' are in bullet points. More than 1 paragraph/bullet point contains unrelated ideas, or more than 1 paragraph does not introduce the topic of the paragraph at the beginning. Some structure but difficult to follow.	3–4 headings used. Either 'Key findings' or 'Limitations' are not in bullet points. At least 1 paragraph/bullet point contains an unrelated idea and/or at least 1 paragraph does not introduce the topic of the paragraph at the beginning. Well organized, follows a logical structure.	Five headings used. 'Key findings' and 'Limitations' are in bullet points. Each paragraph/bullet point contains a clear main point and related ideas. Paragraphs introduce the topic of the paragraph at the beginning. Well organized, follows a logical structure.
Writing mechanics	Considerable writing and grammatical issues that obscure the meaning OR lots of slang and inappropriate word choice.	Multiple sentences/phrases are difficult to read, have slang or inappropriate word choice, or grammar problems obstructing meaning.	Slight difficulty in understanding one or two sentences/phrases but otherwise understandable.	Can read and follow along with minimal effort. Some grammatical or word choice errors are allowable, but they must not obstruct meaning.

11.6.3 M5 References

Morens, D. M., Taubenberger, J. K., & Fauci, A. S. (2008). Predominant Role of Bacterial Pneumonia as a Cause of Death in Pandemic Influenza: Implications for Pandemic Influenza Preparedness. *The Journal of Infectious Diseases, 198*(7), 962-970. doi:10.1086/591708

Noecker, C. A., & Roback, P. (2012). New Insights on the Tendency of NCAA Basketball Officials to Even Out Foul Calls. *Journal of Quantitative Analysis in Sports, 8*(3). doi:doi:10.1515/1559-0410.1402

Roback, P., & Legler, J. (2021a). 6.7.5 Model Discussion and Summary. In *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R* (1. ed., pp. 180-181). Boca Raton: CRC Press.

Roback, P., & Legler, J. (2021b). 11.2 Case Study: College Basketball Referees. In *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R* (1. ed., pp. 374-400). Boca Raton: CRC Press.

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gotzsche, P. C., Vandenbroucke, J. P., & Initiative, S. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet, 370*(9596), 1453-1457. doi:10.1016/S0140-6736(07)61602-X

12

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 semster.

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.

12.1 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.

12.1.1 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).

12.1.2 Submission requirements

Your proposal should be:

- typed (not handwritten)
- one page[1]
- 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.

12.1.3 Rubric

[1] References, 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

12.1.4 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.

12.1.5 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

12.1.6 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.

12.1.6.1 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 …"*

12.1.6.2 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"*

12.1.6.3 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?

12.1.6.4 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."*

12.1.6.5 Time-bounded:

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

Specific	Measurable	Attainable	Relevant	Time-Bound
Make sure your goals are focused and identify a tangible outcome. Without the specifics, your goal runs the risk of being too vague to achieve. Being more specific helps you identify what you want to achieve. You should also identify what resources you are going to leverage to achieve success.	You should have some clear definition of success. This will help you to evaluate achievement and also progress. This component often answers how much or how many and highlights how you'll know you achieved your goal.	Your goal should be challenging, but still reasonable to achieve. Reflecting on this component can reveal any potential barriers that you may need to overcome to realize success. Outline the steps you're planning to take to achieve your goal.	This is about getting real with yourself and ensuring what you're trying to achieve is worthwhile to you. Determining if this is aligned to your values and if it is a priority focus for you. This helps you answer the why.	Every goal needs a target date, something that motivates you to really apply the focus and discipline necessary to achieve it. This answers when. It's important to set a realistic time frame to achieve your goal to ensure you don't get discouraged.

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

12.2 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.

12.2.1 Templates

- template.docx
- template.Rmd, template.pdf, template_rmd.zip

12.2.2 Reflection and evidence components

12.2.2.1 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.)

12.2.2.1.1 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?

12.2.2.1.2 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?

12.2.2.1.3 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 …*

12.2.2.2 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.

12.2.2.2.1 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 …*

12.2.2.3 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

12.2.2.3.1 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	0	Example: No progress this week
	Feb 21		*Reading week*
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] …

12.2.3 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: • activity, • alignment, • SMART lessons
2	Evidence	• Selected screenshots of work done, links, …
3	Timesheet	• Fill provided template or create your own • 1-3 sentences of what you learned

12.2.4 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

13

Knowledge basket: Other

13.1 '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.

13.1.1 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.)

13.2 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.

13.2.1 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.

You are welcome to use any calculator you wish and/or R.

There are 15 points across 16 questions, so you have about 4 minutes per point. The structure is outlined below. You may find having read the syllabus first helpful for question 1.

Question #	Type	General topic area	Points
1	Tick all that apply	Academic integrity	1
2	Single choice (MCQ)	Paraphrasing	1
3	Fill in the blanks (3 numeric)	Rounding	1
4	MCQ	Linear regression assumptions	1
5	MCQ	Linearity	1
6	Numeric	Distributions	1
7	Numeric	Distributions	1
8	MCQ	Distributions	1
9	MCQ	Confidence intervals	1
10	MCQ	Algebra	1
11	MCQ	Likelihood	1
12	Numeric	Proportions	0.5
13	Numeric	Proportions	0.5
14	MCQ	Multiple linear regression	1
15	MCQ	Multiple linear regression	1
16	MCQ	P-values	1

13.2.1.1 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)

13.2.2 Workshop attendance

If you didn't earn a full 1% from the pre-knowledge quiz (either didn't score over 80% or didn't complete it), you **can still earn 0.5 points towards your knowledge basket** by participating in this workshop.

- For your attendance to count you must attend **at least 75% of a single 50-minute workshop** AND **complete the post-event survey**.
- Please arrive on time!
- You can come to both if you want (they will be the same content though!) but cannot double up on points.
- A recording of the session will be posted later, for your reference, but watching that will NOT be eligible for points.
- You can do a **practice version** of the pre-knowledge check as many times as you like. There is also commentary on some of the question after you submit.

13.3 Writing workshop

Information	Note
Name	Writing workshop attendance and participation
Type (Main, Mini or Basket)	Basket
Value	0.25% for attendance AND survey
Due	Attendance on February 16 and submission of survey by 3:03 p.m. on Friday, February 18
Submission instructions	Submission: Via Quercus surveys Untimed, unlimited attempts
Late submissions, accommodations, and extension policy	No late submissions, accommodations, or extensions.

13.3.1 Workshop attendance

- To 'participate fully' you must attend **at least 75% of a single 50-minute workshop** AND **complete the post-event survey**.
- Please arrive **on time**!
- You can come to both if you want (they will be the *same content* though!) but *cannot* double up on points or 'share' attendance across the two.
- A **recording** of the session will be posted later, for your reference, but watching that will NOT be eligible for points.

13.4 5 Ways to Well-being workshop attendance and reflection

Information	Note
Name	5 Ways to Well-being workshop attendance and reflection
Type (Main, Mini or Basket)	Basket
Value	0.5% for attendance AND reflection
Due	Attendance on Wednesday March 30 at 10:10 a.m. ET and submission of survey by 3:03 p.m. on Friday, April 1
Submission instructions	Submission: Via Quercus surveys Untimed, unlimited attempts
Late submissions, accommodations, and extension policy	No late submissions, accommodations, or extensions.

13.4.1 5WtW information

5 Ways To Wellbeing: Investing in practices that are protective of our mental health have never been more important. Drawing on evidence from the study of wellbeing, this participatory webinar will provide an opportunity to explore and practice a cluster of activities that promote resiliency and support us in feeling and functioning better.

To get the most out of this workshop please:

- Be prepared to use your webcam. That said, don't feel you need to make yourself or your space look pretty. We'll be striving for a learning environment that's free of judgment (of others and ourselves).
- Have your cell phone handy.

- Give yourself permission to be present. Don't feel the need to be multi-tasking.
- Bring along your snack, tea, coffee, or other refreshments that you enjoy (this is a self-care workshop after all).

13.4.2 Workshop attendance

- To 'participate fully' you must attend **at least 75% of a single 90-minute workshop** AND **complete the post-event reflection**.
- Please arrive **on time**! There is a Zoom cap of 300, at which point we won't be able to let any more students in.
- There is no recording.
- You CAN earn points for attending BOTH this and the other *different* workshop on the same day.

13.5 Academic resilience workshop attendance and reflection

Information	Note
Name	Academic resilience workshop attendance and reflection
Type (Main, Mini or Basket)	Basket
Value	0.5% for attendance AND reflection
Due	Attendance on Wednesday March 30 at 3:10 p.m. ET and submission of survey by 3:03 p.m. on Friday, April 1
Submission instructions	Submission: Via Quercus surveys Untimed, unlimited attempts
Late submissions, accommodations, and extension policy	No late submissions, accommodations, or extensions.

13.5.1 Workshop attendance

- To 'participate fully' you must attend **at least 75% of a single 90-minute workshop** AND **complete the post-event reflection**.
- Please arrive **on time**! There is a Zoom cap of 300, at which point we won't be able to let any more students in.
- There is no recording.
- You CAN earn points for attending BOTH this and the other *different* workshop on the same day.

13.6 Module check-ins

Information	Note
Name	Module check-in 1–5
Type (Main, Mini or Basket)	Basket
Value	0.1% for completion
Due	Second Fridays of Modules at 3:03 p.m.
Submission instructions	Submission: Via Quercus surveys Untimed, unlimited attempts

Information	Note
Late submissions, accommodations, and extension policy	No late submissions, accommodations, or extensions.

13.6.1 Instructions

Module check-ins are a short questionnaire to help me 'take the pulse' of the class. Each survey will like take only 5–10 minutes and will be available about 6:00 p.m. on the second Wednesday of each module, until 3:03 p.m. ET on the second Friday.

13.7 Team Up! activities

Information	Note
Name	Team Up! activities (aiming for one per module)
Type (Main, Mini or Basket)	Basket
Value	0.5% (graded on quality)
Due	End of tutorial time (first Thursdays)
Submission instructions	Submission: Via Team Up! (links through Quercus) Usually ~50 minutes, one attempt (as a group), but with options to retry incorrect questions (for partial points).
Late submissions, accommodations, and extension policy	No late submissions, accommodations, or extensions.

See the course overview page on Quercus for the links to the versions with answers.

13.8 Graduate school info session

Information	Note
Name	Module 1 graduate school info session and panel
Type (Main, Mini or Basket)	Basket
Value	0.1%
Due	Friday, January 21, 2022 at 3:03 p.m ET
Submission instructions	Submission: Via Quercus
Late submissions, accommodations, and extension policy	No late submissions, accommodations, or extensions.

There will be two versions of this session, each in class time on January 19, 2022, 10:10 to 11:00 a.m. and 3:10 to 4:00 p.m.

These sessions will have a short presentation from the administrators/program managers of some potentially relevant U of T graduate programs as well as a panel discussion with some of the course TAs. They will talk about their experiences, things they wish they'd know, and reference letters.

10:00 a.m. class: Master of Financial Insurance (MFI) and Master of Science in Applied Commuting (MScAC) administrators + TA panel

3:00 p.m. class: Master of Science in Statistics and PhD in Statistics administrator (note: Alison will also share some general advice for if you're looking at program elsewhere as well) + TA panel

To earn the 0.1 points up for attending a grad school information session and panel you must attend for at least 75% of one of the sessions (i.e., you cannot just log in and then log out again) AND complete the short questionnaire. There are three questions, you must answer Q1 and 2, but Q3 is optional.

Notes:

- You *cannot* earn double points for attending both sessions.

- Attendance is not required, there are plenty of other ways to earn knowledge basket points.

- You can attend either session (or both, if you really wanted). There are no section restrictions for STA303.
- The session is *not* being recorded.
- PollEverywhere has nothing to do with your attendance/knowledge basket points. You don't have to engage with it if you do not want to.

13.9 Punctuation art

Information	Note
Name	Punctuation art
Type (Main, Mini or Basket)	Basket
Value	1%
Due	Friday, April 8, 2022 at 3:03 p.m. ET
Submission instructions	Submission: Via Quercus
Late submissions, accommodations, and extension policy	No late submissions, accommodations, or extensions.

13.9.1 Instructions

This is a self-contained mini-code and communication activity. It can be submitted anytime before the due date.

Note: If you like this idea and want to build it out further into your professional development task, or an aspect of it, you *can*, but this task/prompt can only be used once. I.e., if you do a punctuation art based component for your professional development, you cannot also submit this assessment.

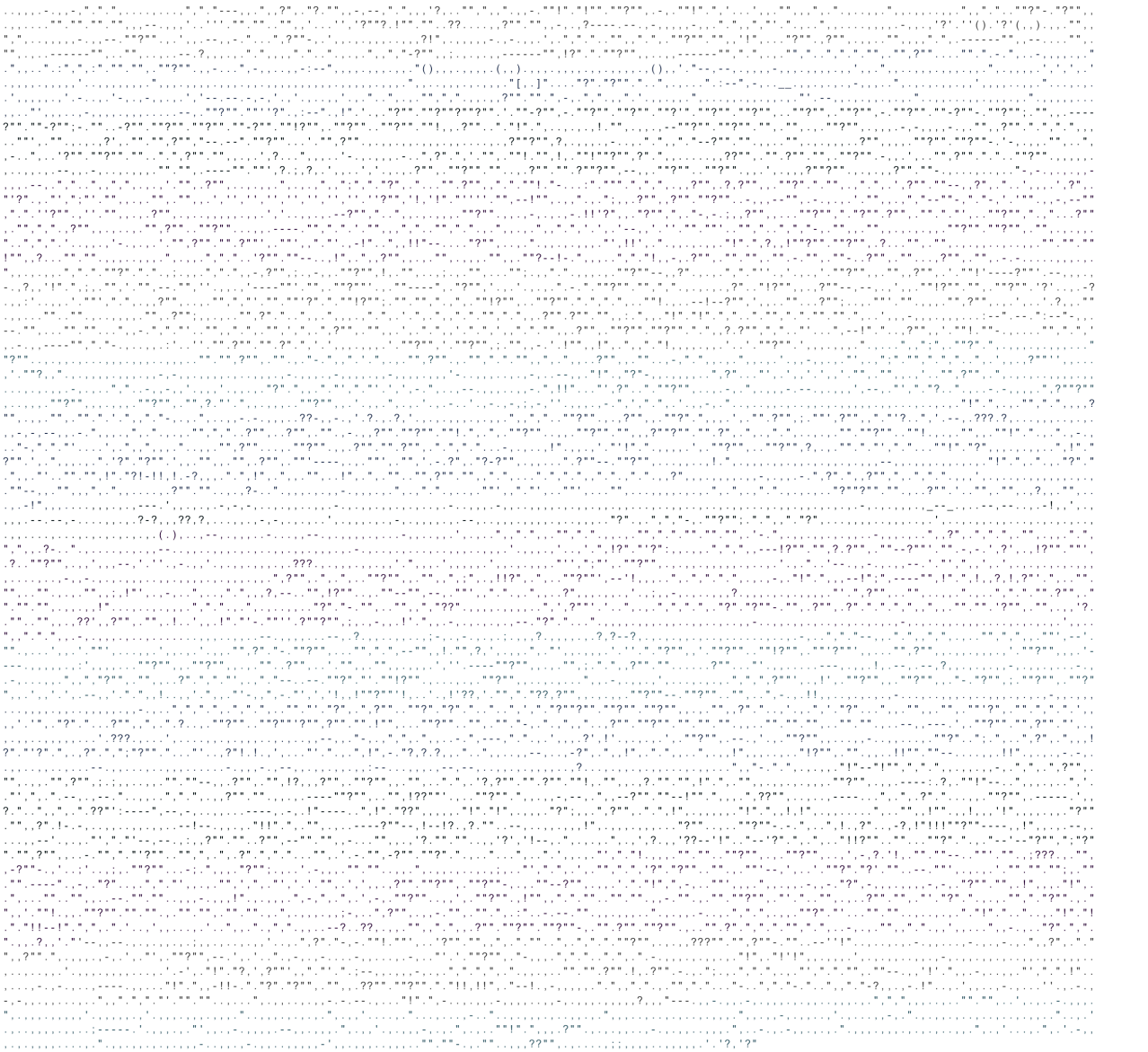
13.9.1.1 Prompt

Use R to create a representation of the entire text of a book, using only the punctuation.

13.9.1.2 Examples

- https://www.theguardian.com/books/gallery/2016/feb/23/say-what-books-with-the-words-removed-punctuation-maps-in-pictures
- https://www.mentalfloss.com/article/75602/what-famous-novels-look-stripped-everything-punctuation

The Hound of the Baskervilles in punctuation



@Liza_Bolton, text source: Projet Gutenburg

13.9.1.3 Skills

- regular expressions
- ggplot (could use base if you prefer)

13.9.1.4 Tips

- Works in the public domain are probably going to be the easiest to access. Project Gutenburg has .txt versions of many classics.

```
# Read in the text of hounds of the Baskervilles and remove the front matter and end matter
text <- readLines("https://www.gutenberg.org/cache/epub/3070/pg3070.txt")[0:7298+61]
```

13.9.1.5 Submission instructions

Submit the following:

- Commented code that produces your punctuation art (Rmd file)
- Your final art output (PNG recommended)
- A PDF one-page description that:
  - introduces what your output shows,
  - how you got the text data (make sure you credit the source),
  - notable steps in preparing the data to create your art, and
  - a brief summary of the skills you learned or demonstrated in this task.
- If you would be willing to have your art shared with future students, please include at the end of your write-up "I give my permission for my punctuation art to be shared [anonymously | and credited to me,]." Update the statement appropriately, e.g., *I give my permission for my punctuation art to be shared and credited to me, Liza Bolton.*. If you have posted it on social media or a personal website, you're welcome to also provide that link.

13.9.1.6 Rubric

	Missing or insufficient 0%	**Partial** 0.5%	**Complete** 1%
Task completion	No submission OR instructions insufficiently satisfied.	Some appropriate work shown, but may be somewhat incomplete or irrelevant.	Prompt/instructions sufficiently satisfied.

13.10 Breathing exercise

Information	**Note**
Name	Breathing exercise
Type (Main, Mini or Basket)	Basket
Value	1%
Due	Friday, April 8, 2022 at 3:03 p.m. ET
Submission instructions	Submission: Via Quercus
Late submissions, accommodations, and extension policy	No late submissions, accommodations, or extensions.

13.10.1 Instructions

This is a self-contained mini-code and communication activity. It can be submitted anytime before the due date.

Note: If you like this idea and want to build it out further into your professional development task, or an aspect of it, you *can*, but this task/prompt can only be used once. I.e., if you do a `gganimate` breathing based component for your professional development, you cannot also submit this assessment.

13.10.1.1 Prompt

Use R to create a box breathing exercise GIF. Box breathing involves inhaling for 4 seconds, holding that breath for 4 seconds, exhaling for 4 seconds, holding for 4 seconds and then repeating as many times as desired.

13.10.1.1.1 Background Breath like a Navy SEAL? Checkout this interesting article about stress and breathing.

13.10.1.1.2 A note If focusing on breathing ends up making you feel anxious, or just isn't the appropriate choice for you and your body right now, remember that you can create this without necessarily having to use it AND that there are lots of other ways to earn your knowledge basket points.

13.10.1.2 Examples

Not created in R:

There is a GIF in the web version. It is viewable at: https://smho-smso.ca/wp-content/uploads/2020/06/Four-square-breathing-EN.gif

There is a GIF in the web version. It is viewable at: https://quietkit.com/img/box-breathing-4x-v03.gif

A few I whipped up on a Saturday evening:

There is a GIF in the web version.

There is a GIF in the web version.

13.10.1.3 Skills

- data simulation
- dynamic data visualization with `ggplot2` and `gganimate`

13.10.1.4 Tips/links

- gganimate website and tutorials
  - Cheat sheet
  - `transision_states` was the `gganimate` function I found particularly helpful.
- `theme_void()` removes all the background and axes labels, which is really helpful for creating an image like these.
- You can use `anim_save` to save the outputs as a GIF file.
- If you want to play with colour palettes, I like playing with https://coolors.co/
  - I used `scale_color_gradient()` for the circle (sets two colours, 'high' and 'low') and `scale_colour_gradientn()` in the GIF with the flying dots (4 colours transitioned between). Here is the documentation for these functions.

13.10.1.5 Submission instructions

Submit the following:

- **Commented** code that produces your breathing exercise (Rmd file)
- Your final activity output, as a GIF
- A PDF one-page description that:
  - introduces what your breathing exercise is and provides some background, and
  - provide a brief summary of the skills you learned or demonstrated in this task.
- If you would be willing to have your exercise shared with current/future students, please include at the end of your write-up "I give my permission for my breathing exercise to be shared [anonymously | and credited to me,]." Update the statement appropriately, e.g., *I give my permission for my* breathing exercise *to be shared and credited to me, Liza Bolton.*. If you have posted it on social media or a personal website, you're welcome to also provide that link.

13.10.1.6 Rubric

	Missing or insufficient 0%	**Partial** 0.5%	**Complete** 1%
Task completion	No submission OR instructions insufficiently satisfied.	Some appropriate work shown, but may be somewhat incomplete or irrelevant.	Prompt/instructions sufficiently satisfied.

13.11 Linear Mixed Models study guide

Information	**Note**
Name	Linear Mixed Models study guide
Type (Main, Mini or Basket)	Basket
Value	1%
Due	Tuesday, March 15, 2022 at 3:03 p.m ET
Submission instructions	Submission: Via Quercus discussion boards
Late submissions, accommodations, and extension policy	No late submissions, accommodations, or extensions.

13.11.1 Instructions

Create a 1 page **study guide** for the topic of **Linear Mixed Models** (Module 3), as relevant to this course, and post it to the class discussion board for this topic.

13.11.1.1 Considerations

- What are the main topics, key vocabulary, useful equations?
- That page can be whatever size you choose, but I'd suggest aiming for a printable letter or legal size page.
- The guide should be submitted in PDF format. PLEASE. I'm always disappointed by the number of folks that miss this instruction.
- Text should *NOT* be hand-written, in order to ensure it is readable for your classmates.
- While not a requirement, consider how you can make your guide most accessible. Image descriptions? Easy to read fonts? Colour-blindness friendly?

13.11.1.2 Rubric

	Missing or insufficient 0%	**Partial** 0.5%	**Complete** 1%
Task completion	No submission OR instructions insufficiently satisfied.	Some appropriate work shown, but may be somewhat incorrect, incomplete or irrelevant.	Prompt/instructions sufficiently satisfied.

13.12 Generalized Linear Models study guide

Information	**Note**
Name	Generalized Linear Models study guide
Type (Main, Mini or Basket)	Basket
Value	1%
Due	Tuesday, March 15, 2022 at 3:03 p.m ET
Submission instructions	Submission: Via Quercus discussion boards
Late submissions, accommodations, and extension policy	No late submissions, accommodations, or extensions.

13.12.1 Instructions

Create a 1 page **study guide** for the topic of **Generalized Linear Models (Module 4)**, as relevant to this course, and post it to the class discussion board for this topic.

13.12.1.1 Considerations

- What are the main topics, key vocabulary, useful equations?
- That page can be whatever size you choose, but I'd suggest aiming for a printable letter or legal size page.
- The guide should be submitted in PDF format. PLEASE. I'm always disappointed by the number of folks that miss this instruction.
- Text should *NOT* be hand-written, in order to ensure it is readable for your classmates.
- While not a requirement, consider how you can make your guide most accessible. Image descriptions? Easy to read fonts? Colour-blindness friendly?

13.12.1.2 Rubric

	Missing or insufficient 0%	**Partial** 0.5%	**Complete** 1%
Task completion	No submission OR instructions insufficiently satisfied.	Some appropriate work shown, but may be somewhat incorrect, incomplete or irrelevant.	Prompt/instructions sufficiently satisfied.

13.13 Hack your class workshop attendance and reflection

Information	Note
Name	Hack your class workshop attendance and reflection
Type (Main, Mini or Basket)	Basket
Value	0.25%
Due	Friday March 4, 2022 at 3:03 p.m. ET
Submission instructions	Submission: Via Quercus
Late submissions, accommodations, and extension policy	No late submissions, accommodations, or extensions.

13.13.1 Instructions

> *Reboot and refresh after Reading Week! Register for this workshop to learn how to Hack Your Class to get the absolute most out of your experience.*

To earn these points, you must:

- Register for and attend at least 75% of the 'Hack your class' workshop, hosted by the Department of Statistical Sciences on **Thursday March 3rd, 10 to 11am ET**, AND
- Submit a short reflection about what you learned from the session on Quercus by 3:03 p.m. ET on Friday, March 4.

Register here

13.14 Sports analytics workshop attendance and reflection

Information	Note
Name	Sports analytics workshop attendance and reflection
Type (Main, Mini or Basket)	Basket
Value	0.25%
Due	Friday March 4, 2022 at 3:03 p.m. ET
Submission instructions	Submission: Via Quercus
Late submissions, accommodations, and extension policy	No late submissions, accommodations, or extensions.

13.14.1 Instructions

To earn these points, you must:

- Register for and attend at least 75% of the 'Sports analytics' workshop, hosted by the Department of Statistical Sciences (and led by two former STA303 students!) on **Thursday, March 3, from noon to 1:30 p.m. ET**, AND
- Submit a short reflection about what you learned from the session on Quercus by 3:03 p.m. ET on Friday, March 4.

Register here

Modules

14

Module 1

Materials for January 10–21, 2022.

14.1 Learning checklist

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

- Understand the structure and policies of STA303.
- Be able to open and work with an RMarkdown document, knit to pdf and export that pdf (RStudio, JupyterHub).
- [STA302 review] Explain the assumptions of linear regression and identify situations where they might be violated.
- Describe the execution and purpose of several common statistical tests (one-sample t-test, paired sample t-test, two-sample t-test, one-way ANOVA), and identify how to conduct them in a linear regression framework.
- Create a reproducible example (reprex) for use in explaining an error being produced by your code.

14.2 Instructor information

Prof. Liza Bolton

Course email: sta303@utoronto.ca

Office 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.

14.3 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.

[1] I used to haaaate 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.

While our course isn't a business, some of the basic parts of this concept apply well to your time at university 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.

14.3.1 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.
  - 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.

14.3.2 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 your 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.

14.3.3 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.

[2] "Dear Madam" and "To my esteemed professor" honestly make me uncomfortable.

14.3.4 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 organized 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 your 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.

14.4 Recap of linear models

14.4.1 Why model?

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

14.4.2 Linear models

In a linear model,

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

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

a.k.a.

DATA = MODEL + ERROR

14.4.3 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,

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

14.4.4 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.

14.4.4.1 Linear

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

14.4.4.2 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.

14.4.5 Optional refresher reading

14.4.5.1 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]

14.4.5.2 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]

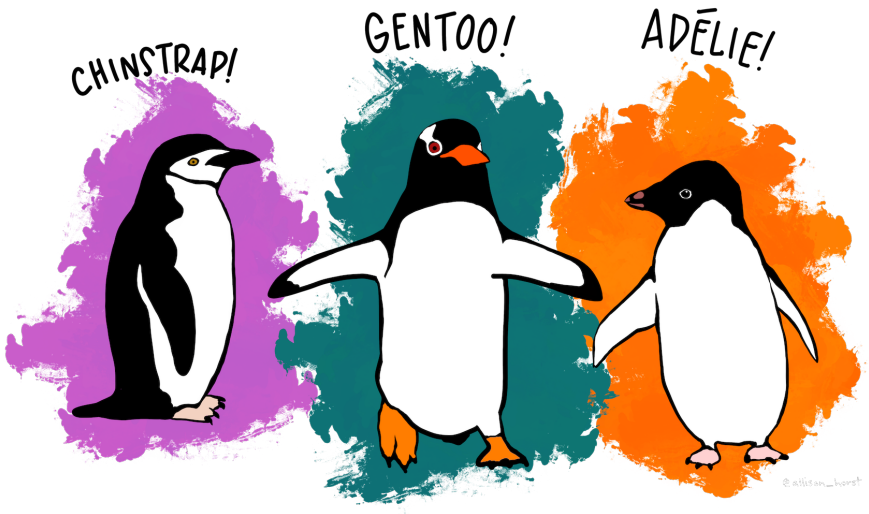
14.4.5.3 Delicious mathematics

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]

14.5 The data for module 1

The data we'll be looking for the next section of this week is from the `palmerpenguins` package by Allison Horst.

```
# install.packages("palmerpenguins")
library(palmerpenguins)
```



Artwork by @allison_horst

There is a GIF in the web version. Not required content.

14.5.1 Let's meet the penguins

```
library(tidyverse)

glimpse(penguins, width = 80) # width here just influences how this is displayed
```

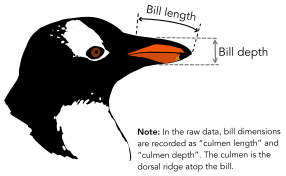
```
## Rows: 344
## Columns: 8
## $ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel~
## $ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse~
## $ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, ~
## $ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, ~
## $ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186~
## $ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, ~
## $ sex               <fct> male, female, female, NA, female, male, female, male~
## $ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

The `glimpse` function comes from the `tibble` package, which is in loaded by the `tidyverse` package.

One of the most common issues students encounter when learning R is that they are trying to use functions that they haven't installed or loaded the package for. Make sure you've read the slides about RStudio on the JupyerHub because packages you install there will need to be **reinstalled each session** if they are not the list of automatically available ones. `tidyverse` is automatically available so you only need to run `library(tidyverse)` or use the namespace for the function `tibble::glimpse()`.

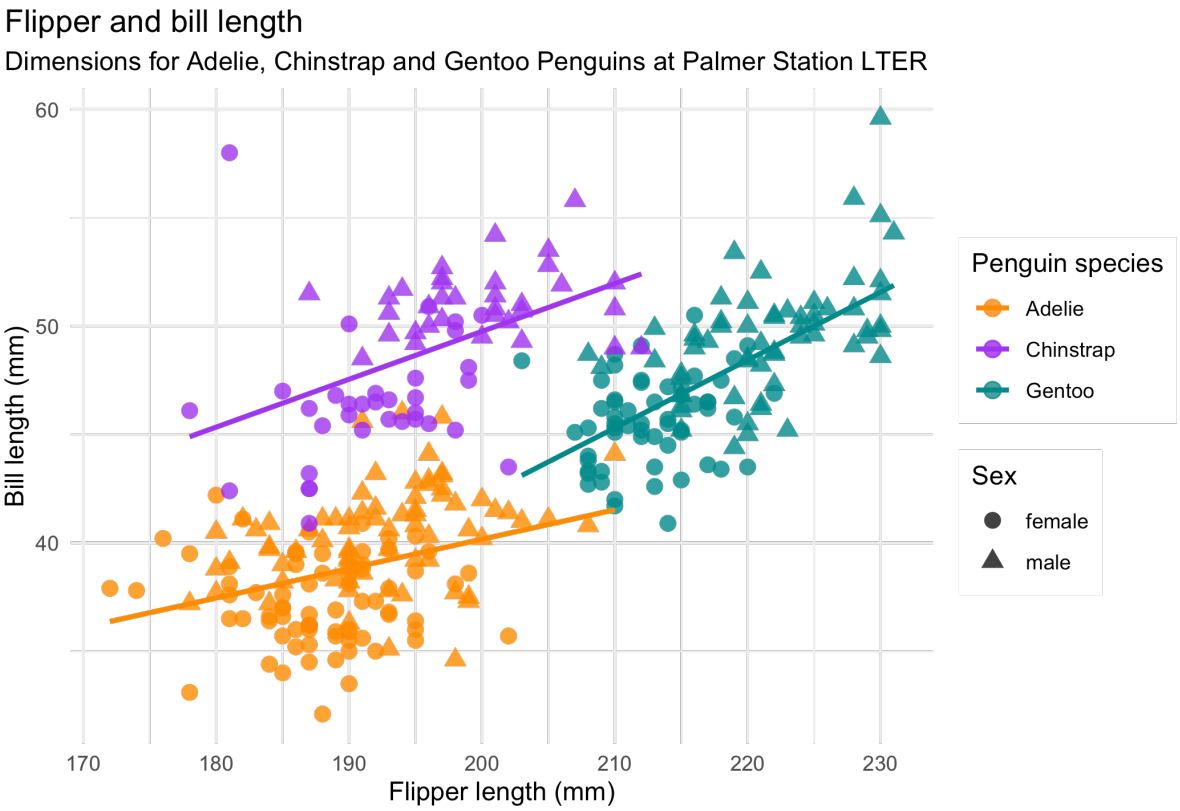
14.5.2 The variables

I hope most of the variables here are clear from their names. In particular, we will be interested in `species`, which has three levels (Adelie, Gentoo, Chinstrap), `sex` (male, female, NA), `body_mass_g` (body mass in grams) and `flipper_length_mm` (length of their flippers measure in mm). We're not using the variables about the bills, but I thought this diagram was interesting anyway.



Artwork by @allison_horst

```
## `geom_smooth()` using formula 'y ~ x'
```



Source: @allison_horst, palmerpenguins package

Here are all our variables of interest in one plot.

Note: In this chart and in future analyses involving `sex`, penguins with an unknown sex have been removed.

14.6 Common statistical tests as linear regression

There is an interactive version of the below notes at this link.

14.6.1 Introduction

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

14.6.1.1 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.

14.6.1.2 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.

14.6.1.3 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.

14.6.2 One-sample t-test

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

14.6.2.1 Use case

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

14.6.2.2 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?)

14.6.2.3 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 likes ours or more extreme.

14.6.2.4 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 you don't have to run any libraries. Use the `t.test()` function run a one-sample t-test.

```
t.test(penguins$body_mass_g, mu = 4000, var.equal = TRUE)
```

```
##
##  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
```

14.6.2.5 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.

```
mod1 <- lm(body_mass_g ~ 1, data=penguins)
summary(mod1)
```

```
##
## 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.

```
mod2 <- lm(body_mass_g-4000 ~ 1, data=penguins)
summary(mod2)
```

```
##
## 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,

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

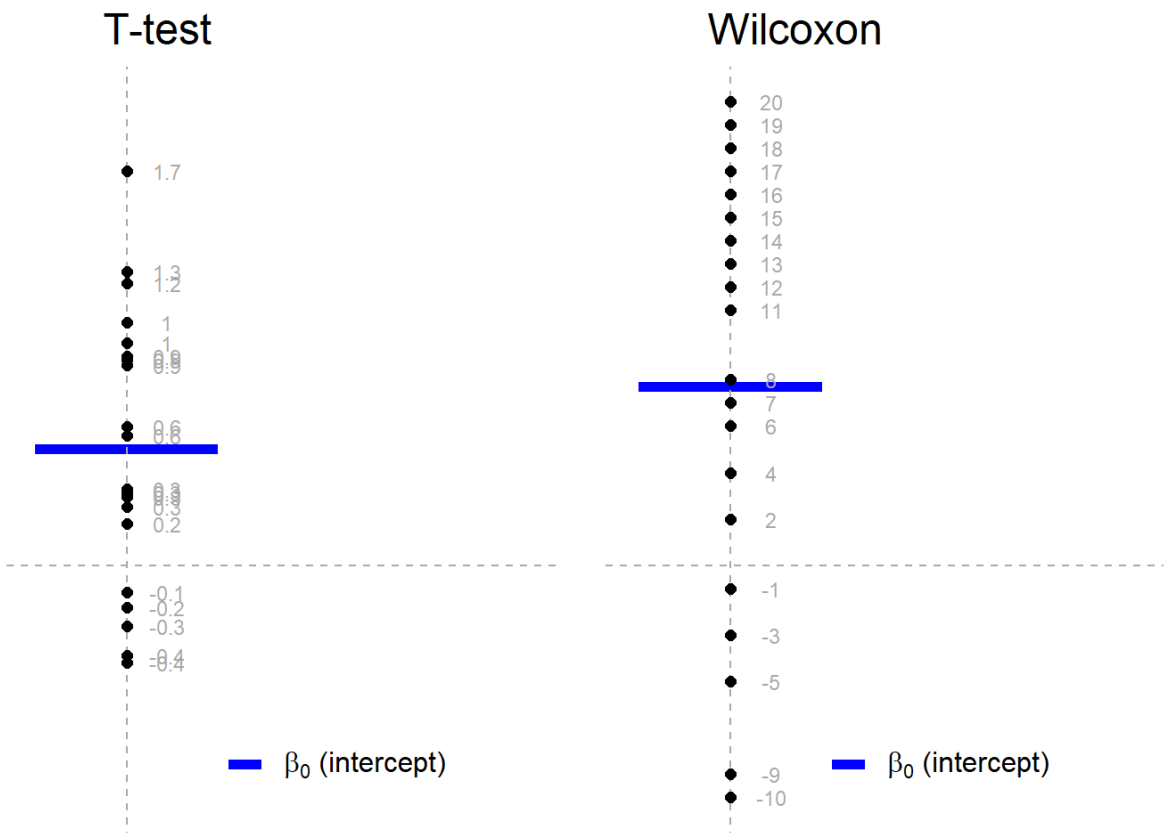
$$H_1 : \mu \neq \text{hypothesized val}$$

are equivalent to our linear regression hypotheses about the intercept,

$$H_0 : \beta_0 = \text{hypothesized val}$$
$$H_1 : \beta_0 \neq \text{hypothesized val.}$$

14.6.2.6 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

14.6.2.7 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))
```

14.6.3 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.

14.6.3.1 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

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

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

14.6.3.2 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*?)

```
mod4 <- lm(body_mass_g ~ flipper_length_mm + species, data=penguins)
summary(mod4)
```

```
##
## 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.

```
head(model.matrix(mod4), n = 20)
```

```
##    (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.

14.6.3.2.0.1 [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 βs, 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.

14.6.4 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.

14.6.4.1 Use case

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

14.6.4.2 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.

14.6.4.3 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 likes ours or more extreme.

14.6.4.4 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 two 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.

```
t.test(body_mass_g ~ sex, data = penguins, var.equal = TRUE)
```

```
##
##  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`.

```
mod5 <- lm(body_mass_g ~ sex, data = penguins)
summary(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 out t-distribution is symmetrical and we're doing a two-tailed test.

14.6.4.5 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
```

14.6.5 ANOVA

14.6.5.1 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…)

14.6.5.2 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://gifst.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.

14.6.5.3 Hypotheses

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

$$H_1 : \text{at least one } \mu \text{ differs from the others}$$

14.6.5.4 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.

14.6.6 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.

14.7 Reproducible examples (reprex)

14.7.1 What is a reproducible example?

Well, it is an example that someone else can reproduce…

Specifically, it is a minimal example needed for someone helping you to reproduce an error or other behaviour you would like to show them with.

This is especially helpful while we are all working online! I can't look over your shoulder in a lab to see what's wrong.

There is a GIF in the web version. Not required content.

14.7.2 Why should you care about creating reproducible examples?

- Useful professional skill for those working with R and other programming languages in graduate school/future employment (concepts transferable to programming in any language).
- The best way to help me (and the TAs and your classmates) help YOU on Piazza.
- The process of preparing a reprex to ask for help forces you to think about your code in a specific way that can help you spot the problem without actually having to ask anyone else (this literally happened to me about 2 hours before writing this slide!)

14.7.2.1 Other things to know

- Please try to use use reprexes when asking for help on Piazza!
  - The outputs from the reprex package work beautifully on Piazza, too.
  - No more upside down photos of computer screens taken from phones, please!

- `reprex` is installed in the JupyterHub and will be loaded whenever we run `library(tidyverse)`
- You won't have to worry about the data side of reprexes as much for this course, as in almost all cases we will all be working with the same data. But! It is a key skill if you will be working with data and programming languages in the future.

14.7.3 Watch the creator of the reprex package explain it

Jenny Bryan is a hero of mine, and this was even BEFORE I knew she was the creator of the `reprex` package.

I think the whole video is excellent, and you're very welcome to watch all of it, but at minimum, please watch the following time intervals:

- 0:00 - 7:20 (background and initial demo)
- 14:08 - 29:48 (philosophy and examples)

The video and slides are on this page: https://reprex.tidyverse.org/articles/articles/learn-reprex.html

14.7.4 Using reprexes on Piazza

Watch this 4 minute video.

Web version embedding for MyMedia is not great, I would recommending opening the link and/or going full screen.

14.8 Reading strategy: previewing and skimming

In graduate school and/or in future employment, it is likely that you will need gather and synthesize information from a range of written sources. To decide if a source is useful to you, and to be as efficient and effective as possible, it is important to have strategies for approaching reading.

In the next module, for the knowledge basket writing task, you will be guided through reading an article called Science isn't broken and asked to respond to it. You can also keep this in mind as a stategry for 'first reads' of course notes, or when re-reading to find some.

Previewing a written text helps you think about the origins and purpose of the text, how it is organised, and where you expect to find information.

Skimming is done after previewing, and is a strategy of reading key paragraphs, like the first and last which usually introduce and conclude the piece, as well as reading the opening sentences of in-between paragraphs and paying attention to figures, captions and any other emphasized text.

14.9 Getting ahead on Module 2

Students who are unfamiliar with the `tidyverse`, especially `dplyr` and `ggplot2`, may find Module 2 to be quite full on. You may wish to start the readings for the next module early in this case.

Wickham & Grolemund. *R for Data Science*. 2017 Chapter 3: Data Visualization and Chapter 5: Data transformation

15

Module 2

Materials for January 24–February 4, 2022.

15.1 Learning checklist

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

- Read csv, xlsx and RDS files in to R.
- Filter, merge, group and summarize data with dplyr functions.
- Create new variables.
- Use pipes %>% to link sequential operations on a dataset.
- Join datasets.
- Evaluate whether a dataset is tidy and alter it if note
- Create appropriate visualizations in ggplot using a range of geometries and aesthetics (scatter plot with geom_smooth, bar chart, histogram).
- Explain common misconceptions about data analysis and statistics in your own words (paraphrasing).
- Recognize when confounding may be an issue.
- Distinguish between an observational study and an experiment.
- Describe a study design which prevents confounding.
- Identify aspects of Informed Consent and recognize gaps in information, comprehension and voluntariness in a given situation.
- Understand the basic principles for web scraping and using APIs.
- Describe and identify selection bias and survivorship bias.
- Demonstrate a general awareness of P-hacking and the reproducibility crisis.
- Describe the idea of 'hypothesizing after knowing result'
- Show why we must consider corrections when performing multiple testing.
- Consider ethical issues in modern data science, including algorithmic transparency, algorithmic bias and what it means when algorithms make the final decision.
- Identify the main components of most statistical reports.
- Implement a previewing and skimming strategy when reading an article.

15.1.1 Key functions

- Visualization
  - `ggplot()`
  - `geom_point`
  - `facet_wrap()`
  - `geom_smooth()`
  - `geom_bar()`
- Transformation
  - `filter()`
  - `select()`
  - `mutate()`
  - `summarise()`
  - `group_by()`
  - `is.na()`
  - `arrange()`
- Loading data into R
  - `readr::read_csv()` for .csv
  - `readxl::read_excel()` for .xlsx
  - `readRDS()`for .RDS
- Viewing data
  - `glimpse()`
  - `head()`
  - `str()`
  - `View()`
- Cleaning up variable names
  - `janitor::clean_names()`
- Introduced in data wrangling video
  - `mutate_if()`
  - `replace_na()`
  - `is.numeric()`
  - `full_join()`
  - `left_join()`
  - `right_join()`
  - `distinct()`
  - `pivot_longer()`
- Working with strings
  - `str_replace()`
  - `str_replace_all()`
  - `str_remove()`
  - `str_c()`
  - `str_detect()`
  - `str_to_sentence()`

15.2 Introduction to Module 2

This module introduces or recaps a range of necessary data wrangling and visualization skills as well as concepts about ethical professional practice for statisticians, and statistical communication information.

With respect to the ethical professional practice components, think of this as just a very high-level survey of these ideas; we aren't going very deep on any of them, but some will set up conversations/tasks later in the course.

- We'll revisit some of these concepts in later classes, like **case-control studies** and logistic regression, **black box algorithms** (very briefly) when we discuss GAMs and issues of confounding, proxies and ethical sourcing of data, throughout.
- We'll touch on **web scraping** in the Team Up! activity for this module.

Folks who have taken or TAed STA130 might be more familiar with some of these concepts than others, but I'm not assuming much prior knowledge.

15.3 Readings for this module

There are several required readings in this module. Keep in mind the previewing and skimming strategies discussed in Module 1.

15.3.1 R for Data Science (+ story time)

Wickham & Grolemund. *R for Data Science*. 2017 Chapter 3: Data Visualization and Chapter 5: Data transformation

Hadley Wickham is a bit of rockstar in the R world.[1] If you took part in the Independent Summer Statistics Community (ISSC), you're probably already well aware of my love for his and Garrett's excellent book, *R for Data Science*. When I brought on a new consultant for my business in New Zealand, he was transitioning from organic chemistry and wanted to become a better data analyst. He worked through this book and in just a few months was leading the analysis for client engagements. It is also the first resource to which I point students who are interested in extra materials for data science.

In the case study section of this module, I'll walk you through a case study using a range of the functions introduced in these chapters and some personal favourites. After these readings and the case study, you should be ready to start your portfolio assessments.

15.3.2 Common misconceptions about data analysis and statistics

Motulsky, H.J., 2014. *Common misconceptions about data analysis and statistics*. https://doi.org/10.1007/s00210-014-1037-6

For the portfolio assessment, you will be asked to write some advice to future you, based on the article *Common misconceptions about data analysis and statistics* by Harvey J. Motulsky.

Focus on what surprised you, what you didn't know before, what you don't want to forget. You can disagree with the article's claims, too.

Direct link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4203998/pdf/210_2014_Article_1037.pdf

[1] My friend literally got him to sign their laptop...

15.3.3 To predict and serve?

Lum, K & Isaac, W. *To predict and serve?* (2016). https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2016.00960.x#sign960-bib-0001

This article is an excellent and accessible introduction to ideas around predictive policing and the current issues. Consider it with the other comments under the Analyzing data section.

(Lum's other work, previously as the Lead Statistician at the Human Rights Data Analysis Group and now at Twitter is awesome!)

There is a GIF in the web version. Not required content.

(If you've seen Minority Report, maybe you were already skeptical about predicting crimes before they happen…but probably for different reasons.)

15.3.4 Science isn't broken

Aschwanden, C. (2015). *Science Isn't Broken: It's just a hell of a lot harder than we give it credit for*. Retrieved from https://fivethirtyeight.com/features/science-isnt-broken

This reading will also be used in the Module 2 writing task. Please read the instructions there before approaching this reading, as it talks you through a practical example of (previewing and skimming)[#previewskim].

15.4 Data wrangling and visualization

15.4.1 Tidy data

15.4.1.1 A bit of advice...and why we need tidy data

> If I had one thing to tell biologists learning bioinformatics, it would be "write code for humans, write data for computers"""Vince Buffalo ((**vsbuffalo?**)(https://twitter.com/vsbuffalo/status/358699162679787521?ref_src=twsrc%5Etfw)), July 20, 2013

This advice applies to everyone learning how to work with data.

The `dplyr` functions we are learning help us **"write code for humans"**, but how do we **"write data for computers"**?

> "It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data" Hadley Wickham, creator of the tidyverse and developer of dplyr

Artwork by @allison_horst

15.4.1.2 Rules of tidy data

There are three interrelated rules which make a dataset tidy:

1. Each **variable** must have its own **column**.
2. Each **observation** must have its own **row**.
3. Each **value** must have its own **cell**.

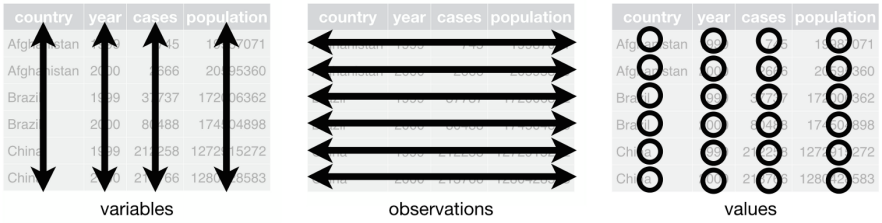


Image from: https://r4ds.had.co.nz/tidy-data.html

15.4.1.3 Recognizing Tidy Data

> "For a given dataset, it is usually easy to figure out what are observations and what are variables, but it is surprisingly difficult to precisely define variables and observations in general." (Wickham, 2014)

A general rule of thumb:

- It is easier to describe functional relationships between variables (e.g., z is a linear combination of x and y, density is the ratio of weight to volume) than between rows.
- It is easier to make comparisons between groups of observations (e.g., average of group A vs. average of group B) than between groups of columns.

(Wickham, 2014)

15.4.1.4 Further reading on tidy data (optional)

While I know many of you will have covered tidy data in STA130, I know others will not be familiar with it. You will need to be able to think about whether or not your data is tidy for several assessments in this course.

Wickham & Grolemund. *R for Data Science*. 2017 Chapter 12: Tidy data

15.4.2 Student grades case study

There are (at least) two ways you could approach this case study:

a) Try it yourself first with the file that has "_guided" in the name and then watch the video.

b) Follow along with the video. I'll be using the file with "_minimal" in the name.

I've tried to design it so either way will work, but I think following along with the video *might* be the best use of your time, if it works for your learning style.

Case study on the Jupyter Hub

If you would rather work locally, there is information on how to download files from the JupyterHub in the Resources section of the course guide.

Note: You'll have to navigate to the Home directory in the Files pane to find this activity in `STA303-w22-activities/m2-casestu`
See the instructions using RStudio with the JupyterHub.

15.4.2.1 Videos

Please note, I filmed these videos last year, so please ignore any assessment/admin comments. The actual R work is all still totally relevant.

- Video 1: Set up
- Video 2: Data wrangling
- Video 3: Data viz

15.5 Ethical professional practice for statisticians

The following sections consider the whys and whats of ethical practice organized under the categories of:

- getting data: confounding and study designs, human research ethics, web scraping and APIs, Indigenous data sovereignty);
- analyzing data: HARKing & multiple testing;
- making decisions with data: algorithmic bias and transparency.

There is a lot of ground covered, though lightly, and it may feel intimidating or overwhelming—there is so much to consider, so much we could get wrong! So, here is some advice from our TA Sonia Markes, who was part of crating this module: *developing the judgment and skills related to these topics takes time and no analysis is ever perfect—we rely on collaboration with other scientists and statisticians who all try to think critically to avoid harm and to improve our collective knowledge and understanding.*

15.5.1 Why should statisticians be ethical?

I hope this questions seems a bit silly to most of you, shouldn't we all strive to be ethical in our professional and personal lives? Well regardless of your personal philosophy, there are additional considerations with respect to our professional ethics.

- **Professional societies:** You may have heard of the Hippocratic oath ("first do no harm") in Medicine and there are similar codes of conduct for Statisticians (see the links to the Canadian and US version below) through our own professional societies.
- **Research ethics boards:** If you are conducting research within universities, it will often have to be approved by a research ethics board. You need to have the skills and knowledge to help appropriately design studies that can be approved under these rules.
- **Legal and business considerations:** Unethical behaviour can also get you are your future employers into trouble. There are of course reputational risks, but there can also be legal risks for creating algorithms that discriminate against protected attributes like gender or ethnicity.

Statistical tools are used to create knowledge. Therefore, anyone using statistical tools is responsible for the knowledge they create through their data collection and analyses.

As ethical statisticians it is important to: - be accurate in our analyses and conclusions - be alert to possible consequences of our results/recommendations on others - be honest in reporting results, even when we don't get the results we hoped for - be respectful of other reasonable results (based on well-conducted research) even if they differ from our own - be mindful of what your data represents, especially if it represents people or their behaviour (a study on people has *subjects*, not objects) - share credit when our work is based on the ideas of others - and more…

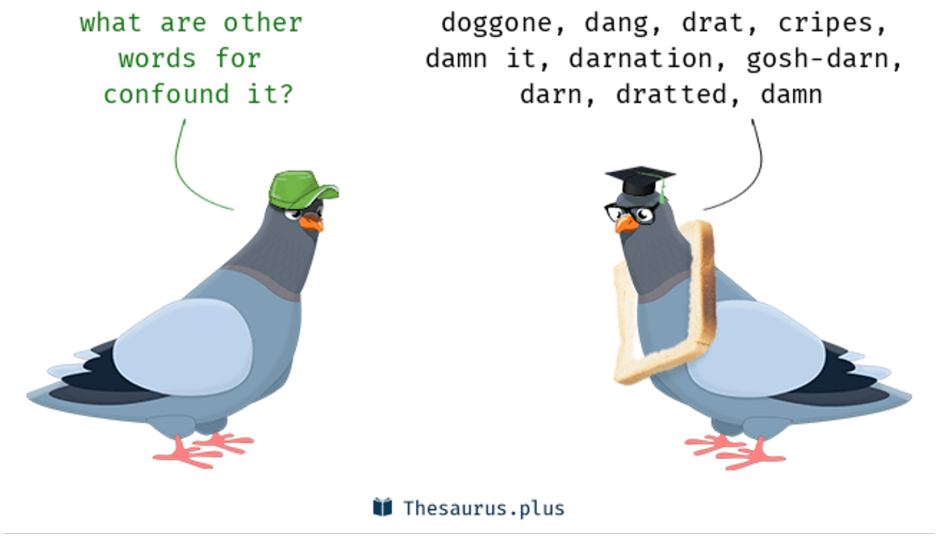
You will be asked to write your own code of ethics in your final project.

15.5.1.1 Relevant codes of conduct for statistical societies/associations in North America

Statistical Society of Canada: https://ssc.ca/sites/default/files/data/Members/public/Accreditation/ethics_e.pdf
American Statistical Association: https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx

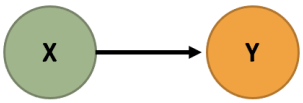
15.5.2 Getting data

15.5.2.1 Confounding and study design

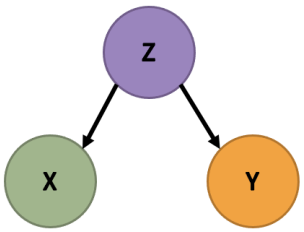


15.5.2.1.1 Confounders (i.e. confounding factors or confounding variables)

Suppose you are interested in the association between a explanatory variable, X, and a response variable Y.



A **confounding variable** (or just 'confounder'), Z, is a variable that influences BOTH the explanatory variable and the response variable.



If we fail to account for our confounding variable, either by not measuring it or not including it, we can come to incorrect conclusions.

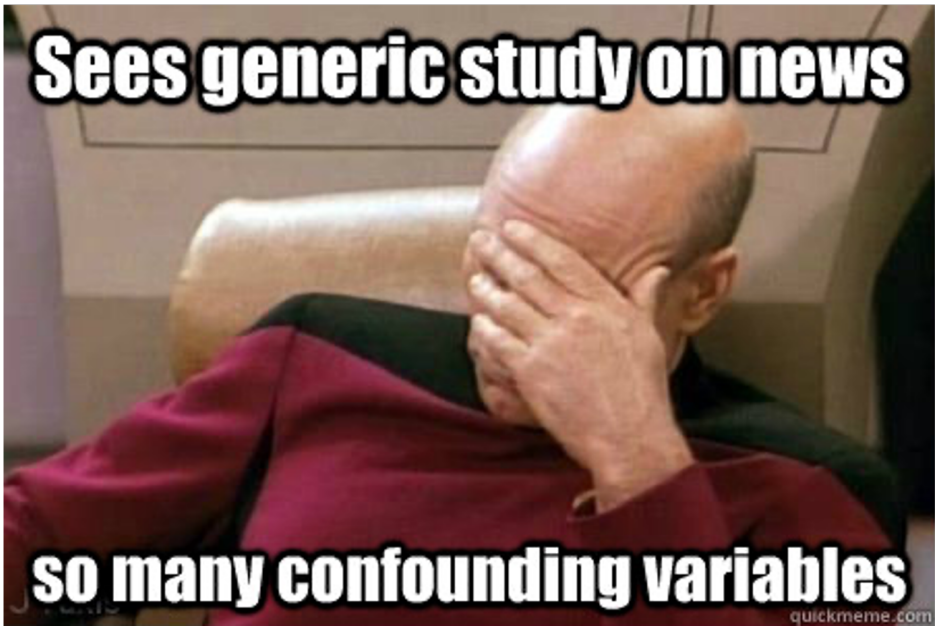
In an **observational study**, variables are "observed" (measured and recorded) without manipulation of variables or conditions by the researcher.

Two variables are **confounded** if their effects on the response variable are mixed together and there is no way to separate them out. If this is the case, we have no way of determining which variable is causing changes to the response.

Example: As ice creams sales rise, so do drownings. But are people drowning because of ice cream? What would be a plausible confounder for this relationship? (Hint:)

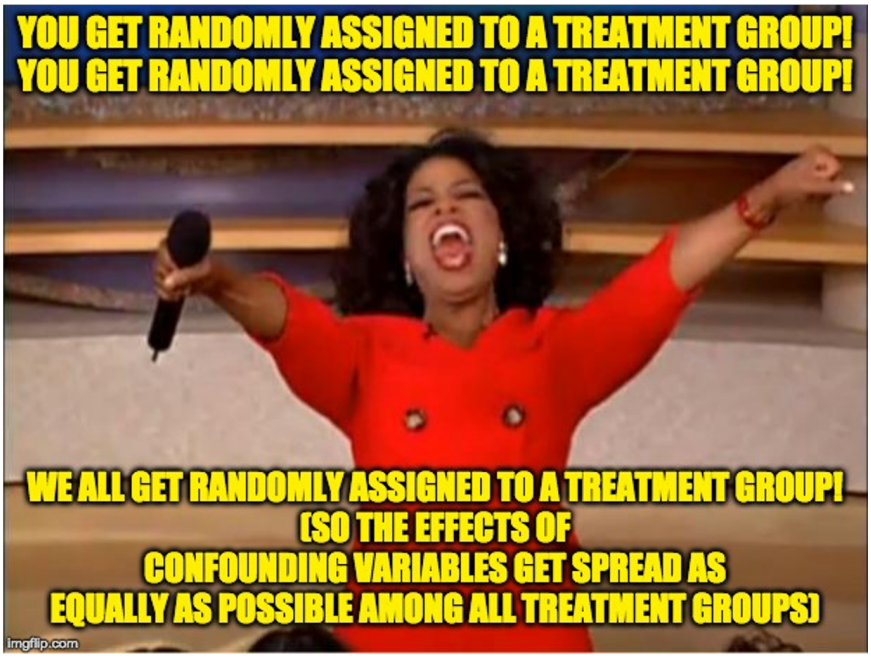
There is no 'test' for confounding. But this is a great thing if you don't want the robots to take your job anytime soon. We need smart people who can think well about confounding.

When we have data from an observational study, we can often only conclude association between variables, not causation. It is worth noting that methods of causal inference are significant area of research for statisticians and others.



So, do we give up on claiming causation?

No!



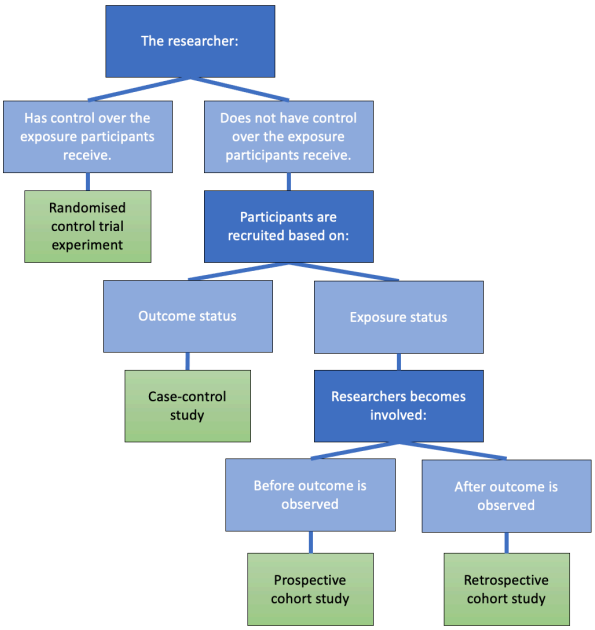
15.5.2.1.2 Designing studies to avoid confounding In an **experiment** (or **randomized trial** or **randomized control trial**) variables and/or conditions are manipulated by the researcher and the impact on other variable(s) is measured and recorded.

The key is to randomly assign some individuals to one treatment (or condition) and randomly assign others to another treatment (sometimes this other treatment is a **control**) *Note: you can have more than two treatments groups too—what is important is that individuals are randomly assigned to them!*

The groups (before treatments are applied) should be very similar to each other with respect to the other variables. Any differences between individuals in the treatment and control groups would just be due to random chance!

If there is a **significant difference** in the **outcome** between the two groups, we may have evidence that there is a **causal relationship** between the treatment and the outcome.

15.5.2.1.3 Common study designs and how to recognize them



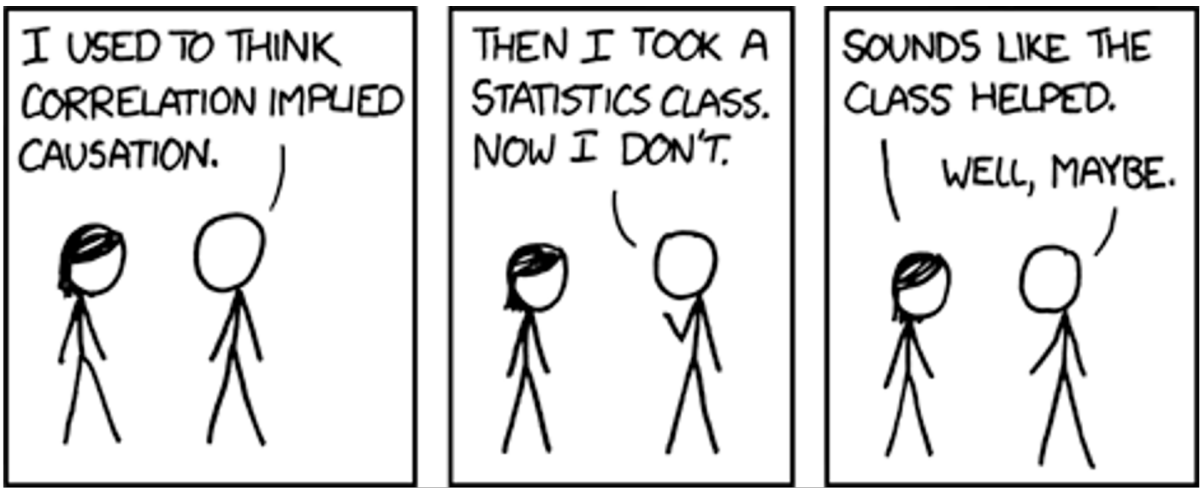
15.5.2.1.4 Causation from observational studies? Although well-designed randomized trials are the gold standard way to establish a causal relationship, observational studies can also help build **evidence** for causation.

15.5.2.1.4.1 Bradford Hill criteria (not assessed)

- Strength of association
- Consistency
- Specificity
- Temporality
- Biological gradient
- Plausibility
- Coherence
- Experiment
- Analogy

> Science is not a magic wand that turns everything it touches to truth. Instead, "science operates as a procedure of uncertainty reduction," said Nosek, of the Center for Open Science. "The goal is to get less wrong over time." This concept is fundamental — whatever we know now is only our best approximation of the truth. We can never presume to have everything right.

From the *Science isn't broken* reading for this module (Aschwanden, 2015).



Source: https://xkcd.com/552/

15.5.2.2 Human research ethics

Ethical codes often emerge out of crisis events.

The Nuremberg code was formulated in August 1947 in Nuremberg, Germany, by American judges[2] sitting in judgment of Nazi doctors accused of conducting murderous and torturous human experiments in concentration camps during the war.

The Nuremberg code codified many of our standard principles of ethical research, including: - research must appropriately balance risk and potential benefits - researchers must be well-versed in their discipline and ground human experiments in animal trials.

15.5.2.2.1 Principles of free and informed consent

15.5.2.2.1.1 Information The research procedure, risks and anticipated benefits, alternative procedures (where therapy is involved), and a statement offering the participant the opportunity to ask questions and to withdraw at any time from the research.

15.5.2.2.1.2 Comprehension The manner and context in which information is conveyed is as important as the information itself. For example, presenting information in a disorganized or rapid manner (with too little time to think about it or ask questions), may limit a participant's ability to make an informed choice.

15.5.2.2.1.3 Voluntariness An agreement to participate in research constitutes a valid consent only if it is voluntary; this requires conditions free of coercion and inappropriate influence.

15.5.2.3 Web scraping and APIs

Web scraping (also known as web harvesting, web crawling or web data extraction) is any method of copying data from a webpage, usually to then store it in a spreadsheet or database.

Downloading a ready-made .csv file hosted by a site wouldn't be considered web scraping. (Although you might find a programmatic way to download many of these could be.)

[2] Did this judgment mean Americans were always getting research ethics right? Definitely not. (Optional) YouTube video about the Tuskegee Study

15.5.2.3.1 What do you need to web scrape?

- Some knowledge of URLs, HTML and CSS
  - URL - Universal Resource Locator
  - HTML - HyperText Markup Language
  - CSS - Cascading Style Sheets
- The `rvest` and `polite` packages in R (or `Rcrawler`, there may be others too) or `Beautiful Soup` for Python
- Professional ethics!

15.5.2.3.2 The Ethical Scraper

I, the web scraper will live by the following principles:

- If you have a public API that provides the data I'm looking for, I'll use it and avoid scraping all together.
- I will always provide a User Agent string that makes my intentions clear and provides a way for you to contact me with questions or concerns.
- I will request data at a reasonable rate. I will strive to never be confused for a DDoS attack.
- I will only save the data I absolutely need from your page. If all I need is OpenGraph meta-data, that's all I'll keep.
- I will respect any content I do keep. I'll never pass it off as my own.
- I will look for ways to return value to you. Maybe I can drive some (real) traffic to your site or credit you in an article or post.
- I will respond in a timely fashion to your outreach and work with you towards a resolution.
- I will scrape for the purpose of creating new value from the data, not to duplicate it.

Source: James Densmore, https://towardsdatascience.com/ethics-in-web-scraping-b96b18136f01

15.5.2.3.3 Terms and Conditions and Robots.txt Many sites give instructions about what you're allowed and not allowed to do on them. One way is through the Terms and Conditions and another is through a file called robots.txt.

T&Cs

Ideally, we should all be reading all the Terms and Conditions of all the websites we use…and of course I'm sure you dooooo.

But when in a hurry, search (CTRL+F or CMD+F) "scrape," "harvest" "crawl" and if none of those come up then "data" and "copied" more generally and that can give you a sense if they prohibit certain uses.

15.5.2.3.4 Robots.txt Most large websites have a robots.txt page to give instructions about what 'robots' are and aren't allowed to visit the page. This is most often used for search engines, but we can check them too. Bad bots can still do what they want.

More on these protocols (and templates you could add to your own site here). http://www.robotstxt.org/robotstxt.html (optional)

An ethical scraper… - …follows the site's terms and conditions and/or robots.txt.
- …uses an API when provided.
- …rate limits their requests.
- I.e., respect a 'crawl limit' suggested by the site, 5 seconds is a polite default if not told otherwise. - …credits their sources.

15.5.2.3.5 Using an API API stands for **a** pplication **p** rogramming **i** nterface.

It is a structured way for data (broadly) requests to be made and fulfilled with computers.

I like this comparison to a restaurant menu. You don't need to know HOW to make crème brûlée to be able to know you WANT it.

If you are using an API, there still may be rules about things like how many requests you can make in a certain time frame and rate limiting. Make sure you're aware of these rules and behave in the spirit of them!

Optional reading (not assessed): https://beanumber.github.io/mdsr2e/ch-ethics.html#sec:terms-of-use

15.5.2.4 Indigenous data sovereignty

Countries and nations tend to want data collected and stored about them/their people to be subject to their laws. You might see examples of this in how government agencies require any cloud storage they use to have the servers be based within their boundaries. This area of thinking is often called **data sovereignty**.

Canada is one of many countries with a history of colonization by settler peoples and the displacement of, discrimination against, and in many cases mass murder of the Indigenous peoples. As countries like Canada go through processes of truth and reconciliation to address these violent and oppressive histories, *indigenous data sovereignty* has also become a growing area of thought.

Why do statisticians needs to know about this? Because we must move from data gathering and analysis as further tools of oppression and be part of honouring the sovereignty of Indigenous peoples and nations over their own data.

1. Be aware of Indigenous rights and interests in relation to data.
2. Understand protocols for consulting with Indigenous peoples about data collection, access and use.
3. Ensure data for and about Indigenous peoples we are given access to is safeguarded and protected.
4. Support quality and integrity of Indigenous data and its collection.
5. Advocate for Indigenous involvement in the governance of data repositories.
6. Support the development of Indigenous data infrastructure and security systems.

Several of these points are generalized from the Māori context in Aotearoa New Zealand. See https://www.temanararaunga.maori.nz/kaupapa.

15.5.2.4.1 How can U of T students develop their knowledge? While there aren't yet specific data sovereignty trainings (we're looking at developing these resources), there are several workshops, events and trainings you have access to.

Go to https://clnx.utoronto.ca and check out the Events & Workshops section.

- St George Calendars: First Nations House - Indigenous Student Services
- Tri-Campus Calendar: Indigenous Cultural Education

Look for workshops with **John Croutch** and/or on the following topics: + Reconciliation: Walking the Path of Indigenous Allyship + Reflecting on Land Acknowledgements + Speaking Our Truths: The Journey Towards Reconciliation | 2 days P1 & P2

As well as being important knowledge for working and living in the Canadian context, as we go forwards there is going to be a competitive *disadvantage* to being ignorant about these issues if seeking work in government, health and social agencies. Many private sector organizations are also making developing these competencies a priority in staff education.

15.5.2.5 Selection bias

Selection bias can occur in a range of ways, but the key feature is that your sample is not representative of the population.

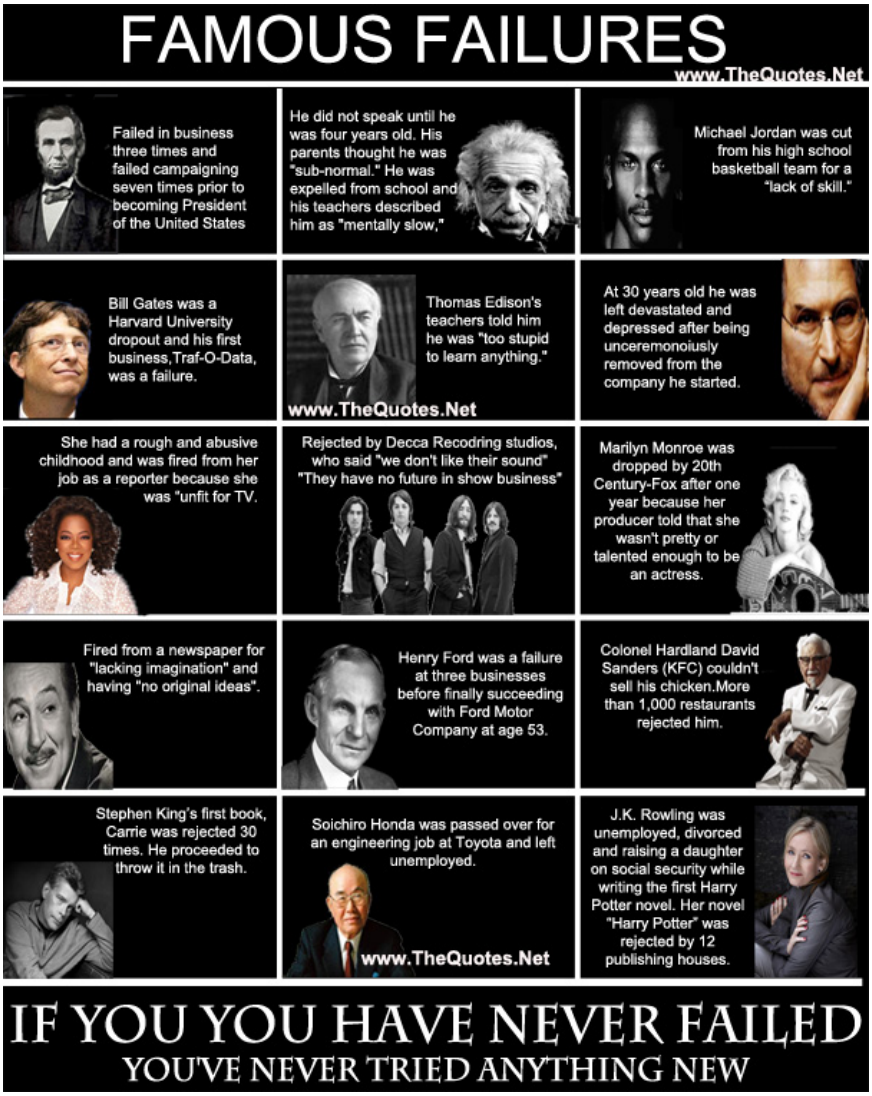
Example: Suppose I want to email out a survey to investigate if U of T students think statistics is important for their future career. I only have the emails for students I teach…in statistics courses. If I randomly sample from this list of students, can I make claims about the population of all U of T students? No.

Example: *The healthy migrant effect.* It has been noticed in many countries that migrants have mortality advantages over local-born populations. While there are several possible things going on and being researched in this area, these findings likely show a component of 'selection bias' in that usually only healthy people can migrate, either due to health screening requirements in the country to which they are migrating, or by people with health complications self-selecting out due to inability/disinclination.

Survivorship bias is a specific type of selection bias.

15.5.2.5.1 Survivorship bias *Note: There is a risk that I'm about to ruin a bunch of 'inspirational' internet content for you.*

If you've spent any time on LinkedIn, and probably lots of other social media sites, you've probably seen an image like this one. Is it inspiring? Sure, maybe…. BUT as an attempted claim about the value of failure it commits the logical error of focusing on just the people who eventually succeeded. I'm sure there are plenty of unemployed, divorced, university dropouts *not* writing *Parry Hotter and the Windows OS*…



Click for source.

15.5.3 Analyzing data

In our readings for this module, the reproducibility crisis, P-hacking and HARKing have been discussed. I won't cover P-hacking or the reproducibilty crisis any further, as these readings are great introductions, but the next few slides talk a little more about **HARKing** and **multiple testing problems**.

- Motulsky, H.J., (2014). *Common misconceptions about data analysis and statistics.* https://doi.org/10.1007/s00210-014-1037-6
- Aschwanden, C. (2015). *Science Isn't Broken: It's just a hell of a lot harder than we give it credit for.* Retrieved from https://fivethirtyeight.com/features/science-isnt-broken

15.5.3.1 HARKing

HARKing is "**H** ypothesizing **A** fter the **R** esults are **K** nown."

I sometimes talk about this as the 'no peeking rule' in setting up hypotheses. For example, you should never pick a one-tailed hypothesis test because of your data, it should only be based on findings from previous studies or a physical theory of a phenomenon.

There is a good introduction to this idea in the Motulsky paper you have been assigned for your portfolio assessment). (Think the XKCD jelly bean comic!)

There are critiques about whether HARKing is as harmful to science as sometimes claimed and a lot of yummy philosophy of science that we won't go into. If you're interested in this area, try reading Rubin (2017), "When Does HARKing Hurt? Identifying When Different Types of Undisclosed Post Hoc Hypothesizing Harm Scientific Progress" (optional).

There is a GIF in the web version. Not required content.

15.5.3.2 Multiple testing problem

One of my favourite examples of positive academic trolling is the dead salmon study. The study used methodology for exploring animal reactions to human emotions expressed in photographs through fMRI scans.

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

If the researchers had ignored the issue of multiple comparisons (there are thousands of areas for which brain activation is measured) they might have ended up with test results that claimed the (dead) salmon was engaging in 'perspective-taking' when shown the photos of the humans. (It was not).

See the poster here: http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf

This is the reason we would want to run an ANOVA when we have more than two groups to compare the means of, instead of doing multiple t-tests between every pair of levels and it is mentioned in conjunction with HARKing in Motulsky paper.

But what is actually going on here?

Suppose you've picked a significance level of $\alpha = 0.05$. When conducting just one test this means we're accepting a 5% risk of making a Type 1 Error, that is, rejecting the null hypothesis when we shouldn't. BUT, if we are conducting several tests at the same time, then we need to think about our **family-wise error rate**, which is our chance of making at least one Type 1 Error across all our tests.

So, if you are making a Type 1 Error 5% of the time, the idea is that 95% of the time you're not. And while 95% is pretty good, if you're doing m independent tests, the not-making-an-error rate becomes 0.95^m, e.g., if you're doing 20 tests, $0.95^{20} = 0.359$, meaning the chance of making at least one 'false discovery' is now $\sim 64\%$

15.5.3.2.1 Correcting for multiple comparisons? There are multiple methods investigators employ in an effort to have their Type 1 Error across multiple tests actually reflect the error rate they are comfortable with.

The simplest but most conservative of these is the Bonferroni correction where you just divide your significance level (e.g., 0.05) by the number of tests you are conducting and use that as the new significance cut off.

E.g. If you'd usually use a 5% threshold and are doing 20 tests, your new threshold is

$$\alpha_{adjusted} = \frac{0.05}{20} = 0.0025.$$

As you can see, this is now a much stronger level of evidence we're requiring against our null hypothesis than when doing a single test.

15.5.4 Making decisions with data

15.5.4.1 Algorithmic bias

Prediction models are taught what they "know" from training data. Training data can be incomplete, biased, or skewed. This can result in **algorithmic bias**.

15.5.4.1.1 Proxy variables There can also be situations where we know we DON'T want to use a variable as part of an algorithm, for ethical and often legal reasons (anti-discrimination laws about gender, race, health status, e.g. American's with Disabilities Act means you can't discriminate against people with mental health conditions). BUT there might be other variables in your data, like certain types of hobbies/memberships, home address, 'personality' quiz questions, that act as 'proxies' for these things, meaning they end up determining outcomes even when you don't want them to.

Optional reading

- Amazon scrapped 'sexist AI' tool, BBC, 2018. https://www.bbc.com/news/technology-45809919
- Amazon discreetly abandoned gender-biased AI-based recruiting tool, HRK News, 2018 https://www.hrkatha.com/recruitment/amazon-discreetly-abandoned-gender-biased-ai-based-recruiting-tool/

15.5.4.2 Should algorithms be transparent?

Some predictive algorithms give us more than just a prediction: they also give us some insight as to what factor(s) influenced the prediction. Examples you might have encountered in your studies already include linear regression models and classification trees.

Other algorithms yield predictions, but no information about how it got from the inputs to the prediction, such as neural networks (you may see these in future courses). These are sometimes called 'black box' algorithms and many machine learning tools fall into this. We'll briefly talk more about this when we discuss generalized additive models later in the course.

What is more important—getting the most accurate predictions, or understanding the factor(s) which influence a prediction?

When you approach a new statistical problem, try to figure out if you are aiming to make **predictions**, make **inferences** about relationships between variables, or provide informative **summaries and descriptions** of the data.

15.6 Statistical communication

This section will be important to review as you work on your final project. You may want to scan it more quickly now and come back in more depth later in the semester.

15.6.1 Prioritize statistical communication as part of your toolbox

While the technical skills of U of T statistics graduates are well received by employers. There is a common theme of communication skills not matching their business needs.

This is part of the thinking behind the emphasis on writing in this course, your professional development task and the project, as well as providing opportunities to 'talk stats' in Team Up! groups.

The section lays out some overarching tips about how to think about your role as a statistical communicator and some general tips about statistical reports, and their contents and structure.

15.6.1.1 What can statisticians learn about communication from Winston Churchill's mother's dinner conversation?

In the lead up to a British election during the Victorian era, Jennie Jerome (Churchill's mum) attended a dinner with the leader of the Conservatives, Benjamin Disraeli, and the leader of the Liberals, William Ewart Gladstone. She said:

> "When I left the dining room after sitting next to Gladstone, I thought he was the cleverest man in England. But when I sat next to Disraeli, I left feeling that I was the cleverest woman."

Who won that election?

Disraeli.[3]

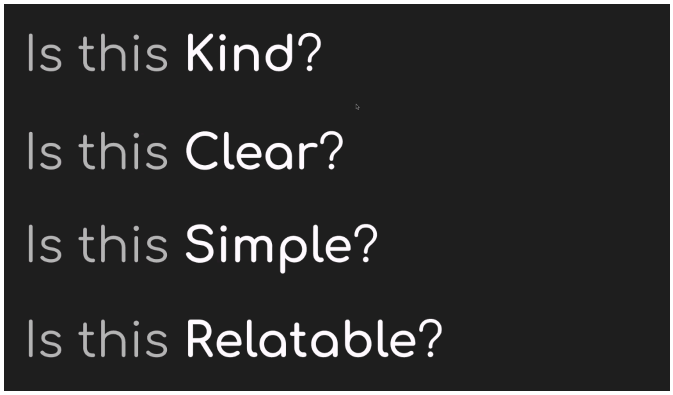
So, what *can* statisticians learn about communication from Winston Churchill's mother's dinner conversation?

I think of this story often when working with clients. Especially early in my career, while I was still building my confidence, there was always the impulse to 'show off' my stats knowledge through jargon in some attempt to show that I was 'worth the money' clients were paying.

And sure, you need to have substance and skill to do the job well. But always think of your job when communicating with non-statisticians as additionally one of **education**.

You want repeat business? Or to impress your non-statistician manager? Then communicate in ways that helps them to 'aha!' moments. **Have them leave you feeling cleverer about statistics than they thought they were coming in.**

Below is a screenshot from a talk Chelsea Parlett-Pelleriti, @ChelseaParlett gave at the Toronto Data Workshop. These are the questions she asks herself when communicating about statistics and ones I think we should all ask ourselves, too.



Parlett-Pelleriti, C. (2020) *Talking to non-statisticians about statistics.* Toronto Data Workshop presentation, September 24.

15.6.2 What do statisticians usually *write*?

Statisticians must often write reports to summarize the results of scientific research or data analyses:

Academia, health sciences, etc.

- Write "methods" section of collaborative papers.

[3] (Optional) Read more here: https://www.businessinsider.com/charismatic-leadership-tips-from-history-2016-10

- Present to research groups.

Industry

- Create slide decks (either stand-alone or for presentation) summarizing results of data analyses / research for use within the company.
- Write reports summarizing results of data analyses / research for use within the company.

15.6.3 Common report components

15.6.3.1 Brief overview (ONE of the following, not both)

15.6.3.1.1 Abstract (or summary) These are common academic articles.

- Short and to the point
- Self-contained (i.e., someone could read this and only this and get the gist of your work)
- Emphasize conclusions rather than specific methods
- Approximately 100–200 words
- No figures, no tables, no/few references.

15.6.3.1.1.1 'Plain English' summaries These are also becoming popular in some academic journals. Think of this as something a high school student could read and understand the purpose of the article. Nice suggestions + example here (optional).]

15.6.3.1.2 Executive summary You might see this more commonly in business or government contexts. It is a summary aimed at busy and **non-technical** 'executives' (for example) and would accompany or preface a more detailed technical report. Your ability to communicate flexibly to different audiences is very important here.

- Short and to the point
- Self-contained
- Use bullet points.
- Emphasize conclusions/recommendations.
- *Might* include the most important figures and tables, very carefully chosen.

15.6.3.2 Introduction

Purpose: provides background information about the project/research. * Why was the investigation done? * What research questions are considered in the report (scope)?

The final paragraph of the introduction should outline what is in the rest of the report. Think of it as the road map.

Important: If you are consulting for a client/collaborator,[4] don't pretend to be an expert if you are not one! The purpose of this section is to provide context (as you understand it) for the rest of your report, so that the report is a complete document, and it is clear what you will and will not address.

[4] If you're interested in statistical consulting, STA490 is a great course (I'm biased), though often limited to students in the Methods and Practice Specialist and with very strict prerequisites

15.6.3.3 Data summary (optional)

Useful if you feel you need more details than you can reasonably give in the Introduction and/or what will be discussed in a Methods section.

Purpose: describing the data, but not answering the research question (which you will put in the "Results" section).

Might include: * Basic summary tables, graphs, and descriptive statistics.
* Changes made to the data to make it suitable for analysis.
* Discussion of decisions around missing data.
* Large tables or series of graphs should go in an *appendix*, not in the main body of the report.

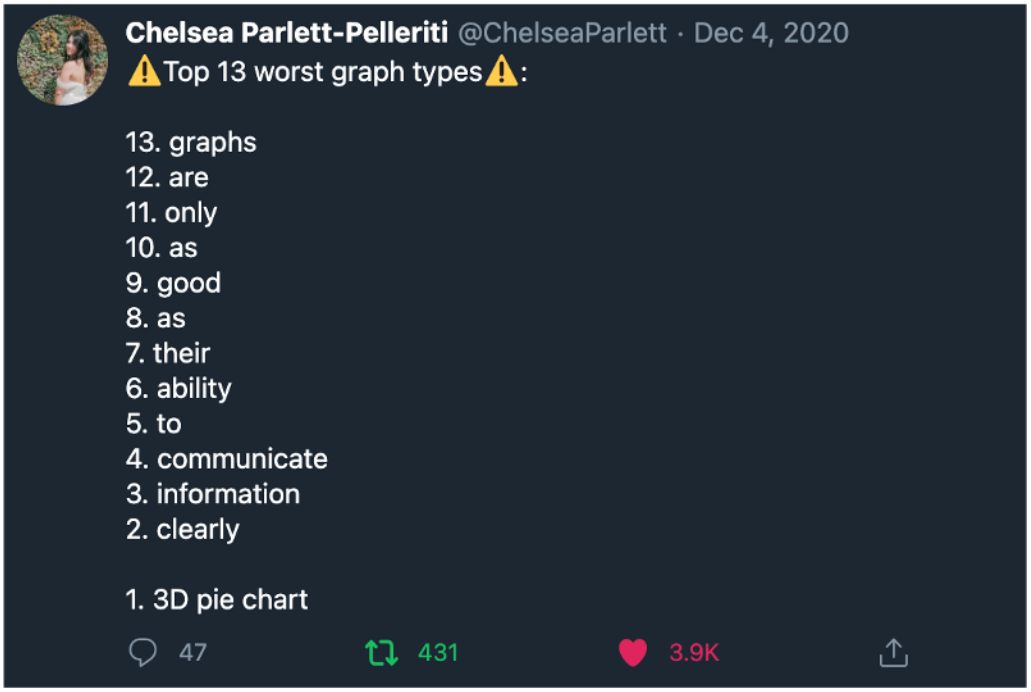
15.6.3.4 Methods

- Briefly describe the design of the scientific study, in the context of the analysis.
  - If you're writing for a collaborator, you don't need to tell the collaborator what *they* did. Use this section for what you did!
  - Resist the temptation to make a diary. This shouldn't be a digest of everything you've done (only the key analyses which you wish to comment on in the results section).
- Do NOT include trivial information (e.g., "The data in Excel were saved in a text file and read into R").
- Be clear which methods were used for what purpose, related to the context of the project. This will likely need to be clarified further in the results section.
- The required level of detail depends on the statistical method.
  - Basic/standard methods (e.g., t-tests, linear regression): do not need to explain.
  - More advanced methods: explain the general motivating ideas in words; you are not writing a statistics textbook!
- A separate methods section is traditional in many scientific journals, but can be difficult (and frustrating) to write if it is too disconnected from the analysis. *Sometimes* it is better to describe the methods and results together.

15.6.3.5 Results

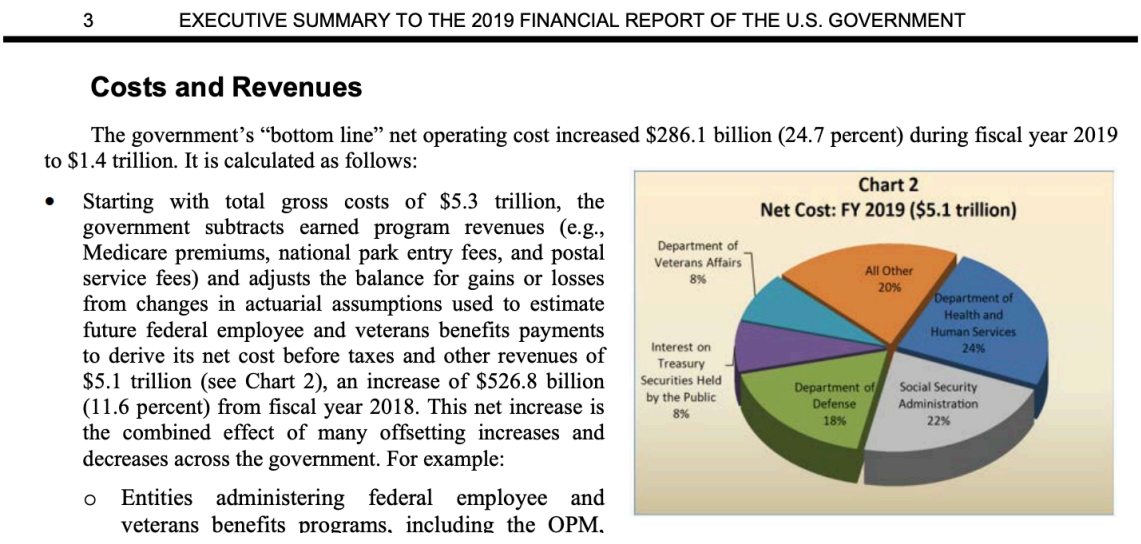
- Make sure this section is clearly organized, point-by-point, in paragraphs.
- Computer output NEVER goes in the body of the report, and RARELY in an appendix, though this depends on your audience/future users.
  - I would encourage you to consider providing code in 'supplementary materials' as this can aid others with reproducing your results and learning from your work. Privacy and sensitivity of data considerations apply.
- Include graphs and small tables that tell the main story; less relevant graphs/tables should go in an appendix.
- If you include a figure/table, you must refer to it in the body of the report (NO lonely figures/tables!). Describe in words what the table/graph shows.
- If you find you are repeating all of the information that is in the graph or table, consider whether or not you need the graph or table. The purpose of figures and tables is to illustrate a pattern, which should be clear for someone looking at it.
- Figures and tables require clear labels and captions (not too small!).
  - General advice/convention is that tables are read from top to bottom, and so **table captions go on top of the table**. Figures on the other hand are read 'bottom to top' and so **figure captions go below the figure**. If your image is not in a report (i.e. an info-graphic like our charts from the data exploration assessment) you might prefer to have a clear title (and subtitle) at the top instead of a caption.
- You may want to group similar graphs into a figure with sub-figures (e.g., Fig 1a, 1b —be sure that the labels/captions are large enough to read though)

- **Describe results in words** and include, parenthetically, the relevant statistics such as means with standard errors (or confidence intervals) or the test statistic with p-value.
  - **Round appropriately!** What is an interesting or useful level of 'granularity' for this study?



Source: https://twitter.com/ChelseaParlett/status/1335009390567456768

Aside: Someone at the US Treasury is my arch nemesis, and they don't even know it.



3 EXECUTIVE SUMMARY TO THE 2019 FINANCIAL REPORT OF THE U.S. GOVERNMENT

Costs and Revenues

The government's "bottom line" net operating cost increased $286.1 billion (24.7 percent) during fiscal year 2019 to $1.4 trillion. It is calculated as follows:

- Starting with total gross costs of $5.3 trillion, the government subtracts earned program revenues (e.g., Medicare premiums, national park entry fees, and postal service fees) and adjusts the balance for gains or losses from changes in actuarial assumptions used to estimate future federal employee and veterans benefits payments to derive its net cost before taxes and other revenues of $5.1 trillion (see Chart 2), an increase of $526.8 billion (11.6 percent) from fiscal year 2018. This net increase is the combined effect of many offsetting increases and decreases across the government. For example:
  - Entities administering federal employee and veterans benefits programs, including the OPM,

Source: https://fiscal.treasury.gov/files/reports-statements/financial-report/2019/executive-summary-2019.pdf

15.6.3.6 Conclusion/discussion

- **Summarize** what was learned regarding the research question.
- Discuss possible **weaknesses/limitations** in the study and/or analysis and specific suggestions for future considerations.

- If appropriate (often done in academic articles) discuss **strengths**/positive points of difference of this study also.
- Explain what the work was able to achieve and not able to achieve.

15.6.3.7 References

- Only include references that have been cited in the text.
- There are many different styles. I often use APA, but don't really have any strong preference. Be consistent and head any instructions relevant to where you are publishing and/or norms in the company/industry/discipline you are working in.

15.6.3.8 Appendices

- Include tables and figures which would **interrupt the flow** of the report if they were included in the main body (e.g., less relevant, but still interesting)
- Keep in mind this is ***not* a dumping ground**. Only include graphs/tables which would be useful/interesting to your audience, and make sure any figures/tables you include there stand alone (e.g., with detailed titles and captions) and are supported by text if necessary.
- If you include an appendix, you should **point to it in the body of the text**, so that the reader knows when it is relevant to flip to it, if they want more information.

15.6.3.9 Final notes on organization

- The overall structure of most articles and reports?
  - *Say what you're going to do and why,*
  - *do it,*
  - *say what you did (and what it means).*
- The **ordering** of ideas should be consistent throughout your report. E.g., if there are multiple research questions in the introduction, they should be addressed in the same order in the methods, results, and conclusion. This is such a simple thing, but easy to forget in the moment.
- You *could* also combine the methods/results sections and **organize your report by the question being answered** (with one subsection for each question).
  - In each subsection, state the methods used and the results. Having separate sections for methods and results makes more sense if the same methods are used repeatedly.
- Many people only read the introduction and conclusion (or at least they start there…) so make these sections are complete and self-contained.

15.6.4 Paraphrasing

This section has been prepared by our Head TA, Amin.

STA303 and Amin's roll is supported by the Faculty of Arts and Science's English Language Learning and Writing teams. If you have particular challenges or questions about writing and communication, let us know! We may be able to direct you to specific resources and/or make STA303 specific ones.

This section will help you with your portfolio assessment and may also be useful as you take notes on the readings for this module.

Paraphrasing is re-writing someone else's statement in your own words. To paraphrase well you need to (1) **provide a reference** (avoid plagiarizing) and (2) **understand** the text and **write** in your own words and create new sentence structures. Recommended resource: https://advice.writing.utoronto.ca/using-sources/paraphrase/.

15.6.4.1 Example activity prompt

Suppose you were asked to paraphrase the below text and provide your conclusions from the observations in the text and the graph in 100 to 200 words. The report is from Statistics Canada (Provencher, Milan, Hallman, & D'Aoust, 2018):

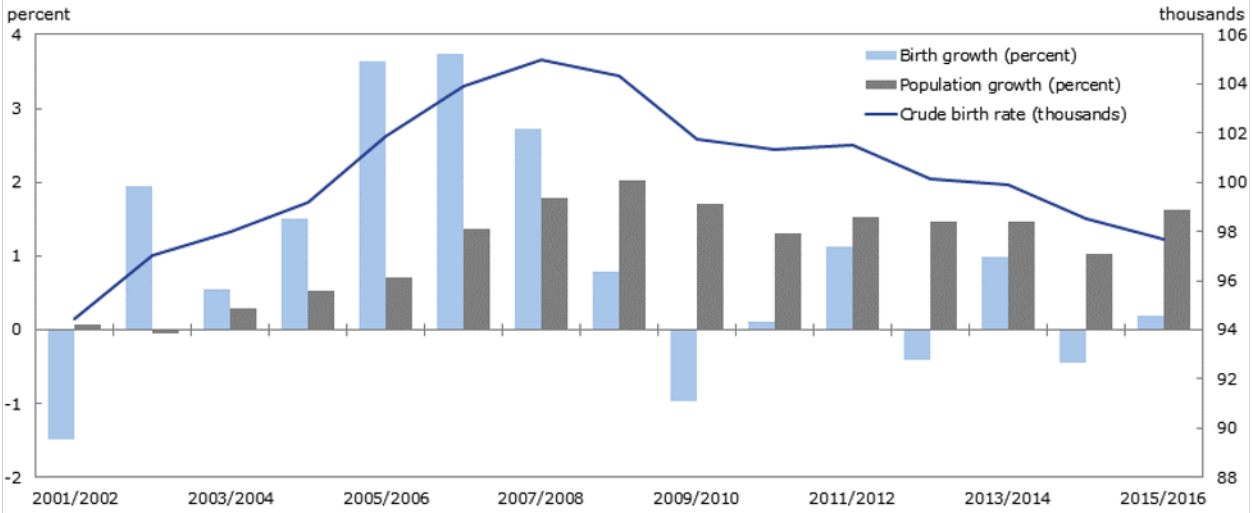
Note: To avoid plagiarism, I have:

(1) placed quotation marks ("") around the verbatim text copied and pasted below, and

(2) provided a reference in the previous sentence (Provencher, … 2018) and the complete citation at the end under "References."]

> "The annual growth in the number of births in the past five years has fluctuated, compared with the period from 2002/2003 to 2008/2009, when it remained positive and stronger (2.1% on average). Between 2011 and 2012 and between 2013 and 2014, the number of births rose around 1.0% across Canada, while they fell 0.4% from 2012 to 2013 and from 2014 to 2015. There was only a small increase of 0.2% between 2015 and 2016. Since 2001, the total number of births fell only four times, in 2001/2002 and 2009/2010, and more recently in 2012/2013 and 2014/2015."

Figure 3
Annual variation in the number of births and in the population of women aged 25 to 34, and the crude birth rate of women aged 25 to 34, Canada, 2001/2002 to 2015/2016



Note: Data used are the most recent available at time of dissemination.
Sources: Statistics Canada, Canadian Vital Statistics, Births Database, 2001 to 2016, Survey 3231 and Demography Division, Demographic Estimates Program (DEP).

Source: Provencher, Milan, Hallman, & D'Aoust, 2018

The following provides guidance on what **poor**, **adequate** and **excellent** paraphrasing could look like. You will be putting this into practice with your portfolio writing task.

15.6.4.1.1 Sample partial/complete answers with overall gradings

15.6.4.1.1.1 Poor

- Copying/pasting verbatim is NOT paraphrasing.
- In the following, there is no change in structure, just a few words replaced.

"The annual growth yearly increase in the number of births in the past last five years has fluctuated varied, compared with in contrast to the period from…"

15.6.4.1.1.2 Adequate

- Changes in wording and structure but text is vague and does not help the reader understand the main points in context of location or time, nor does it provide the source of information.

The number of births has fluctuated over time. There are times when the number is relatively high and sometimes it is relatively low. Not all increases were large with some being as little as 0.2%…

This is adequate in that it is not obviously an academic integrity offence but it would not be particularly useful.

15.6.4.1.1.3 Excellent

- Changes in wording and structure. Provides a good overview.
- Conclusion with appropriate hedging: "…may have been a contributing factor…"
- A shorter version without explaining all the points would have been okay too.
- Stating the reason for observation is unknown and may require further inquiry would have been okay too.

Based on the 2018 report provided by Statistics Canada, yearly birth growth in Canada was initially increasing from 2001/2002 up till 2008/2009 but then slowed down. At its peak, yearly birth growth reached nearly 4% while at its lowest, after 2008/2009, birth growth was even slightly negative at times. The number of births fell 4 times from 2001 to 2016, with 3 out of 4 of these times being in the second half of 2001 – 2016. All these observations point toward a possible change at that time to disrupt the trends midway in 2008. The timing of the lowering of birth numbers coincides with the 2008 financial crisis in Canada and this may have been a contributing factor as families may have not had the financial certainty to have kids. A close examination of contributing factors through research is needed to see if this is true.

15.6.4.2 How to think about paraphrasing?

Based on a method suggested by (Swales & Feak, 2012), identify:

- important points from the text,
- relationships,
- linking phrases,
- connectors, and
- synonyms.

Below is an example of how you might have taken notes on the example paragraph to aid your paraphrasing. You *don't* need to submit anything like this as part of your portfolio assessment. Submit the response to the prompt only. Please be aware that if you do not paraphrase correctly, you may receive a zero on the ENTIRE writing component.

15.6.4.2.1 Example using the Swales & Feak (2012) approach Important points (your notes)

- Number of births has fluctuated in past 5 years.
- The first half of 2001 – 2016 period saw an average increase in births.
- The second half of 2001 – 2016 period saw an average increase in births.
- Number of births fell 4 times from 2001 to 2016, with 3 out of 4 of these times being in the second half of 2001 – 2016.

Relationship between points

- Chronological order (more to less)
- Asymmetry in distribution of events

Linking phrases and expressions to use in paraphrasing

- but
- while

Verbs that might establish other relationships in paraphrasing

- Increase
- Decrease [not used, slowing down used instead]
- Coincides

Possible synonyms to use when paraphrasing

- fluctuated → varied [+between, +around], changed [not used]
- compared with → in contrast to [not used]
- annual→ yearly
- rose / remained positive and stronger → increased

Other considerations

- Who is the audience?

You are ready to re-write the report in your own words!

15.6.4.3 References

Provencher, C., Milan, A., Hallman, S., & D'Aoust, C. (2018). Report on the Demographic Situation in Canada. Fertility: Overview, 2012 to 2016. Statistics Canada. Retrieved from http://www150.statcan.gc.ca/n1/pub//91-209-x/2018001/article/54956-eng.htm

Swales, J., & Feak, C. B. (2012). Academic writing for graduate students : essential tasks and skills (3rd ed.). Ann Arbor: University of Michigan Press.

16

Module 3

Materials for February 7–February 18, 2022.

16.1 Learning checklist

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

- Describe grouping structures for correlated data.
- Discuss partitioning of sources of variability in mixed models.
- Identify situations where linear mixed models may be needed.
- Create plots to explore hierarchical data.
- Write the formula for LMMs.
- Fit LMMs with the `lmer` function using random slopes and random intercepts, as appropriate.
- Interpret the results of `lmer` outputs.
- Conduct appropriate tests to compare LMMs.
- Polish your professional writing.

16.2 Introduction

This module recaps a few topics you should have encountered in previous courses: * statistical independence and correlated data,
* specifically, *from your understanding of the context/collection of data, can you identify whether observations are likely to be independent?*, and
* maximum likelihood/maximum likelihood estimation .

It also introduces an extension to regression, the **linear mixed model**, that uses maximum likelihood estimation to allow us to more flexibly deal with correlated data, that would violate our independence assumptions.

16.2.1 How deep are we going on likelihoods?

We'll need the concept of maximum likelihood for most of the models we'll meet in this course. That said, **proofs about MLEs are not the focus of *this* course.**

- You should have encountered the concept of likelihood in *STA238: Probability, Statistics and Data Analysis II* OR *STA261: Probability and Statistics II*.
  - Having taken one of these is a pre-req for STA302, which is the pre-req for STA303.

- For more depth on this important and useful topic (mathematically and/or computationally) consider taking:
  - *STA355: Theory of Statistical Practice*, and/or
  - *STA410: Statistical Computation*, and/or
  - *STA422: Theory of Statistical Inference* and/or
  - *STA452: Mathematical Statistics I + STA453: Mathematical Statistics II*

These recommendations are based only on the calendar entries for these courses, the latest syllabi of these courses are probably your best guide to what is covered.

16.2.1.1 Maximum likelihood review

$$L(\theta) = f(x|\theta)$$

* Likelihood helps us understand how well our model fits our data. * Maximizing the likelihood function estimates the coefficient values for our model that make *the data we actually observed the most likely.* * That might feel a little backwards, as you could say the chance of observing out data is 100%, because we did…observe it…but hopefully the idea of picking the coefficient estimates for our model that make our data the most likely makes some sense.

16.2.1.2 Properties of maximum likelihood estimators

For large sample sizes:

- Bias goes to 0 (as sample size increases)
- Approximately minimum variance
- Approximately normal distribution (usually)

Another nice feature of MLEs are that they are '**invariant**' under transformation. This means that if $\hat{\theta}$ is an MLE of θ, and g is a 1-1 function, then $g(\hat{\theta})$ is an MLE of $g(\theta)$. This will come in handy in future modules when we want to transform our coefficient estimates with link functions.

16.2.1.3 Suggested reading

As this should mostly be a recap of likelihood, this is a great article on which to practice your previewing and skimming strategies.

You may want to pay special attention to the likelihood ratio test.

Chapter 2 (§ 2.1–2.9) of Roback, P. & Legler, J. Beyond Multiple Linear Regression. (2021). Retrieved from: https://bookdown.org/roback/bookdown-BeyondMLR/ (Links to an external site.)

16.2.2 Model comparison more generally

Thank you to TA Ian for developing this section! He applies statistics in his own research in ecology.

- Model selection is a key component of the development of any statistical model. In model development, there may be multiple plausible hypotheses that can be evaluated simultaneously by parameterizing different models with the same data. These candidate models can differ in both fixed and random effects however the models generally have a similar structure (e.g., multiple combinations of a set of fixed effects). To test the hypotheses with the candidate models, one must determine and compare the support for each model to determine which model(s) fit the observed data the best. While the likelihood-ratio test is suitable when comparing two models to one another, there are certain situations when the goal is to compare multiple (> 2) models to identify the model that best fits the data.

- There are many different criteria for model selection that can be used. Measures such as Akaike Information Criterion (AIC), adjusted R^2, and the Bayesian Information Criterion (BIC) measure the goodness of fit of the model to the data whereas others are focused on calculating the predictive performance of the model (e.g., prediction error rate, root mean square error). Model selection criteria focus on model fit however some will also consider model complexity as model fit generally increases with an increase in the number of model parameters. The selection of the appropriate criteria is generally dependent on the type of model and the purpose of model development (i.e. inference or prediction).

- For this course, we will use likelihood ratio tests as well as the information-theoretic (I-T) approach using AIC for model selection. AIC measures the lack of model fit to the data based on the log-likelihood of the model while penalizing for the number of parameters in the model. Lower AIC values indicate that the model strongly fits the observed data whereas a higher value indicates a poor model fit. **The absolute AIC values can be challenging to interpret** so the values are compared relative to one another. Among the candidate models, identify the lowest AIC value ($AIC_{min}\}$) and calculate the differences between AIC_{min} and the AIC values of every candidate model (AIC_i) to obtain AIC differences (ΔAIC) for each model. In general, large values of ΔAIC (> 10) indicate no support for the model, models with intermediate values (4-7) have little support, and models with small values (0-3) have strong support. These categories are arbitrary and may differ among studies.

- There are scenarios where multiple models have similar support and there is no considerable difference among the model selection criteria. These models would then represent plausible models and support their associated hypotheses. Additionally, model weights can be calculated to evaluate the relative support of each model by the data. Larger weights indicate that the model is strongly supported by the data whereas small values suggest the model poorly fits the data.

16.2.2.0.1 References Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behavioral ecology and sociobiology, 65(1), 23-35.

16.3 Correlated data

16.3.1 Reading

Chapter 7 (§ 7.3 and 7.8) of Roback, P. & Legler, J. Beyond Multiple Linear Regression. (2021). Retrieved from: https://bookdown.org/roback/bookdown-BeyondMLR/

16.3.2 Key vocabulary

- **Fixed effects**: These are non-random quantities. All the coefficients you were estimating in STA302 are examples of fixed effects as you were not treating β_1 as a random variable, for example.
- **Random effects**: These ARE random quantities. These model parameters are treated as random variables.
- **Mixed effects model**: A model that includes both fixed and random effects as its parameters. These are also called hierarchical models and just mixed models. They are NOT the same as *mixed methods*, which is a 'mix' of *quantitative* and *qualitative* research methodology.
- **Nested/nesting design**: Observational units are grouped within grouping units. There may be multiple levels of grouping.
- **Crossed effect design**: Every observational unit experiences every level of the treatment variable.
- **Observational units**: The person or thing on which your outcome of interest is measured. In an experiment we might also call this the 'experimental unit' or some might say 'statistical unit.' In Chapter 7 of Beyond Multiple Linear Regression, these are referred to as 'level-one observational units.'
  - E.g. In a survey of our class, each student is the observational unit. When I am preparing my final grades, each assessment for each student is my unit of observation. This links back to tidy data (week 2). We want each observational unit to have its own row and only that one row.

- **Grouping units**: How your observational units are grouped together. In Chapter 7 of Beyond Multiple Linear Regression, these are referred to as 'level-two observational units,' but you could have even more levels of grouping. Groups within groups, etc.

16.4 Statdew Valley interactive

Click this text to go to the interactive

Note: A 'static' version of this interactive follows, to aid review, but I recommend you work through the interactive the first time.

16.4.1 Welcome to Statdew Valley!

You're grandmother has asked you to take over her old farm plot in *Statdew Valley*. Armed with statistical knowledge, you set out to begin your new life!

There are two of datasets for you to analyze to better understand the operation of your farm.

Note: The final 'hint' for code chunks will be the solution.

Yes, this is totally an homage to Stardew Valley. Have I played? No, but à la Girlfriend Reviews on YouTube, this is based on "what it's like to live with someone who plays Stardew Valley."

16.4.2 Optional: Create your 16 bit character

You can use R for so much more than just running regressions! This section is **totally optional**, and definitely **not assessed.**

1. **Copy this code into a separate .Rmd**. It won't work from within a Shiny App (this interactive).

2. Save the .Rmd and make sure it is in a folder, not in your Home directory. Rmd files in your Home directory will cause interactives to fail.
3. Choose an image to pixelate to get your character. You'll need to have a URL for it (or you could edit the code for a local file). I think it is easiest to upload to GitHub (or Google Photos) and then right click to copy the image address.
4. Play with the Shiny app interface for the pixelart package to help you get the settings right.
5. Update these settings in the code below.
6. [Super super optional] Post your character on the optional discussion board for this week. If you have a Stardew Valley character or some other pixel art of yourself, you can share that, too. This is just for fun, no requirement to participate. Please be sensible about making sure any images are *appropriate.* Don't use images of other people, etc.

```
# Install pixel art
# More info: https://privefl.github.io/pixelart
devtools::install_github("privefl/pixelart")

# Install tidyverse if you haven't
# install.packages("tidyverse")

# Install magick if you haven't
# install.packages("magick")

library(magick)
library(pixelart)
library(tidyverse)

# Upload a profile picture of yourself to GitHub or another service where you can then get the image addre

# Here is an example of mine
url = "https://github.com/sta303-bolton/other/blob/main/lb-for-statdew-valley.png?raw=true"

# Run Shiny app for pixel art models
# This should open in a new browser window
# You can use it to play around with the settings and once you're mostly happy, update the settings below
pixelart::run_pixelart()

url = "https://github.com/sta303-bolton/other/blob/main/lb-for-statdew-valley.png?raw=true"
resize1 = 300
resize2 = 40
ncolors = 15
color_bg = "#FFFFFF"
saturation = 70
degrees = 0
left = 0
top = 0
right = 0
bottom = 0


# I made some changes to the plot_color_matrix function in the pixel art package to produce the look I wan
# Most of this code is the same as from that package. Credit to: Florian Privé, author and creator of the
plot_color_matrix <- function(raster){
    rows <- seq_len(nrow(raster))
    cols <- seq_len(ncol(raster))
```

```
    cbind(expand.grid(y = rev(cols), x = rows), expand.grid(color = t(raster),
        stringsAsFactors = FALSE)) %>% ggplot() + geom_tile(aes_string("x",
        "y", fill = "I(color)")) + coord_equal() + theme_void()
}

im0 <- url %>% magick::image_read() %>% magick::image_background(color_bg) %>%
        crop(left = left, right = right, bottom = bottom, top = top) %>%
        magick::image_rotate(degrees) %>% magick::image_modulate(saturation = 120)

im1 <- downsize(im0, resize1)

# this can take a while to run
kmeans <- kmeans_colors(im1, ncolors)

im2 <- downsize(im0, resize2)

plot_color_matrix(colors_kmeans(im2, kmeans))
```

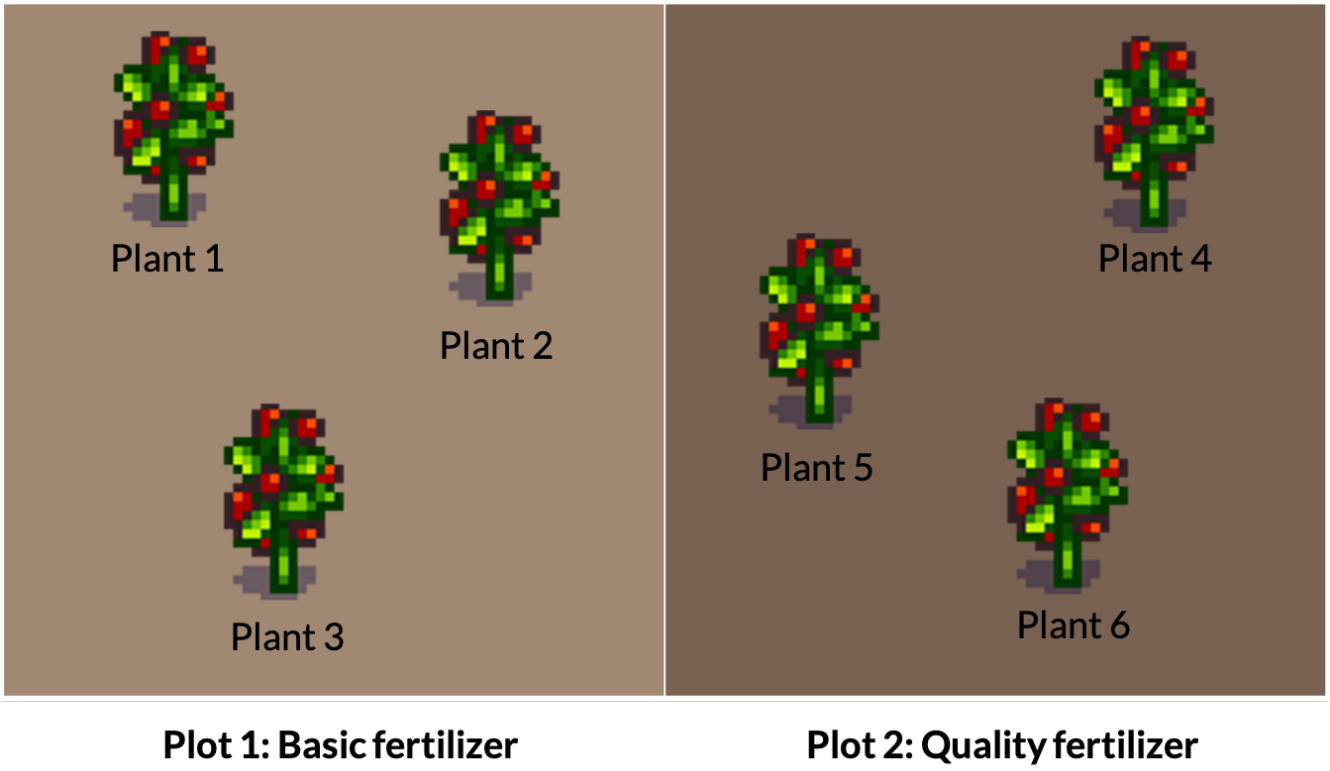
Is this GOOD character pixel art? No. But I think it is cool that you can do image manipulation in R!

16.4.3 Task 1: Tomatoes (part 1)

Suppose your farm has a small experimental tomato patch. You are interested in knowing if there is any difference between the weight of tomatoes produced from plants that are fertilized with basic fertilizer vs tomatoes produced from plants that are fertilized with quality fertilizer.

Tee patch has two plots, one for each fertilizer type. In each plot there are three tomato plants (all of the same species). Other growth influencing factors, like light exposure and precipitation are the same for each plot.

4 tomatoes were chosen at random from each plant and weighed (reported in grams). This data is stored in `tomatoes`.



Layout of tomato patch

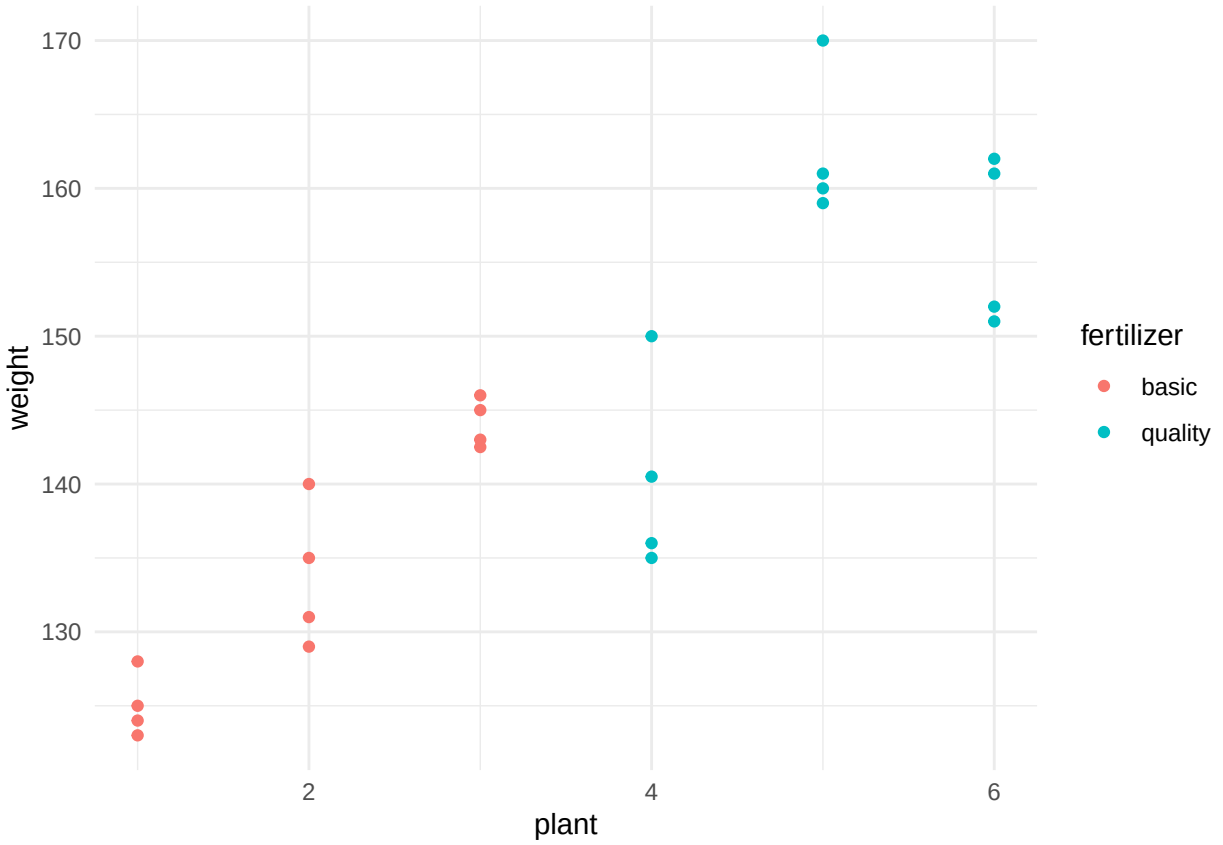
```
glimpse(tomatoes)
```

```
## Rows: 24
## Columns: 3
## $ plant      <int> 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5,~
## $ fertilizer <chr> "basic", "basic", "basic", "basic", "basic", "basic", "basi~
## $ weight     <dbl> 123.0, 124.0, 128.0, 125.0, 135.0, 140.0, 129.0, 131.0, 146~
```

16.4.3.1 Plot the tomat-ah dat-ah

Create a plot with `plant` on the x axis, `weight` on the y axis and coloured by `fertilizer` type. Choose a sensible geometry. I also used `theme_minimal()` because I like the way it looks. This is just an exploratory plot for our benefit, so it can be a bit rough. No title, etc.

```
tomatoes %>%
  ggplot(aes(x = plant, y = weight, colour = fertilizer)) +
  geom_point() +
  theme_minimal()
```



Note: We're going to take a pretty STA302 approach here. We'll start out by doing something WRONG and then improve it in the next section, still under the least squares framework. We can do this BECAUSE this is what is called a 'balanced data' design. That is, there is an equal number of observations in each combination of levels: 4 tomatoes per plant per fertilizer type plot.

16.4.3.2 The WRONG approach: fixed effects only

Fertilizer is our variable of interest, so we need to include it in the model, but we might expect that tomatoes from the same plant are more similar to each other in weight than tomatoes from different plants. This also seems believable based on the plot. So, let's start by proposing a model like this:

16.4.3.2.1 Model

$$y_i = \alpha_j + \beta_k + \epsilon_i$$

where y_i is the weight of the i^{th} tomato from plant k in fertilizer plot j. Here, α_j is the population mean weight for fertilizer plot j and β_k is the difference of plant k from that mean.

The two longer horizontal lines in this graph represent α_{basic} and $\alpha_{quality}$ respectively. Each plant also has a shorter horizontal line shown which represents the mean weight of tomatoes from that plant. The distances (shown with the 6 vertical lines) are the β values discussed above. (You're not required to be able to create this plot right now. If you're curious you can see the code in the .Rmd version of this activity.)

Now, this probably looks a little different to how you've been writing models, but I hope it makes sense as describing the data that we're interested in.

Question: Could we use this model to predict the average weight of tomatoes from a new plant (e.g. plant 7) grown in the quality fertilizer?

Answer: No We have no coefficient for plant 7, so no, this model as it is currently won't help us make predictions for any plants other than the ones we've observed."

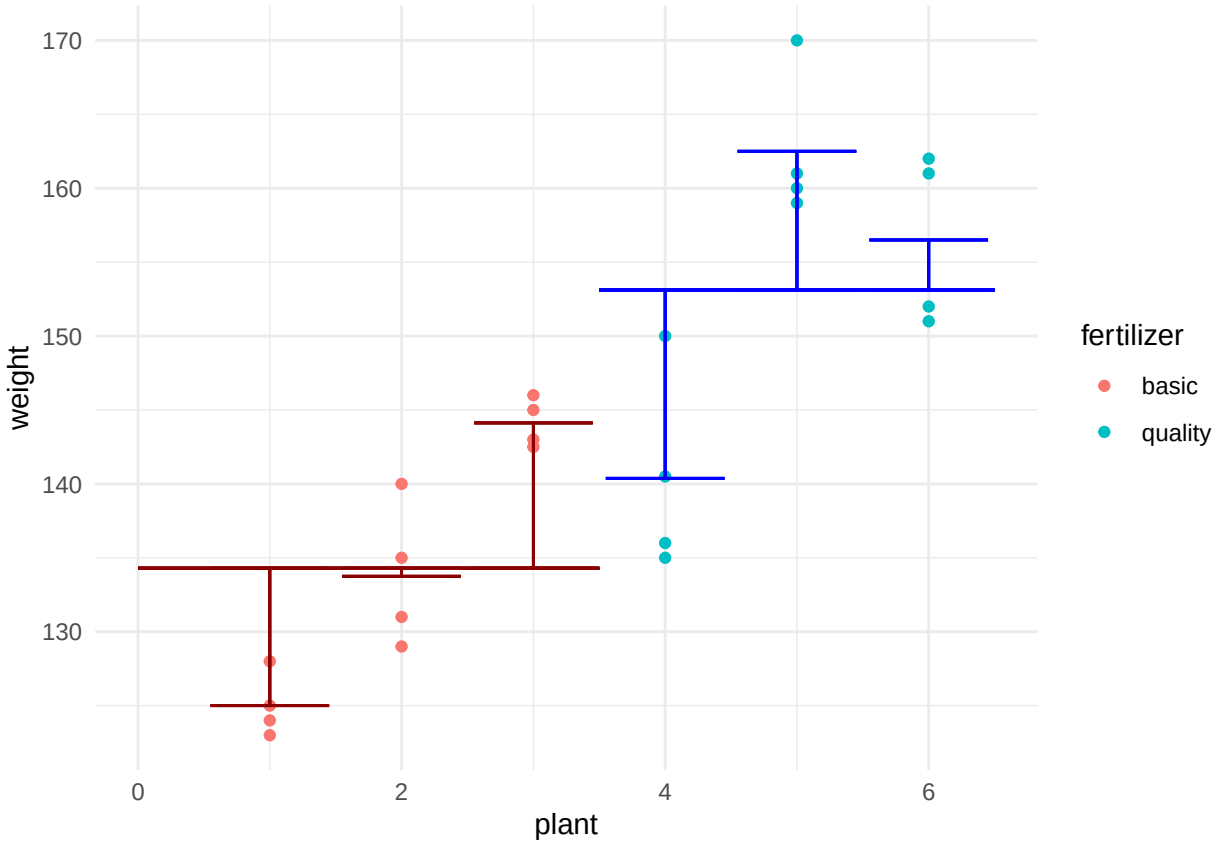


Figure 16.1: *Figure: Visual representation of the equation for this data*

Question: Is this model identifiable as currently set out? That is, is there only one possible set of values that we could set these parameters to to get the correct response value?

Answer: No. The β_k s are completely nested within our fertilizer levels, with three plants in the 'basic' fertilizer plot and three in the 'quality fertilizer plot.' Why is the a problem? In trying to fit these values, you'll end up with an infinite number of possible values, because you can add any value to α_{basic} and subtract the same value from β_1, β_2 and β_3 to get the same value of y. (Likewise for the 'quality' fertilizer plot and associated plants). Our model conceptualized in this way will not be identifiable."

16.4.3.3 Let's try some models: Part 1

1. Fit a model called `tom1` with `weight` as the outcome and `fertilizer` and `plant` as the predictors. Make sure `plant` is treated as a categorical (factor) variable, not numeric.
2. Fit a model called `tom2` with `weight` as the outcome and and just `fertilizer` as the predictor.
3. Compare these models using ANOVA.

```
tom1 <- lm(weight ~ as_factor(plant) + fertilizer, data = tomatoes)
tom2 <- lm(weight ~ fertilizer, data = tomatoes)
anova(tom1, tom2)
```

```
## Analysis of Variance Table
##
## Model 1: weight ~ as_factor(plant) + fertilizer
## Model 2: weight ~ fertilizer
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     18  411.62
## 2     22 2192.29 -4   -1780.7 19.467 2.41e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
result_tom1 <- anova(lm(weight ~ as_factor(plant) + fertilizer, data = tomatoes),  lm(weight ~ fertilizer,

pval <- sprintf("%.2e", result_tom1$`Pr(>F)`[2])
```

Question: Is there any evidence against the hypothesis that a model with fertilizer alone fits the data as well as the model with both plant and fertilizer?

A. "Yes, we have evidence against the fertilizer only model being just as good.
B."No, we can proceed with the fertilizer only model.

Answer: The p-value from our ANOVA above, 2.41e-06, suggests that there is strong evidence against there being no plant effect. Which means that we can't separate whether the fertilizer is having an effect or not, with this model.

To emphasize the problem further, let's fit another model

1. Fit `tom3`, that predicts `weight` with `plant` (as a factor) as the only independent variable.
2. Run an anova to compare `tom1` and `tom3`

```
tom1 <- lm(weight ~ as_factor(plant) + fertilizer, data = tomatoes)
tom3 <- lm(weight ~ as_factor(plant), data = tomatoes)
anova(tom1, tom3)
```

```
## Analysis of Variance Table
##
## Model 1: weight ~ as_factor(plant) + fertilizer
## Model 2: weight ~ as_factor(plant)
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1     18 411.62
## 2     18 411.62  0         0
```

Question: Think carefully about `tom1` and `tom3`. Which ONE of the following statements is TRUE based on the results of your ANOVA, above?

A. We have very strong evidence against the claim that the plant only model fits the data as well as the model with plant and fertilizer.
B. We have no evidence against the claim that the plant only model fits the data as well as the model with plant and fertilizer, and so it appears there isn't a difference between the fertilizers. C. Because of the nesting of plant within fertilizer, these models are actually really the SAME model and comparing them tells us nothing about the fertilizer.

Answer: C. These models actually are the SAME in terms of the information they have, because of the total nesting of plants within fertilizer plots. You'll notice the residual degrees of freedom are the same. So it is true that comparing them tells us nothing about fertilizer effects.

16.4.3.4 Summary

Approaching our model in this way has several problems.

- Our model isn't **identifiable.** In fact, if you run `summary(tom1)`, you'll see that the fertilizer coefficient is NA because of this! Oops.
- What we can model isn't that useful to us, because it doesn't **generalize** to any new plants. (E.g., we can't say anything about a plant 7.)
- If we just fit the model is fertilizer, we know our **independence assumption** is violated and don't know how much of a problem that will be for us. (Violated because we know observational units, tomatoes, are not independent when they come from the same plant.)

So, do we give up on our first farming task?

Nope!

16.4.4 Task 1: Tomatoes (part 2)

16.4.4.1 The RIGHT approach: mixed effects

To approach this problem better, let us take a step back and realize that we really want to treat `fertilizer` and `plant` ***differently***. Fertilizer is a classic *fixed effect* kind of variable. We want to be able to treat the parameter as fixed and we want to be able to use it to generalize so we can make claims about whether or not there is a difference between the fertilizers. Meanwhile, we're not *really* that interested in the effects of each of these plants. We want information that will help us make decisions for future plants. What if we plant 20 new tomato plants? We want to be able make data driven claims in that case, too, not just for these 6 plants. BUT we also can't ignore the plant effects (independence assumption, as discussed).

Our solution? **Treat plant effect as a random variable**. We can think of these 6 plants as drawn from the population of all tomato plants (of this species, in these conditions). We can then understand each of their effects as an independent observation from this distribution.

Decision: We will model `fertilizer` as a fixed effect and `plant` as a random effect.

16.4.4.1.1 Model This looks like:

$$y_i = \alpha_j + b_k + \epsilon_i$$

where, again, y_i is the weight of the i^{th} tomato from plant k in fertilizer plot j. Here, α_j is again the population mean weight for fertilizer plot j BUT $b_k \sim N(0, \sigma_b^2)$ and $\epsilon_i \sim N(0, \sigma^2)$. b_k and ϵ_i are mutually independent.

Because we're in 'statistical easy mode' (balanced design) with this example, we can actually achieve all of this in the framework of ordinary least squares you learned in STA302. FIRST we have to do some data manipulation.

Question: Can you guess what we need to do to our data?

A. Calculate average weight by fertilizer type and refit with formula `weight_avg ~ plant`
B. Calculate average weight by plant and refit with formula `weight_avg ~ fertilizer`
C. Calculate average weight by plant and fertilizer and refit with formula `weight_avg ~ 1`

Answer: We want to aggregate the data such that we have one value for each plant, the average weight of tomatoes from that plant. Then we will fit the model with the formula `weight_avg ~ fertilizer`.

1. Create a new dataset called `tom_agg` from the `tomatoes` dataset.
2. Group by both fertilizer and plant. Because plant is fully nested in fertilizer, this is just a convenient way to keep the fertilizer variable in the dataset.
3. Summarize so you get the mean tomato weight for each plant. Call this variable `weight_avg`.
4. Run `glimpse(tom_agg)`.

Before you run the code, see if you can predict what the output should look based on these instructions.

Question: What will the new `tom_agg` data set look like?

A. There will be one row for each plant, so 6 rows total. Three columns, for fertilizer, plant and weight_avg.", correct = TRUE), B. There will be one row for each plant and fertilizer combination, so 2*6 = 12 rows total. Three columns, for fertilizer, plant and weight_avg."), C. There will still be 24 rows and the original three columns, but now there will be an additional `weight_avg` column."), D. We can't predict how many rows there will be, but there should be 3 columns.

Answer: Run the code and see! (or see below)

```
tom_agg <- tomatoes %>%
  group_by(fertilizer, plant) %>%
  summarize(weight_avg = mean(weight), .groups = "drop")

glimpse(tom_agg)
```

```
## Rows: 6
## Columns: 3
## $ fertilizer <chr> "basic", "basic", "basic", "quality", "quality", "quality"
## $ plant      <int> 1, 2, 3, 4, 5, 6
## $ weight_avg <dbl> 125.000, 133.750, 144.125, 140.375, 162.500, 156.500
```

Now that you've created the new appropriately aggregated dataset, `tom_agg`, use it to fit a model called `tom_final` that predicts `weight_avg` from `fertilizer`. Then run `summary()` on it. Calculate a 95% confidence interval for the effect of using quality fertilizer over basic fertilizer.

```
tom_final <- lm(weight_avg ~ fertilizer, data = tom_agg)
summary(tom_final)
```

```
##
## Call:
## lm(formula = weight_avg ~ fertilizer, data = tom_agg)
##
## Residuals:
##        1        2        3        4        5        6
##  -9.2917  -0.5417   9.8333 -12.7500   9.3750   3.3750
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        134.292      6.091  22.048  2.5e-05 ***
## fertilizerquality   18.833      8.614   2.186   0.0941 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.55 on 4 degrees of freedom
## Multiple R-squared:  0.5445, Adjusted R-squared:  0.4306
## F-statistic: 4.781 on 1 and 4 DF,  p-value: 0.09406
```

```
confint(tom_final)
```

```
##                        2.5 %    97.5 %
## (Intercept)       117.381037 151.20230
## fertilizerquality  -5.081909  42.74858
```

Question: Which one of the following statements about our results is correct?

A. We have weak evidence against the claim that fertilizer type has no effect on tomato weight.
B. It is plausible, at the 95% confidence level, that on average, quality fertilizer produces tomatoes that are 10 grams heavier than tomatoes grown in basic fertilizer.
C. At the 5% significance level we fail to reject the hypothesis that the average weight of tomatoes grown each fertilizer type is the same.
D. Using a 5% significance level here, there is a chance that we are making a type II error.
E. More than one of the above.

Answer: All of these are actually correct! A) If using a 'strength' approach instead of a reject/fail to reject one, we often say a p-value between 0.1 and 0.05 is 'weak evidence.' B) 10 is in our 95% confidence interval, so it is a 'plausible' or believable value, because a CI is a range of plausible values for our parameter. C) Our p-value is larger than 0.05, so we fail to reject it at the 5% level. D) Because we are failing to reject the null, there is a chance that we're wrong, and we SHOULD have rejected the null. Failing to reject the null when you should is a type II error. I remember this with 'fail **2** reject -> type **2**'

16.4.4.2 Variance of our random effect

When we are conducting mixed effects models, we're often interested in the variability of our random effects. So, we want to estimate σ_b^2 from the model above.

Firstly, we can think about what variability that isn't explained by our fixed effects. This is our *residual* variance. We get this from squaring the residual standard error in `tom_final`. This value is actually partitioned into two independent sources of variability, $\hat{\sigma}^2$, the variance of our random error, and $\hat{\sigma_b}^2$, the variance of our random effect.

While you can derive this yourself if you like, let's also take as true the following:

$$\frac{\text{RSS}_{tom\_final}}{4} = \hat{\sigma_b}^2 + \frac{\hat{\sigma}^2}{4}$$

.

The following just reruns the final code from the tasks above so I can use it.

```
tom_agg <- tomatoes %>%
  group_by(fertilizer, plant) %>%
  summarize(weight_avg = mean(weight), .groups = "drop")

tom_final <- lm(weight_avg ~ fertilizer, data = tom_agg)
summary(tom_final)
```

```
##
## Call:
## lm(formula = weight_avg ~ fertilizer, data = tom_agg)
##
## Residuals:
##        1        2        3        4        5        6
##  -9.2917  -0.5417   9.8333 -12.7500   9.3750   3.3750
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        134.292      6.091  22.048  2.5e-05 ***
## fertilizerquality   18.833      8.614   2.186   0.0941 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.55 on 4 degrees of freedom
## Multiple R-squared:  0.5445, Adjusted R-squared:  0.4306
## F-statistic: 4.781 on 1 and 4 DF,  p-value: 0.09406
```

```
# Get the RSE and square it
overall_var <- summary(tom_final)$sigma^2
overall_var
```

```
## [1] 111.2917
```

Secondly, we actually have an estimate of $\hat{\sigma}^2$ from the model we fit right at the beginning, `tom1`. This is the variance we can't account for even after fitting both `plant` and `fertilizer` as fixed effects.

$\hat{\sigma}^2 = \frac{\text{RSS}_{tom1}}{\text{DF}_{tom1}}$

(This is just from our usual linear model theory. $\hat{\sigma} = \text{RSE} = \sqrt{\frac{\text{RSS}}{DF}}$, so $\hat{\sigma}^2 = \text{RSE}^2 = \frac{\text{RSS}}{DF}$)

```
tom1 <- lm(weight ~ as.factor(plant) + fertilizer, data = tomatoes)
summary(tom1)
```

```
##
## Call:
## lm(formula = weight ~ as.factor(plant) + fertilizer, data = tomatoes)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.500 -2.938 -1.062  2.156  9.625
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)        125.000      2.391  52.279  < 2e-16 ***
## as.factor(plant)2    8.750      3.381   2.588  0.01857 *
## as.factor(plant)3   19.125      3.381   5.656 2.30e-05 ***
## as.factor(plant)4   15.375      3.381   4.547  0.00025 ***
## as.factor(plant)5   37.500      3.381  11.090 1.78e-09 ***
## as.factor(plant)6   31.500      3.381   9.316 2.63e-08 ***
## fertilizerquality       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.782 on 18 degrees of freedom
## Multiple R-squared:  0.9047, Adjusted R-squared:  0.8783
## F-statistic: 34.19 on 5 and 18 DF,  p-value: 1.395e-08
```

```
# Get the RSE and square it
error_var <- summary(tom1)$sigma^2
error_var
```

```
## [1] 22.86806
```

The unexplained variability in our final model can be partitioned into the plant-to-plant variability and the random error variability. The random error variance is divided by the degrees of freedom of the final model.

$$\text{RSE}^2_{tom\_final} = \frac{\text{RSS}_{tom\_final}}{4} = \hat{\sigma_b}^2 + \frac{\hat{\sigma}^2}{4}$$

$$= \hat{\sigma_b}^2 + \frac{\text{RSS}_{tom1}}{18} \cdot \frac{1}{4}$$

$$\hat{\sigma_b}^2 = \text{RSE}^2_{tom\_final} - \text{RSE}^2_{tom1} \cdot \frac{1}{4}$$

```
# Let's calculate the variance of the plant effect
plant_effect_var <- overall_var - error_var/4
plant_effect_var
```

```
## [1] 105.5747
```

So, $\hat{b} \sim N(0,\ 105.57\ )$. As you might expect, a lot of the variance in tomato weight that we can't account for just based on the fertilizer is due to plant-to-plant variation

16.4.5 Task 2: Life is sweet as honey

Your grandmother was a bit eccentric when it came to farm management. She swore that her bees produced more honey if she played them music in the morning. In fact, she even mentioned that she thought they particularly liked K-pop.

She left you some data that she thinks proves this. While the sample size isn't that large, you decide to take a look anyways. The data is called `honey` and records the honey produced each month for each of her 6 hives (in grams). Each hive listened to each of the song options (Butter by BTS, Adagio in B minor by Mozart and No music) for a month, three times. The order was randomized for each hive. It took 9 months to collect this data and the information your grandmother left you says honey production should have been fairly constant over this period and that all the hives experienced the same weather conditions, food access, etc.

(Note: this is definitely not accurate to how beekeeping or weather actually works…but ignore this for our purposes.)

```
glimpse(honey)
```

```
## Rows: 54
## Columns: 3
## $ hive    <fct> 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 1, 1, 1,~
## $ song    <fct> No music, No music, No music, No music, No music, No music, No~
## $ honey_g <dbl> 196.56, 199.58, 200.71, 195.80, 199.58, 200.71, 226.80, 227.55~
```

16.4.5.1 Wrangle and plot the data

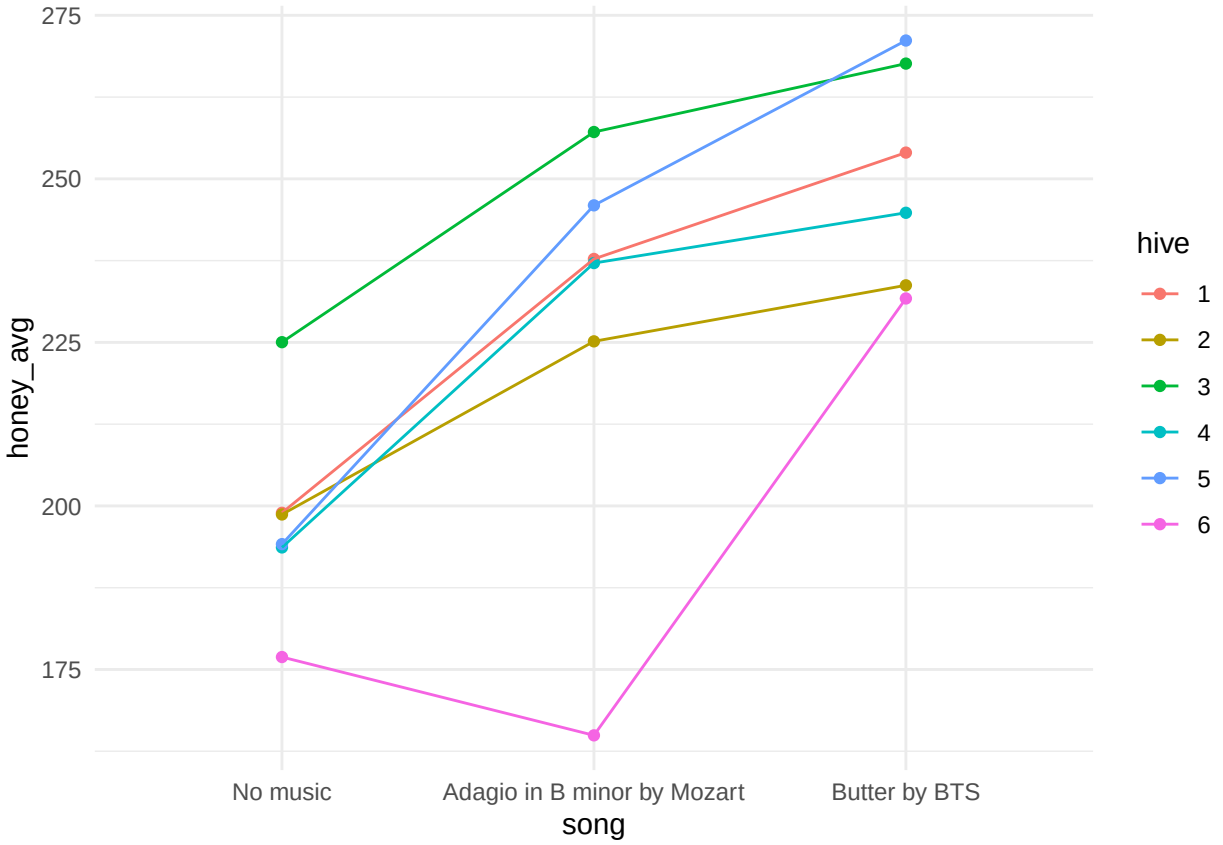
1. Wrangle the honey data so that the `song` variable is a factor in this order: No music, Adagio in B minor by Mozart and Butter by BTS.
2. Make `hive` a factor variable also.
3. Save this over the original honey dataset.
4. Then, using your new version of the `honey` dataset, create a new dataset called `honey_agg_int`, where you group by `hive` and `song` and summarize to find the average amount of honey produced by each hive while listening to a given song. Call your averaged variable `honey_avg`.

```
honey <- honey %>%
  mutate(song = fct_relevel(song, "No music", after = 0)) %>%
  mutate(hive = as_factor(hive))

honey_agg_int <- honey %>%
  group_by(hive, song) %>%
  summarize(honey_avg = mean(honey_g), .groups = "drop")
```

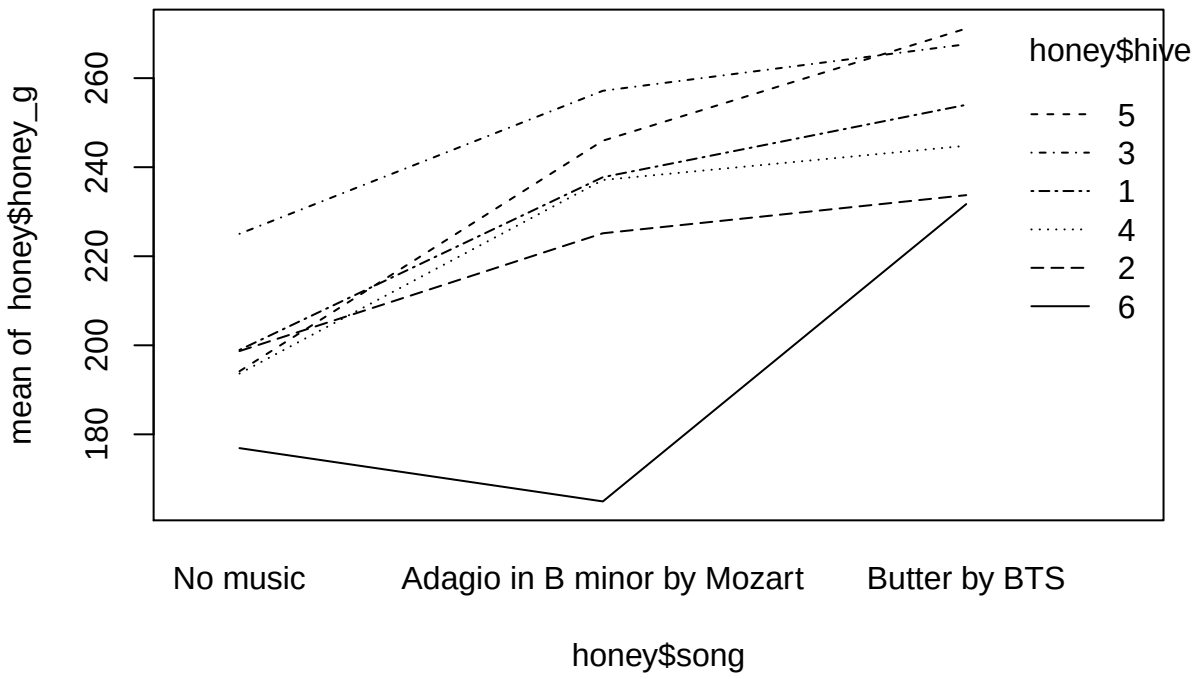
1. Plot the `honey_agg_int` data with `song` on the x axis and `honey_avg` on the y axis.
2. In your aesthetic, you should also colour by hive and group by hive. (See the hint if you get stuck).
3. Use the point and line geometries.
4. Use `theme_minimal()` if you like.

```
# This assumes you've run the correct code for the wrangling above
honey_agg_int %>%
  ggplot(aes(x = song, y = honey_avg, colour = hive, group = hive)) +
  geom_point() +
  geom_line() +
  theme_minimal()
```



We've actually made an interaction plot for our data. There is a function in the core `stats` package that also does this for us.

```
interaction.plot(honey$song, honey$hive, honey$honey_g)
```



Question: What has us worried about the independence assumption if we were just going to do regression as usual?

A. We don't expect honey production while listening to the same song to be independent.
B. We don't expect production for a given hive to be independent from month to month.
C. We don't expect these observations to be independent because they all experienced the same weather conditions and food availability.
D. No concerns about the independence assumption.

Answer: B. We expect that some hives will be more productive than others, naturally; a 'hive' effect. We want to estimate if there is in fact an effect of song, so that isn't an independence violation. We're glad the other conditions that could affect honey production were similar across the whole period for all hives as this means any difference we see is likely due to song effect (once we account for hive effect), not weather fluctuations.

This data is an example of **crossed effects**, because every hive experiences every level of the 'treatment,' i.e. each hive listens to each song.

Question: What would a completely nested version of this study look like?

A. This is both nested and crossed already.
B. Assign hives 1, 2 and to listen to randomly swap between no music for and then Mozart for 9 months. Hives 4, 5 and 6 would be similar but BTS instead of Mozart.
C. Repeated measures (x9) of honey production on each hive, listening to the same song, for 9 months.
D. Not possible to design a nested version.

Answer: C. You can't be nested and crossed for the same grouping variable. A fully nested version of this would be much like our tomato study, with two hives assigned to each song.

16.4.5.2 Model formula

This looks like:

$$y_{ijk} = \mu + \alpha_i + b_j + +(\alpha b)_{ij} + \epsilon_{ijk}$$

where y_{ijk} is the amount of honey produced (in grams) in the k^{th} month by the j^{th} hive while listening to song i. Here, μ is the grand mean of honey production, α_i are the I fixed effects for `song` and b_j are the random effects for hive j. $(\alpha b)_{ij}$ are the IJ interaction terms for the interaction between the the hive and the song. What does this represent here? Well, maybe some hives respond to a certain song quite differently to other hives. $(\alpha b)_{ij}$ is a random effect because any term involving a random effect must also be a random effect. $(\alpha b)_{ij} \sim N(0, \sigma^2_{\alpha b})$, $b_j \sim N(0, \sigma^2_b)$ and $\epsilon_{ijk} \sim N(0, \sigma^2)$. All the random effects are mutually independent random variables.

16.4.5.3 Fit the models

1. Fit an interaction model (main effects and interactions) to predict honey production based on `song` and `hive`. Call it `bee_int_mod`.

2. Fit a main effects (no interactions) model to predict honey production based on `song` and `hive` (no interaction). Call it `bee_main`.

3. Compare these two models with ANOVA.

```
bee_int_mod <- lm(honey_g ~ song*hive, data = honey)
bee_main <- lm(honey_g ~ song + hive, data = honey)
anova(bee_int_mod, bee_main)
```

```
## Analysis of Variance Table
##
## Model 1: honey_g ~ song * hive
## Model 2: honey_g ~ song + hive
##   Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
## 1     36  475.5
## 2     46 6570.0 -10   -6094.5 46.141 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question: What do you think we can claim from the analysis above?

A. It seems unlikely that the variance for the interaction terms is 0.
B. There is no evidence against the hypothesis that the main effects model fits the data as well as the interaction model.
C. These models are not identifiable.

Answer: A. Because the the model with the interaction explains the data significantly better than the main effects only model, we can say it seems like the variability explained by the interaction is going to be important, i.e. we need interaction terms because they aren't all the same value (which is how we'd have zero variance). These models are identifiable. (Run a summary on them, you'll get no errors.)

16.4.5.4 Model with our aggregated data

1. Run a new model using the `honey_agg_int` data, where `honey_avg` is the response and song and hive are fixed effects, no interactions. Save the model as `bee_agg_mod`.
2. Run an ANOVA on `bee_agg_mod`.

```
bee_agg_mod <- lm(honey_avg ~ song + hive, data = honey_agg_int)
anova(bee_agg_mod)
```

```
## Analysis of Variance Table
##
## Response: honey_avg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## song       2 8360.1  4180.1  20.576 0.0002855 ***
## hive       5 5914.7  1182.9   5.823 0.0089508 **
## Residuals 10 2031.5   203.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The very low p-values you'll see above suggest that we can also reject $H_0 : \sigma_b^2 = 0$ and $H_0 : \alpha_{BTS} = \alpha_{Mozart} = \alpha_{no\ song} = 0$ because it seems plausible that there is hive-to-hive variation, and that there are differences in honey production between the songs.

16.4.5.5 Describing our residuals

Recall the formula for the residual sum of squares (RSS):

$$RSS = \Sigma_{i=1}^{n} {e_i}^2 = \Sigma_{i=1}^{n} (y_i - \hat{y}_i)$$

where:

- e_i is the residual for observation i
- y_i is the observed value of the response variable for the i^{th} observation
- $\hat{y}_i$ is the **predicted** value (from the model) for the i^{th} observation

```
rss_int <- sum((honey$honey_g - predict(bee_int_mod))^2)
```

16.4.5.5.1 Calculate the residual sum of squares for the `bee_int_mod` model. Save it as `rss_int`.

$$RSE = \sqrt{\frac{RSS}{n - p - 1}}$$

where:

- RSS is the residual sum of squares
- n is the number of observations in your dataset (that are being used for the model, must be complete cases for the variables being used)
- p is the number of coefficients being estimated in our model (not including the intercept)

16.4.5.5.2 Calculate the RSE for the `bee_int_mod` model.

- Start by getting the sample size from the data (i.e., don't 'hard code' it). Save it as `n_int`.
  - Hint: `nrow()` is a useful function for finding the number of rows of a dataset.

- Calculate the degrees of freedom for your model. This is 'easy' (for a given value of easy…) to do by hand for balanced designs, but not something you'll do for more complicated models in practice.
  - Hint: $n - p - 1$
  - Save the value as `df_int`
- Finally, use these objects and `rss` calculated above to calculate the RSE. Save the object as `rse_int`.
  - You can check if you are right by comparing to the RSE from the fitted model: `Residual standard error: 3.634 on 36 degrees of freedom`
  - You can also check it directly with `summary(bee_int_mod)$sigma`

```
# Number of observations
n_int <- nrow(honey)
n_int
```

```
## [1] 54
```

```
# Degrees of freedom for the model
df_int <- n_int - nrow(summary(bee_int_mod)$coefficients)
df_int
```

```
## [1] 36
```

```
# Calculate RSE
rse_int <- sqrt(rss_int/(df_int))
rse_int
```

```
## [1] 3.63432
```

```
# Check RSE from the model
summary(bee_int_mod)
```

```
##
## Call:
## lm(formula = honey_g ~ song * hive, data = honey)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4400 -1.9533  0.1883  1.7300  9.0700
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         198.9500     2.0983  94.816  < 2e-16 ***
## songAdagio in B minor by Mozart      38.8067     2.9674  13.078 3.05e-15 ***
## songButter by BTS                    55.0633     2.9674  18.556  < 2e-16 ***
## hive2                                -0.2533     2.9674  -0.085  0.93244
## hive3                                26.0833     2.9674   8.790 1.73e-10 ***
## hive4                                -5.2933     2.9674  -1.784  0.08288 .
## hive5                                -4.7867     2.9674  -1.613  0.11546
## hive6                               -22.0500     2.9674  -7.431 9.01e-09 ***
```

```
## songAdagio in B minor by Mozart:hive2 -12.3433     4.1966  -2.941  0.00568 **
## songButter by BTS:hive2               -20.0333     4.1966  -4.774 2.99e-05 ***
## songAdagio in B minor by Mozart:hive3  -6.6800     4.1966  -1.592  0.12018
## songButter by BTS:hive3               -12.4767     4.1966  -2.973  0.00523 **
## songAdagio in B minor by Mozart:hive4   4.6633     4.1966   1.111  0.27384
## songButter by BTS:hive4                -3.9067     4.1966  -0.931  0.35809
## songAdagio in B minor by Mozart:hive5  12.9800     4.1966   3.093  0.00382 **
## songButter by BTS:hive5                21.9200     4.1966   5.223 7.58e-06 ***
## songAdagio in B minor by Mozart:hive6 -50.7767     4.1966 -12.100 3.02e-14 ***
## songButter by BTS:hive6                -0.2500     4.1966  -0.060  0.95283
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.634 on 36 degrees of freedom
## Multiple R-squared:  0.9904, Adjusted R-squared:  0.9858
## F-statistic: 217.9 on 17 and 36 DF,  p-value: < 2.2e-16
```

```
# We can pull the value directly from the model summary with $sigma
summary(bee_int_mod)$sigma
```

```
## [1] 3.63432
```

16.4.5.6 Calculating our random effect variances

16.4.5.6.1 Overview In this section I'm just going claim a couple equations for how we can partition (separate into components) the unexplained variation in the models we have fit.

We will have a variance component for every one of our random effects (and interactions between a fixed and random effect are also random effects themselves!) as well as the overall residual variance.

We can use this to claim what proportion of variance is explained by our different random effects and what proportion remains unexplained by our variables.

I practice, we'll let R handle this for us, especially as things will get a little messier when we leave the safety of balanced designs. See the Wood reading referenced in the credits if you want to cover this further now. Optional, *not* assessed.

Basically, I want you to take away from this section that **we can estimate how much our different sources of variation are each contributing to our overall variation.**

16.4.5.6.2 Unexplained variance after fitting the interaction model As we saw above, our unexplained variability, after fitting the interaction model (the most complicated linear model possible for this data), `bee_int_mod`, gives us our σ^2, i.e., the variability of the error term is $\sigma^2 = 3.63^2$.

Save the squared residual standard error from the `bee_int_mod` model as `var_int`.

```
var_int <- summary(bee_int_mod)$sigma^2
var_int
```

```
## [1] 13.20828
```

16.4.5.6.3 RSE for main effects model on aggregated data Next, we want to consider the what portion of the variability is explained by the interaction of hive and song. Our `bee_agg_mod` residual variance will come from two sources: the variance that CAN be explained by the interaction of hive and song AND the variability we cannot explain (even with out most complicated model), our random error.

$$\text{RSE}^2_{bee\_agg} = \sigma^2_{ab} + \frac{\sigma^2}{K}$$

So, RSE_{bee_agg} is the `sigma` value calculated from our main effects only model on the aggregated data, and K is the number of observations at each level of our interaction, i.e., the number of levels we have averaged over in aggregating this data. We want this value as we can use it to take the average of the squared errors from our interaction model (remember that $sigma^2$ is the RSE squared.)

We can get K from the context, by understanding that each hive was exposed to each song for 3 months and honey production was measured monthly. OR we can remember that this design is balanced and divide the total sample size ($n = 54$) by the number of combinations (levels of interaction) which will be 6 hives times 3 songs, which is then $\frac{54}{6 \cdot 3} = 3$.

Doing some rearranging, we now can calculate σ^2_{ab} from other values we have available to us.

$$\sigma^2_{ab} = \text{RSE}^2_{bee\_agg} - \frac{\sigma^2}{K} =$$

198.75

Calculate σ^2_{ab} and save it as `var_ab`.

```
var_ab <- summary(bee_agg_mod)$sigma^2 - var_int / 3
var_ab
```

```
## [1] 198.7458
```

16.4.5.6.4 Hive-to-hive variance Finally, to get σ^2_b we do one final aggregation and model.

Start by creating a new dataset called `honey_agg_hive`.

- Start with the `honey` data
- Group by `hive`
- Use summarize to get the `honey_avg_hive`
  - You can use `.groups = "drop"` as an argument in your `summarize` call if you are getting a message about dropping groups.

```
honey_agg_hive <- honey %>%
  group_by(hive) %>%
  summarize(honey_avg_hive = mean(honey_g))

honey_agg_hive
```

```
## # A tibble: 6 x 2
##   hive  honey_avg_hive
##   <fct>          <dbl>
## 1 1               230.
## 2 2               219.
```

```
## 3 3              250.
## 4 4              225.
## 5 5              237.
## 6 6              191.
```

Once you have your `honey_agg_hive` dataset, use it to fit an **intercept-only** model, saved as `bee_hive_mod` that effectively just calculate one coefficient, the grand mean of honey production in our data. The **RSE** for this model is really the same as the standard deviation of average honey production, hive-to-hive.

```
bee_hive_mod <- lm(honey_avg_hive ~ 1, data = honey_agg_hive)
summary(bee_hive_mod)
```

```
##
## Call:
## lm(formula = honey_avg_hive ~ 1, data = honey_agg_hive)
##
## Residuals:
##        1        2        3        4        5        6
##   4.7669  -6.2787  24.4646  -0.2743  11.6135 -34.2920
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  225.473      8.107   27.81 1.12e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.86 on 5 degrees of freedom
```

```
# Notice that the RSE here is just the SD of the average honey production hive-to-hive
sqrt(var(honey_agg_hive$honey_avg_hive))
```

```
## [1] 19.85727
```

So now we can take our hive-to-hive variability subtract the variability explained by the song and hive main effects, and find the variability in honey output hive-to-hive not explained by song or song/hive interactions. We are dividing by I (the number of songs) because we are averaging across them.

$$\sigma_b^2 = \text{RSE}_{bee\_hive}^2 - \frac{\text{RSE}_{bee\_agg}^2}{I}$$

```
# hive to hive variation, not explained by song or song/hive interaction
var_hive <- summary(bee_hive_mod)$sigma^2 - (summary(bee_agg_mod)$sigma^2)/3
var_hive
```

```
## [1] 326.595
```

```
# variance in honey production explained by the interaction of song and hive
var_ab <- summary(bee_agg_mod)$sigma^2 - var_int / 3
var_ab
```

16.4.5.6.5 Summary of all our calculated error variances

```
## [1] 198.7458
```

```
# variance in honey production explained by hive to hive variance
var_hive <- summary(bee_hive_mod)$sigma^2 - (summary(bee_agg_mod)$sigma^2)/3
var_hive
```

```
## [1] 326.595
```

```
# residual variance (unexplained)
var_int <- summary(bee_int_mod)$sigma^2
var_int
```

```
## [1] 13.20828
```

We can all talk about the proportion of variance explained by each of our sources of variation. This is also called intraclass correlation coefficient (ICC) or the variance partition coefficient (VPC). We can calculate it by finding the sum of all these variances, and setting that value as our denominator.

```
tibble(Source = c("song:hive", "hive", "residual"),
  `Proportion of variance explained` = c(
    var_ab / (var_ab + var_hive + var_int),
    var_hive / (var_ab + var_hive + var_int),
    var_int / (var_ab + var_hive + var_int))) %>%
  knitr::kable()
```

Source	Proportion of variance explained
song:hive	0.3690394
hive	0.6064349
residual	0.0245257

Question: Which of our random effects explains the greatest variance in the honey production, after controlling for our fixed effects?

A. Variations in the interaction between the hives and songs.
B. Random noise.
C. Hive-to-hive variability.
D. Impossible to say.

Answer: C. Of the three variances for our random effects, the variance for the hive effect is the largest, and so explains the greatest proportion of the variability in honey production not explained by the song differences."

16.4.5.7 Summary

Note the below notation is is based on using least squares and treating all variables as fixed effects. This isn't the way you would explain a linear model in practice, but I hope helps you see how we are thinking about partitioning our variance. It also helps remind us that regression is just fancy averaging.

Link to tweet

Note: The · you'll see in some of the following subscripts indicate where we are aggregating and averaging across that level of the data.

16.4.5.7.1 Full interaction model (includes main effects) (`bee_int_mod`)

In R	Notation	RSE
`lm(honey_g ~ song*hive, data = honey)`	$y_{ijk} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + \epsilon_{ijk}$ **Traditional linear regression** $y_{ijk} = \beta_0 + \beta_1 \text{song}_i + \beta_2 \text{hive}_j + \beta_3 \text{song}_i \text{hive}_j + \epsilon_{ijk}$ These are equivalent, just with different contrasts (how we set up our model matrix).	RSE^2 is σ^2 is the u after fitting the mo model.

16.4.5.7.2 Main effects only model, on hive/song aggregated data (`bee_agg_mod`)

In R	Notation	RSE
`lm(honey_g ~ song + hive, data = honey_agg)`	$y_{ij\cdot} = \mu + \alpha_i + b_j + \epsilon_{ij}$ where $\epsilon_{ij} = (\alpha b)_{ij} + \frac{1}{K}\Sigma_{k=1}^{n}$ **Traditional linear regression** $y_{ij\cdot} =$ $\beta_0 + \beta_1 \text{song}_i + \beta_2 \text{hive}_j + \epsilon_{ij}$	The RSE^2 here is the unexplained variation in the average honey production after accounting for the main effects of song and hive. We will use it to help find $\sigma_{\alpha b}$.

16.4.5.7.3 Intercept only model for hives averages (`bee_hive_mod`)

In R	Notation	RSE
`lm(honey_g ~ 1, data = honey_agg_hive)` (note aggregated dataset)	$y_{\cdot j\cdot} = \mu + \frac{1}{I}\Sigma_{i=1}^{I}\alpha_i + \epsilon$ where: $\epsilon_j = b_j + \frac{1}{I}\Sigma\epsilon_j$ **Traditional linear regression** $y_{\cdot j\cdot} = \beta_0 + \epsilon_j$	The RSE^2 here is hive-to-hive variation in average honey output (without taking into account the interaction between hive and song). We will use it to help find σ_b.

16.4.6 Where to next?

We've done pretty much everything we can while remaining in the ordinary least squares context, and even in the simplest examples, all the aggregating and model fitting we're doing will get a little tedious. Now we will turn to fitting **linear mixed models** more in more sophisticated ways with a specialist package.

The models we'll be fitting will rely on maximum likelihood methods.

16.4.7 Credits

- The commentary for this activity draws heavily on section 2.1 of Wood, S. Generalized Additive Models: An Introduction with R, 2nd Edition. 2017. https://ebookcentral-proquest-com.myaccess.library.utoronto.ca/lib/utoronto/detail.action?docID=4862399 (requires you to log in with your UTORid).
- Overall concept based on the Stardew Valley game.
- Stardew Valley images from https://stardewcommunitywiki.com.

16.5 Vocal pitch case study: Part 1 (LMMs)

16.5.1 Motivation

You're interested in a language and especially want to investigate how **voice pitch** (perceived "highness" or "lowness" of a voice) is related to politeness for speakers of that language.

16.5.2 Design of the experiment

- You recruit 6 speakers of the language as subjects in your study.
- There are 3 males and 3 females.
- Each subject is asked to respond to 14 hypothetical situations.
  - 7 of these situations are *formal* situations that usually require politeness (e.g., giving an excuse for being late to your scary boss), and
  - 7 were more *informal* situations (e.g., explaining to a friend why you are late).
- For each hypothetical situation you measure the subjects voice pitch as they respond. You also take note of each of your subjects' sex, since you know that's another important influence on voice pitch.

Note: Much of the content here is adapted from Winter, B. (2013). **Linear models and linear mixed effects models in R with linguistic applications**. arXiv:1308.5499. The language considered was Korean.

Photos from unsplash.com

16.5.3 Read in the data and explore it

For convenience, I have created a new variable called `condition`, it is just a modification of the `attitude` variable, that takes the level `formal` if the particular measurement was under formal conditions and `informal` if the measurement was under informal conditions. This variable contains the same information as the `attitude` variable, just coded how I want it to be, so I remove the `attitude` variable at the end.

```
# data wrangling
polite_data = read_csv('data/m3/politeness_data.csv') %>%
  mutate(condition = as_factor(if_else(attitude == "pol", "formal", "informal"))) %>%
  mutate(sex = as_factor(gender))
```

```
glimpse(polite_data, width=85)
```

```
## Rows: 84
## Columns: 7
## $ subject   <chr> "F1", "F1", "F1", "F1", "F1", "F1", "F1", "F1", "F1", "F1", "F1",~
## $ gender    <chr> "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", ~
## $ scenario  <dbl> 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 1, 1, 2, 2, 3, 3, 4, 4,~
```

```
## $ attitude  <chr> "pol", "inf", "pol", "inf", "pol", "inf", "pol", "inf", "pol", "i~
## $ frequency <dbl> 213.3, 204.5, 285.1, 259.7, 203.9, 286.9, 250.8, 276.8, 231.9, 25~
## $ condition <fct> formal, informal, formal, informal, formal, informal, formal, inf~
## $ sex       <fct> F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F,~
```

```
head(polite_data)
```

```
## # A tibble: 6 x 7
##   subject gender scenario attitude frequency condition sex
##   <chr>   <chr>     <dbl> <chr>        <dbl> <fct>     <fct>
## 1 F1      F             1 pol           213. formal    F
## 2 F1      F             1 inf           204. informal  F
## 3 F1      F             2 pol           285. formal    F
## 4 F1      F             2 inf           260. informal  F
## 5 F1      F             3 pol           204. formal    F
## 6 F1      F             3 inf           287. informal  F
```

Without doing any analysis, how might this data violate our assumptions for a linear model?

16.5.4 Recall: Linear regression assumptions

1. **Errors are independent (observations are independent)**
2. Errors are identically distributed and the expected value of the errors is zero, $E[\epsilon_i] = 0$
3. Constant variance (homoscedasticity), $var[\epsilon_i] = \sigma^2$.
4. A straight-line relationship exists between the errors ϵ_i and responses y_i

1.-3. are usually expressed as assuming the errors are i.i.d normally distributed with mean of zero and variance σ^2,

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

16.5.4.1 Why we can't assume independence

Each subject has 14 measurements of their pitch (measured as the frequency of the vibration of your vocal cords) in different scenarios and different levels of politeness (formal/informal).

With this kind of data, since each subject gave multiple responses (a "repeated measures" design), we can see that this would violate the independence assumption that's important in linear modelling: **multiple responses from the same subject cannot be regarded as independent from each other**.

In this scenario, every person has a *slightly different* voice pitch. This personal factor affects all responses from the same subject, thus responses are inter-dependent (correlated) within subjects rather than independent. I.e., if you have a generally low voice, even if you change your pitch when you are being informal versus formal, you probably still have a generally low pitch.

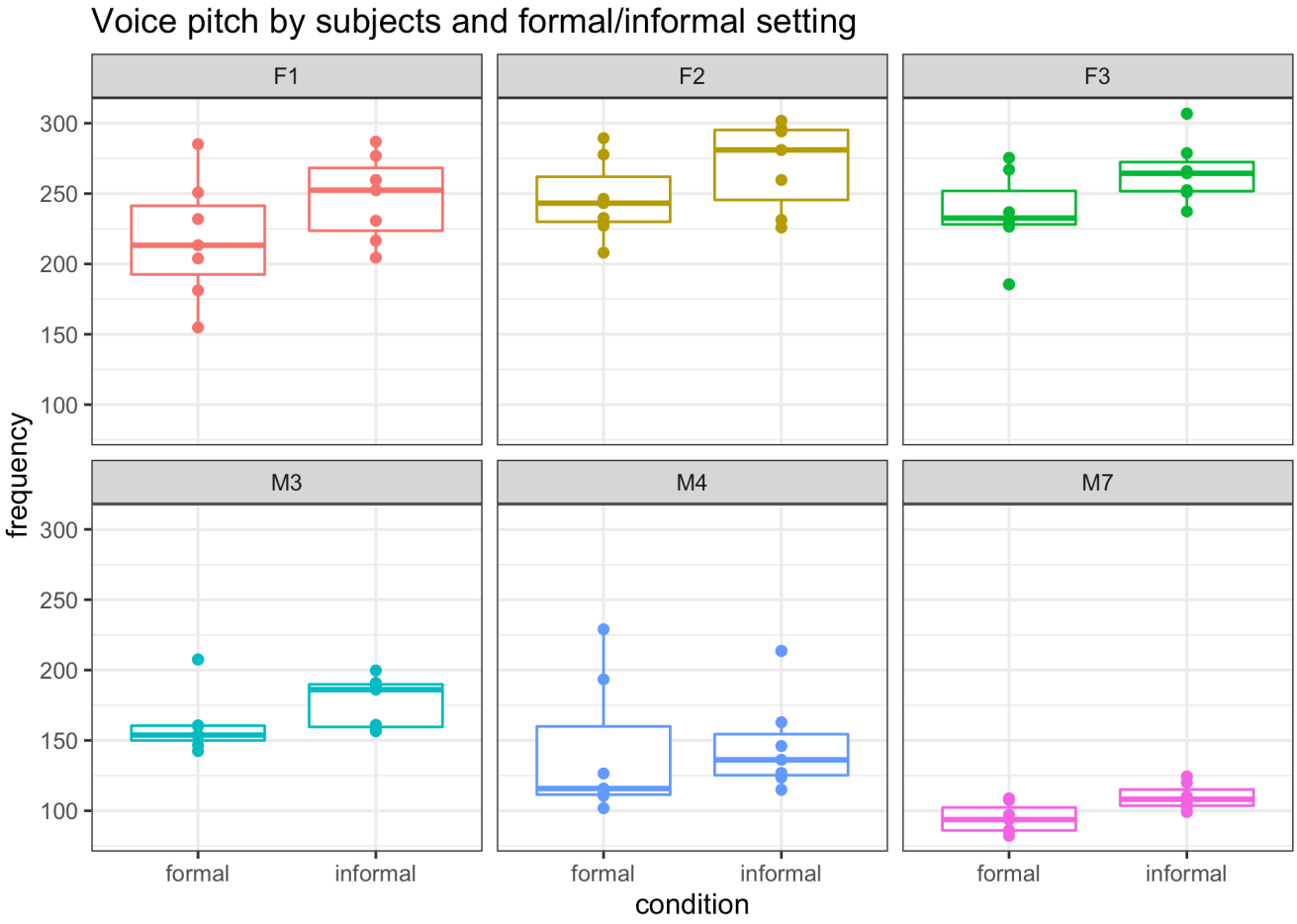
There also differences, on *average* in pitch by sex, so this is something we will also want to account for in our model, later.

So what do we do?

16.5.4.2 Let's start with a picture

Female subjects have an ID starting with "F" and males subjects have an ID starting with "M." **What do you see?**

```
polite_data %>%
  ggplot(aes(x = condition,
             y = frequency,
             color=subject)) +
  geom_boxplot() +
  geom_point() +
  facet_wrap(~subject, nrow = 2) +
  ggtitle(label="Voice pitch by subjects and formal/informal setting") +
  theme_bw() +
  theme(legend.position = "none") # remove legend because we have facet names
```



16.5.4.2.1 Key themes from our plot

1. Males tend to have lower voices than females.
2. Within the male and the female groups, there is plenty of individual variation. Some subjects have relatively high pitch for their sex and others have relatively low values, regardless of the formal/informal version of the scenario.
3. Subjects seem to generally use a higher pitch when speaking informally than they do when speaking formally.

Before we can go further with this case study, let's establish a few things...

16.6 Linear mixed models

16.6.1 Thoughts on plots for hierarchical/correlated data generally

- Make use of faceted plots! You may have also seen these called lattice plots or trellis plots. These plots show you relationships in your data, conditioned on one or more of your other variables. Facet based on your grouping units.
- When working with ggplot you can facet by adding the `facet_wrap` command. You can control which variables you condition on and how many rows and columns the facets are organised into. (There are lots of other great features we won't go in to in depth here.)

E.g. `plot + facet_wrap(~one_var)` or `plot + facet_wrap(~one_var + two var, nrow = 3)`

Assume `plot` is a ggplot object.

These names come from the appearance of these plots being similar to the lattice of a garden trellis. They also show you different *faces* (facets) of your data.]

16.6.2 Assumptions

Linear mixed models assume that:

1. There is a continuous response variable.
2. We have modelled the dependency structure correctly (i.e. made correct choices about our random variables).
3. Our units/subjects are independent, even through observations within each subject are taken not to to be.
4. Both the random effects and within-unit residual errors follow normal distributions.

5. The random effects errors and within-unit residual errors have constant variance .

While it is possible to some extent to check these assumptions through various diagnostics, a natural concern is that if one or more assumptions do not hold, our inferences may be invalid. Fortunately it turns out that linear mixed models are robust to violations of some of their assumptions.

16.6.3 Our model set up

16.6.3.1 Our STA302 linear model

$$y = X\beta + \epsilon$$
$$\epsilon \sim N(0, \Lambda_\theta)$$

16.6.3.2 Linear mixed model

$$y = X\beta + Zb + \epsilon$$
$$b \sim N(0, \psi_\theta), \quad \epsilon \sim N(0, \Lambda_\theta)$$

- y is vector of outcomes for subject i,
- X and Z are model matrices for the **fixed** and **random** effects, respectively,
- the vector β describes the effect of covariates on the mean/expectation of the outcome, b is the random effects for the units (assumed to be normally distributed with mean zero),
- ϵ is the residual errors, normally distributed with a given variance and the errors within units are mutually independent. In this course we won't deal with more complicated situations than a simple error vector that is distributed $N(0, \sigma^2)$.

16.6.3.3 Alternative formulations

Where you have observations Y_{ij} for repeated measures j on individuals i, you may also see:

$$Y_{ij} = X_{ij}\beta + \epsilon_{ij}$$

Where $\epsilon_{ij} = U_i + Z_{ij}$ and $Z_{ij} \sim N(0, \tau^2)$.

These errors are normally distributed BUT correlated.

OR

$$Y_{ij}|U_i \sim N(\mu_{ij}, \tau^2)$$
$$\mu_{ij=X_{ij}\beta+U_i}$$
$$[U_1, \ldots, U_M]^T \sim MVN(0, \Sigma)$$

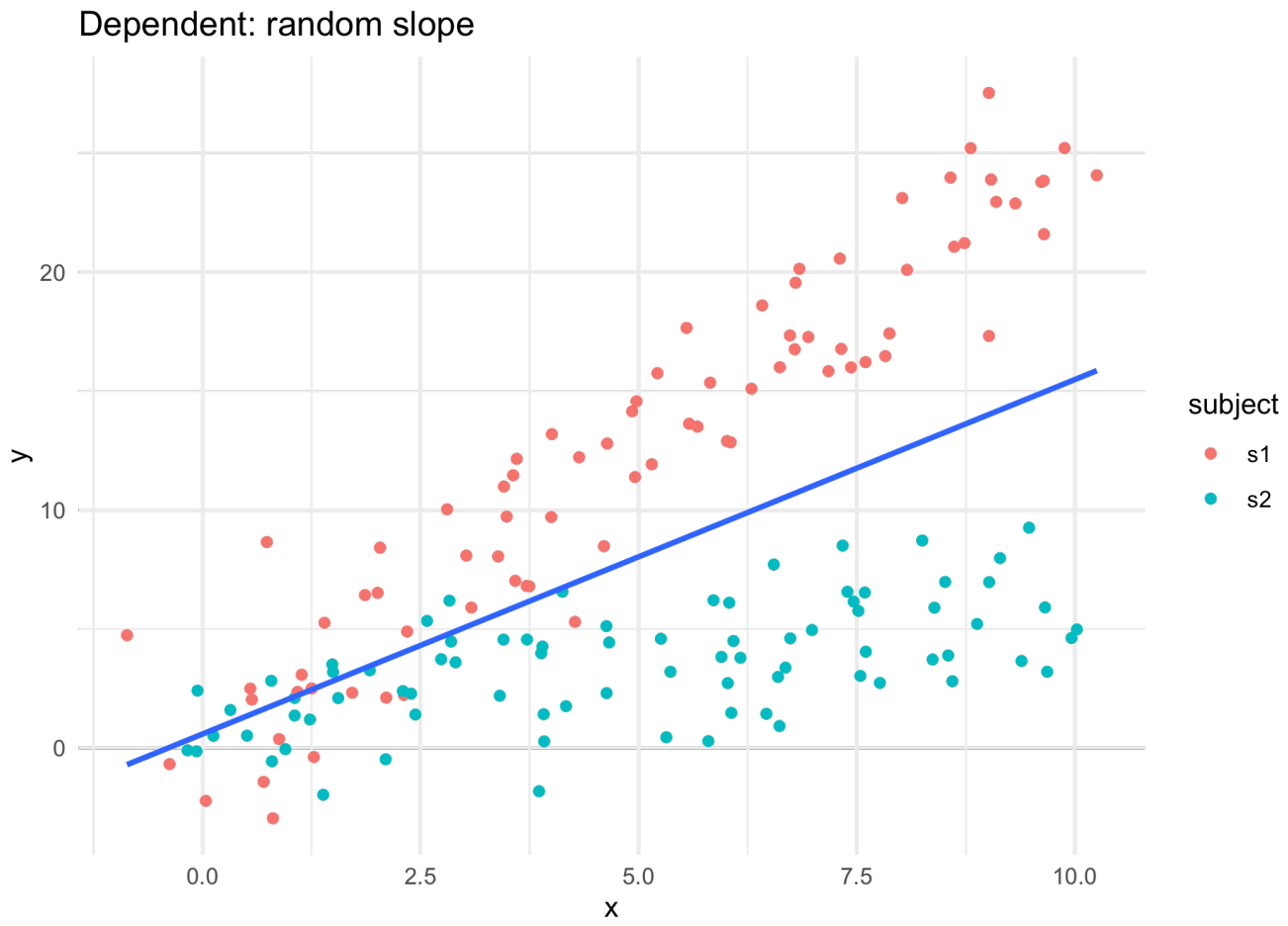
- fixed effects: $X_{ij}\beta$ (what you're used to)
- random effects: U_i for i in 1 to M (new part that makes this a linear mixed model)

16.6.4 What can correlated errors look like?

16.6.4.1 Random intercepts

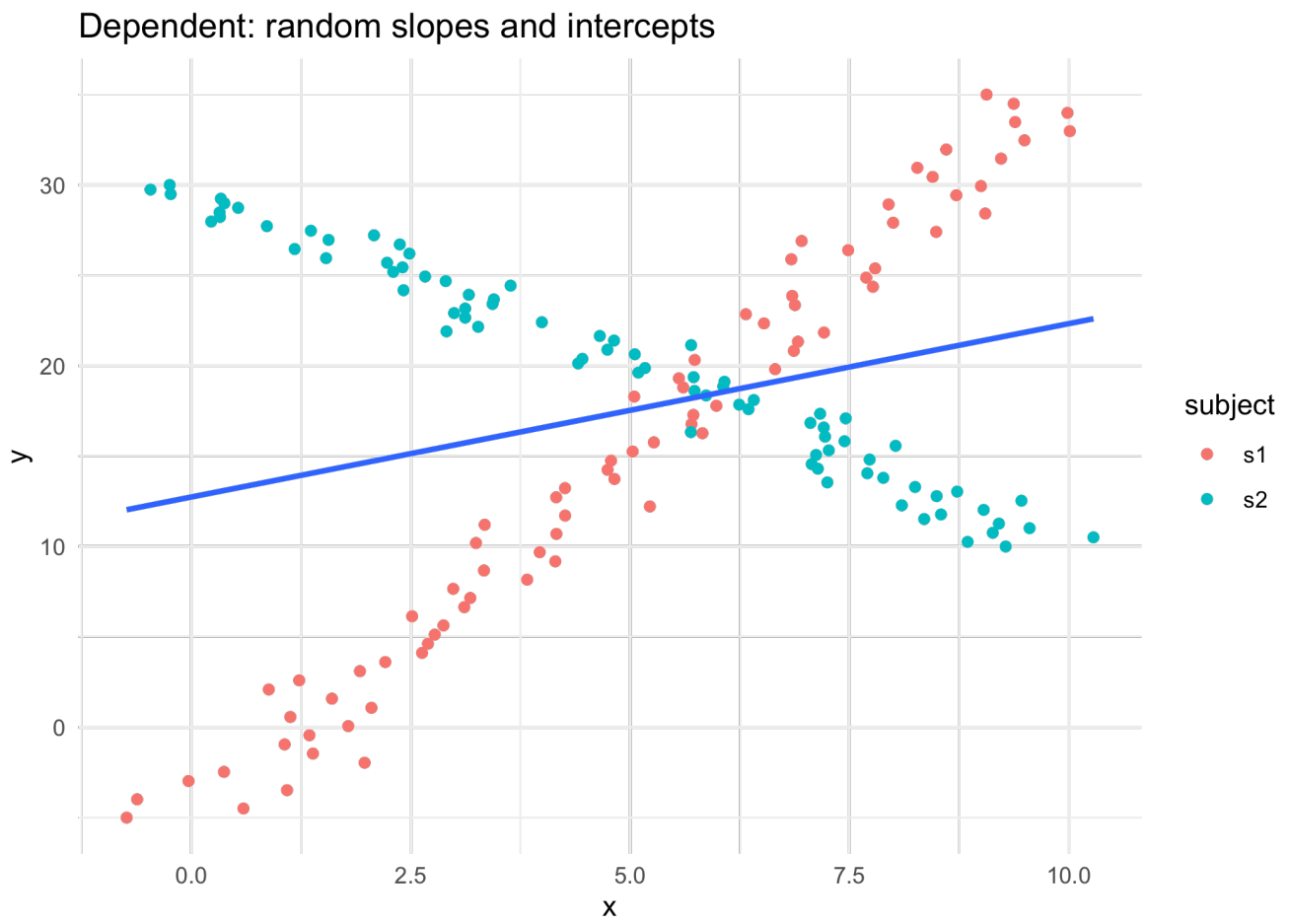
This is one example of dependence. Here, if a subject is above average in one observation, they're likely to be above average in their other observations.

16.6.4.2 Random slopes



In this case, not only would the residuals for subject 1 tend to be positive while the residuals for subject 2 would tend to be negative, the difference would get bigger as x increases. You don't want that kind of pattern in the residuals!

16.6.4.3 Random slopes and intercepts



Subjects appear to have different slopes AND intercepts.

16.6.5 How do we tell R which situation we're in?

We'll be using the `lmer()` function from the `lme4` package.

Many R packages provide a "vignette" to help you understand how to use the package and the theory behind it.

```
# This line only installs lme4 if you haven't got it installed already
if(!("lme4" %in% installed.packages())) install.packages("lme4")

install.packages("lme4")
vignette("lmer")
```

16.6.5.1 Vignette activity

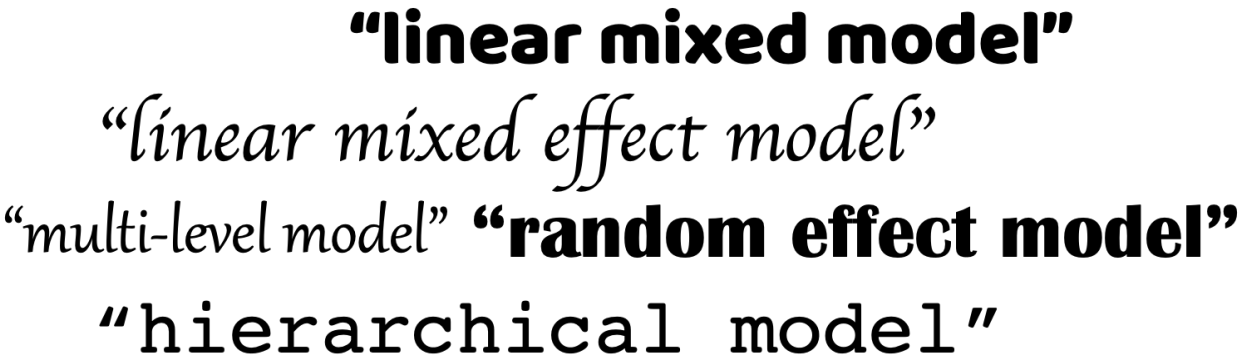
Have you ever looked at a package vignette before?

Use the vignette for this package (see the code above) to try to fill out the syntax for the following situations. You may find Table 2 on page 7 of the vignette quite helpful.

Type of dependency	lme4 syntax
Random intercept	
Random slope	
Random slope AND intercept	

16.6.6 Additional considerations

16.6.6.1 Linear mixed models come with lots of different names



16.6.6.2 Fixed vs. random effects

Fixed effects are the things you care about and want to estimate. You likely chose the factor levels for a specific reason or measured the variable because you are interested in the relationship it has to your response variable. Random effects can be variables that were opportunistically measured whose variation needs to be accounted for but that you are not necessarily interested in (e.g. spatial block in a large experiment). The levels of the random effect are likely a random subset of all possible levels (although as a rule of thumb, there should usually be at least 5). However, if the experimental design includes nesting or non-independence of any kind, this needs to be accounted for to avoid pseudoreplication.

16.6.6.3 REML or ML

The math behind maximum likelihood (ML) and restricted maximum likelihood (REML) is beyond what I will assess you on in this course. The main difference between the two is how they estimate the variance parameters. They will estimate βs the same way.

REML is effectively a two-stage approach where estimates of the random effects are conditioned on the fixed effects. REML accounts for the degrees of freedom lost when estimating the fixed effects and provides us with unbiased estimates of our variance components. ML estimates of the variance components tend to be biased downward (smaller than they should be), making us more likely to pick models with simpler random effects structures than we should.

We **prefer ML** if we need to compare two **nested** models based on their *fixed effects*, as we can't do likelihood ratio tests with models fit with REML *unless* the fixed effects are exactly the same. We **prefer REML** when there is a large number of parameters or if our main goal is estimates of our model parameters (random and fixed). Which is *usually* our goal.

For this course, we'll usually use REML, unless we are trying to compare nested models with the same random effects. There are often not major differences between the two methods, anyways.

16.6.6.4 Confidence intervals and tests

You can create Wald confidence intervals for fixed effects the way you're used to from STA302. BUT for our random effects things can get a bit more awkward.

A variance must be in $[0, \infty)$, so what happens if you get an estimate for $\widehat{\sigma_b^2}$ that is $0.02 \pm 0.04 = (-0.02, 0.06)$? We can get into trouble if our variance components are close to the boundary of the parameter space or if a test requires us to fix them at these boundaries.

Despite these difficulties, if our interval estimates of our variance components are pretty safely away from 0, then we can conclude the associated random effect is needed in the model. If it seems that a variance component is effectively 0, than we can conclude we don't need the associated random effect. We can also get this from likelihood ratio tests, where big or small p-values lead to clear-cut conclusions, but we face more challenges for making inferences when our p-value is close to our decision criteria.

16.7 Vocal pitch case study: Part 2

Recall:

$$Y_{ij} = X_{ij}\beta + Z_i b + \epsilon_{ij}$$

- Y_{ij} is the vocal pitch for the ith subject on the jth vocal response.
- $X_{ij}\beta$ has an intercept and an effect for sex and condition (formal/informal)
- $Z_i b$ represents subject i's 'baseline' pitch
- ϵ_{ij} is random noise (and potentially other unmeasured confounders)

16.7.1 Modelling individual means with random intercepts

We can consider each subject's mean vocal pitch and in our model we will assume different random intercepts for each subject. The mixed model estimates these intercepts for us.

We are going to use the package `lme4` and the function `lmer()` and then fit a model with only the random effect of subject (each subject gets their own intercept).

16.7.1.1 Random intercept only model

```
# install.packages("lme4")
library(lme4) # load the package

sub_only = lmer(frequency ~ (1 | subject),  data = polite_data)
```

```
summary(sub_only)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: frequency ~ (1 | subject)
##    Data: polite_data
##
## REML criterion at convergence: 828.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4980 -0.6071 -0.1817  0.6433  2.7520
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subject  (Intercept) 4003     63.27
##  Residual              940     30.66
## Number of obs: 84, groups:  subject, 6
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   192.66      26.04   7.397
```

16.7.1.2 What does this model do?

`(1 | subject)` is the R syntax for a random intercept. It asks that we assume there is a different intercept for each subject. The '1' stands for the intercept and the term to the right of the '|' should be a nominal or factor variable to be used for the random effect. You can think of this formula as telling your model that it should expect that there's going to be multiple responses per subject, and these responses will depend on each subject's baseline level. This effectively resolves the non-independence that stems from having multiple responses by the same subject.

16.7.1.3 Interpreting this output

From top to bottom:

- The model was fit using restricted maximum likelihood (REML) - what is that? We'll see soon.

- Then we are shown our formula again and the data this model was fit with.
- The next two parts we won't use much: there is much deeper you could go on the fitting and convergence of these models, but that is not within our scope, and we expect to see the scaled residuals centered near to 0.
- The next part gives us the estimated variances for the random effects in the model. Recall the variances we calculated diretly in the Statdew Valley activity.
- Finally, we have the fixed effect portions of the model, with a separate intercept but nothing else yet because we haven't added any fixed effects.

16.7.1.4 Adding our fixed effects

We're interested in the effects of a formal/informal scenario and sex on the pitch of the subject's voice, so we'll add these as fixed effects.

Question of interest: What is the influence of formal vs informal condition on voice pitch in speakers of this language, while accounting for sex and variation within subjects?

16.7.1.5 Model with condition and sex

```
w_fixed = lmer(frequency ~ condition + sex + (1|subject), data=polite_data)
summary(w_fixed)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: frequency ~ condition + sex + (1 | subject)
##    Data: polite_data
##
## REML criterion at convergence: 795.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3629 -0.5677 -0.2082  0.4615  3.2900
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subject  (Intercept) 599.5    24.49
##  Residual             844.6    29.06
## Number of obs: 84, groups:  subject, 6
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)        237.054     15.166  15.631
## conditioninformal   19.864      6.342   3.132
## sexM              -108.660     20.974  -5.181
##
## Correlation of Fixed Effects:
##             (Intr) cndtnn
## condtnnfrml -0.209
## sexM        -0.691  0.000
```

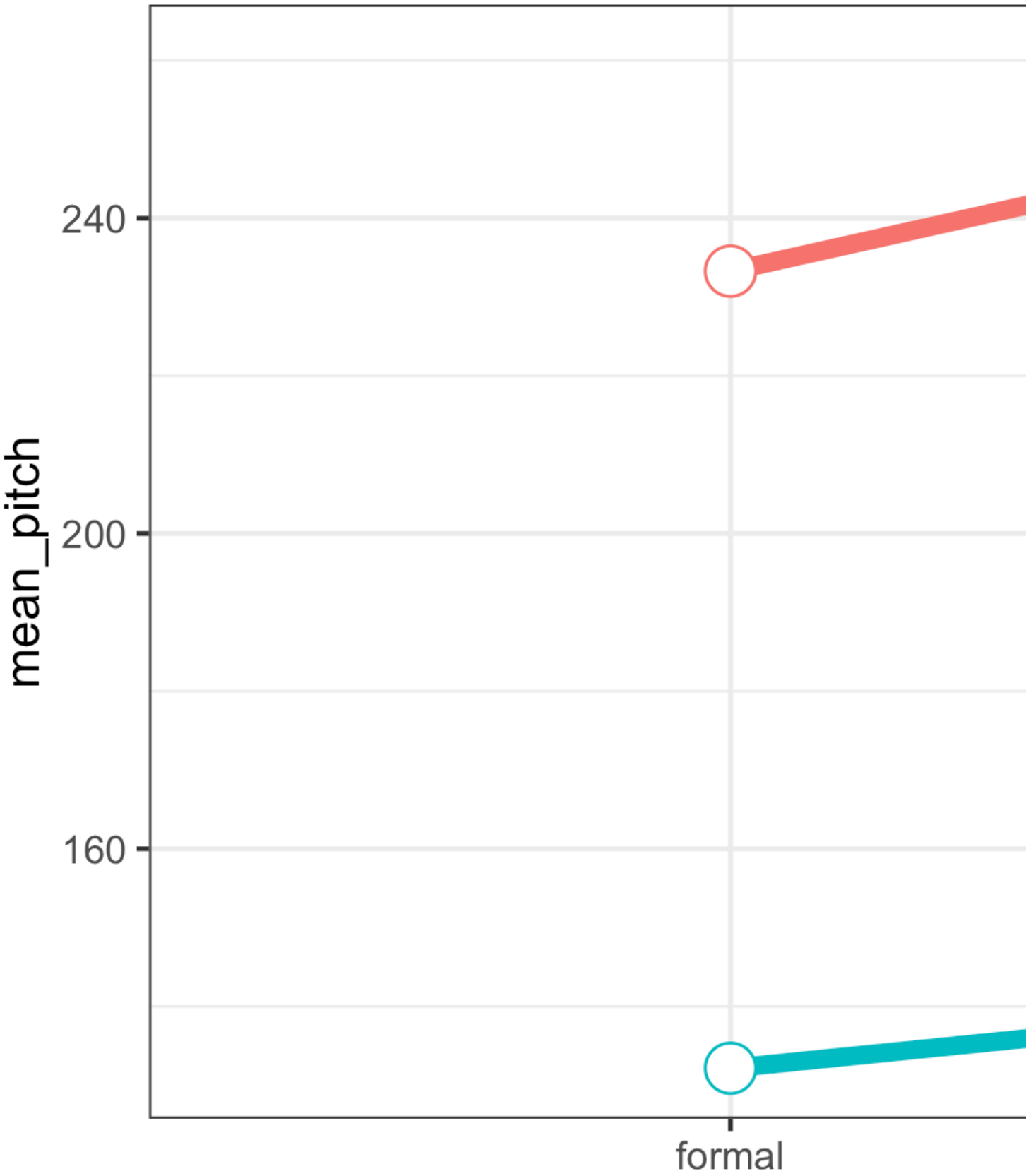
You should be able to interpret the fixed effects in the way you are used to from linear models. Interpret the fixed effects coefficients below.

```
fixef(w_fixed)
```

```
##       (Intercept) conditioninformal              sexM
##         237.05357          19.86429        -108.65952
```

Males have lower pitches than females by an average of 109 Hz and in informal situations speakers have a higher pitch by about 20 Hz on average.

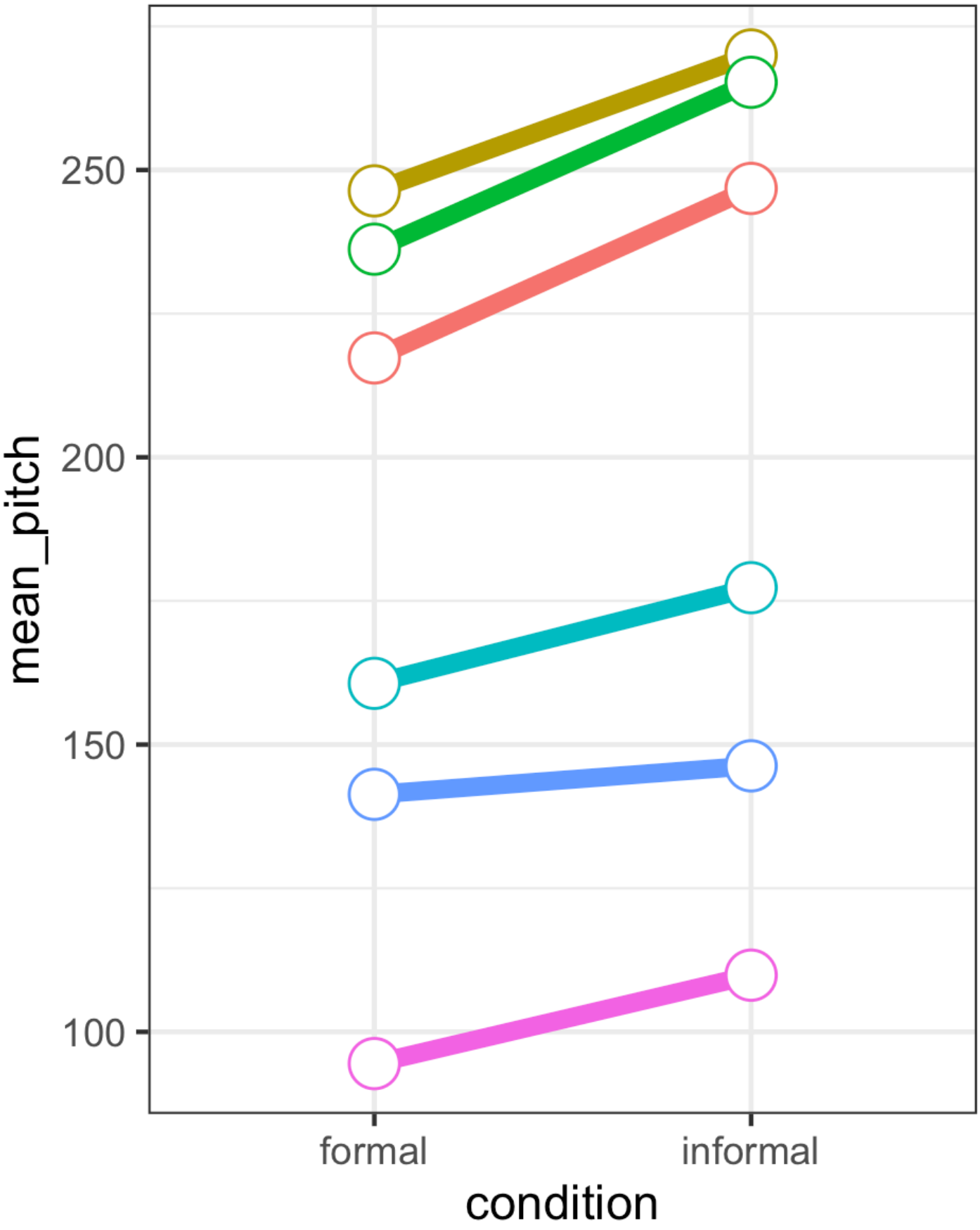
16.7.1.6 Aside: Summary of means



This gives us a general idea of what is going on…but doesn't account for variability at all! Maybe these differences don't actually mean anything in the context of the variability of our data. That is why we want to take a modelling approach.

16.7.1.7 Random slopes for condition

Previously we assumed that the effect of formal/informal conditions were the same for all subjects (one coefficient for this variable). However, the effect of the condition might be *different for different subjects*; that is, there might be a condition and subject interaction. For example, it might be expected that some people are more polite in formal scenarios, others less. We need a random slope model, where subjects and items are not only allowed to have differing intercepts, but where they are also allowed to have different slopes for the effect of formality on pitch.


mean_pitch
250
200
150
100
formal
informal
condition

Do you think the slopes are different for different subjects?

Note: The following model generates that warning `boundary (singular) fit: see ?isSingular` because our random effects are very small. This is an indication that we may be overfitting. We are including a random effect that we might not need to...(which, spoilers! is what we'll find out at the end of this secion using our model comparison).

```
w_slopes = lmer(frequency ~ condition + sex + (1 + condition | subject),  data = polite_data)
```

```
## boundary (singular) fit: see ?isSingular
```

```
summary(w_slopes)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: frequency ~ condition + sex + (1 + condition | subject)
##    Data: polite_data
##
## REML criterion at convergence: 795.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3537 -0.5822 -0.2209  0.4785  3.2847
##
## Random effects:
##  Groups   Name              Variance Std.Dev. Corr
##  subject  (Intercept)       621.9309 24.939
##           conditioninformal   0.8045  0.897   -1.00
##  Residual                   844.3846 29.058
## Number of obs: 84, groups:  subject, 6
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)        237.737     15.279  15.560
## conditioninformal   19.864      6.352   3.127
## sexM              -110.027     20.946  -5.253
##
## Correlation of Fixed Effects:
##             (Intr) cndtnn
## condtnnfrml -0.246
## sexM        -0.685  0.000
## convergence code: 0
## boundary (singular) fit: see ?isSingular
```

These new condition slopes are looking pretty similar...

```
coef(w_slopes)
```

```
## $subject
##    (Intercept) conditioninformal      sexM
## F1    223.1947          20.38733 -110.0272
## F2    247.4326          19.51559 -110.0272
## F3    240.4958          19.76508 -110.0272
```

```
## M3    266.7502          18.82080 -110.0272
## M4    243.5792          19.65418 -110.0272
## M7    204.9719          21.04275 -110.0272
##
## attr(,"class")
## [1] "coef.mer"
```

Let's compare some models. `lmtest::lrtest` will run a likelihood ratio test for us.

What is your conclusion? Do we need the random slopes?

```
lmtest::lrtest(w_fixed, w_slopes)
```

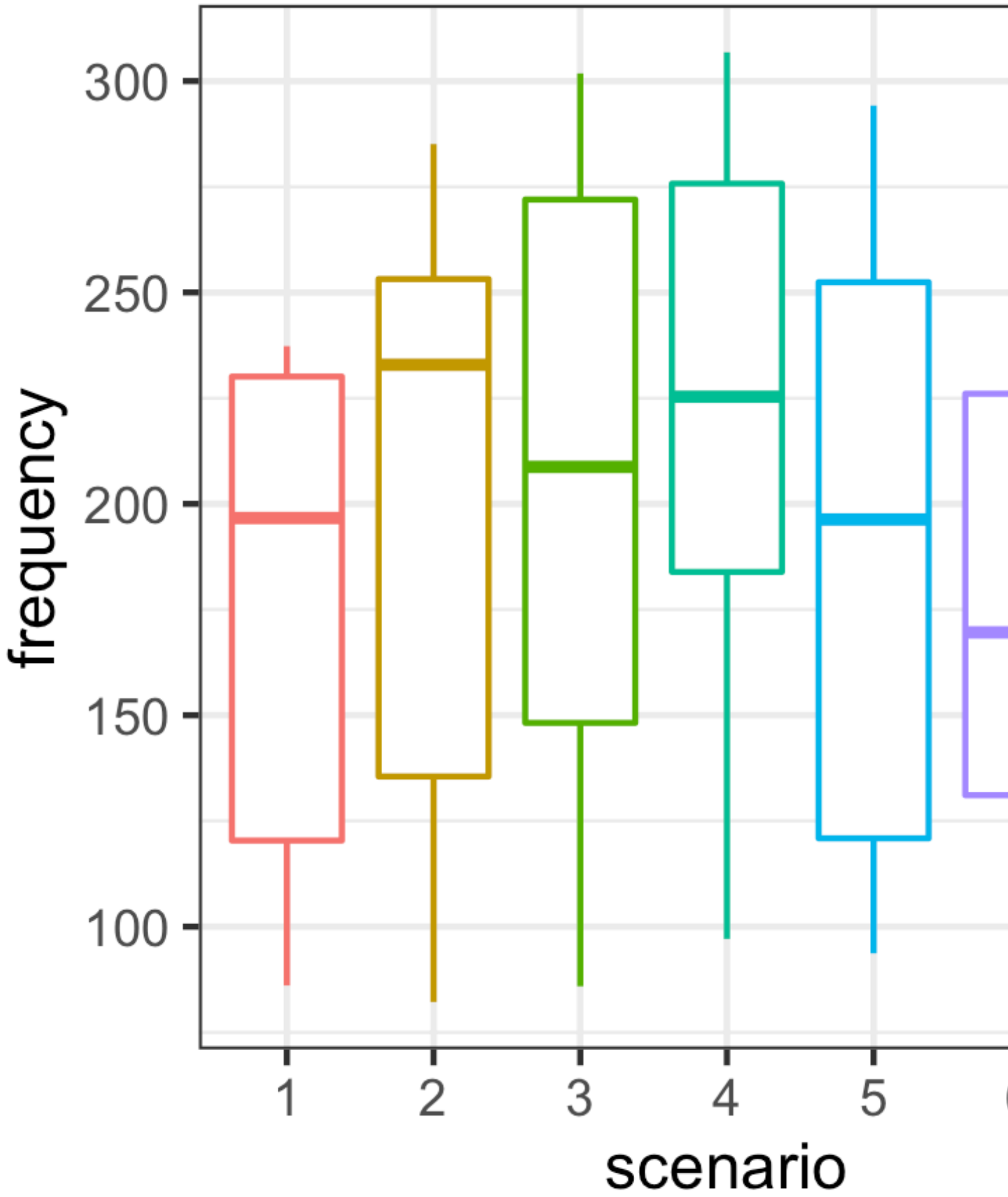
```
## Likelihood ratio test
##
## Model 1: frequency ~ condition + sex + (1 | subject)
## Model 2: frequency ~ condition + sex + (1 + condition | subject)
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   5 -397.9
## 2   7 -397.9  2 0.0119     0.9941
```

Note: We can happily do a likelihood ratio test because we have the same fixed effects and are only comparing nested random effects here. See the additional considerations section for why I mention this.

Conclusion from this test: The slopes are an unnecessary complication to our model. (I.e., No evidence against the hypothesis that the simpler model explains the data just as well.)

16.7.1.8 Scenario random intercepts

Does adding random intercepts for the scenarios improve the model?


300
250
200
150
100
frequency
1
2
3
4
5
scenario

```
w_scen = lmer(frequency ~ condition + sex + (1|subject) + (1|scenario), data=polite_data)
summary(w_scen)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: frequency ~ condition + sex + (1 | subject) + (1 | scenario)
##    Data: polite_data
##
## REML criterion at convergence: 784
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2717 -0.6326 -0.0833  0.5260  3.5191
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  scenario (Intercept) 222.0    14.90
##  subject  (Intercept) 614.4    24.79
##  Residual             637.0    25.24
## Number of obs: 84, groups:  scenario, 7; subject, 6
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)        237.054     16.101  14.723
## conditioninformal   19.864      5.507   3.607
## sexM              -108.660     20.974  -5.181
##
## Correlation of Fixed Effects:
##             (Intr) cndtnn
## condtnnfrml -0.171
## sexM        -0.651  0.000
```

Does adding random intercepts for the scenarios improve the model?

```
lmtest::lrtest(w_scen, w_fixed)
```

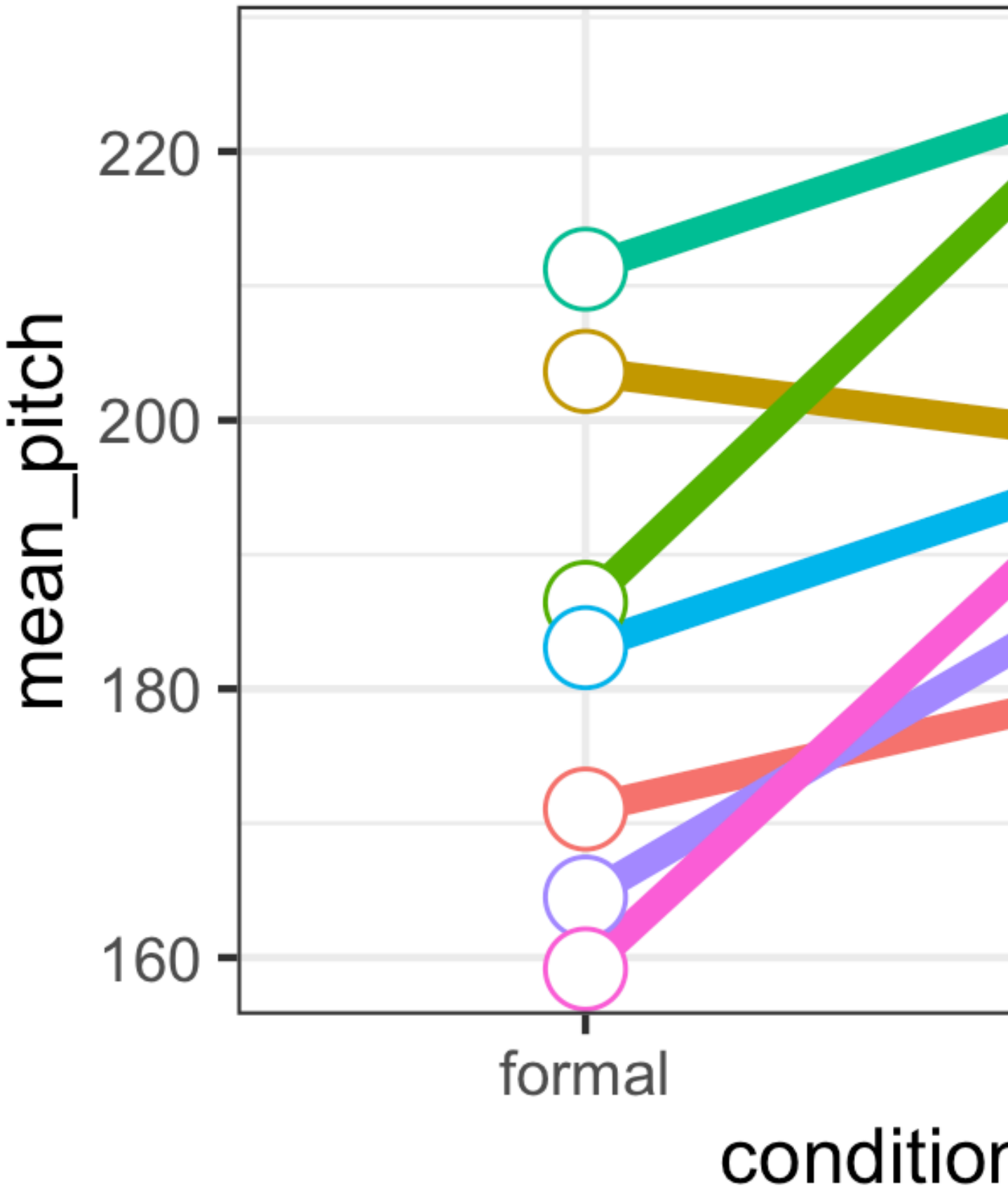
```
## Likelihood ratio test
##
## Model 1: frequency ~ condition + sex + (1 | subject) + (1 | scenario)
## Model 2: frequency ~ condition + sex + (1 | subject)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -391.98
## 2   5 -397.90 -1 11.851  0.0005765 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion from this test: Yes! Including a random effect for scenario does explain the data better. We have very strong evidence against the hypothesis that the simpler model fits the data just as well.

16.7.2 Scenario random slopes

Does adding random slopes for the scenarios improve the model?

```
w_scen_slope =
  lmer(frequency ~ condition +
                   sex + (1|subject) +
                   (1 + condition|scenario),
       data=polite_data)
```


220
200
180
160
mean_pitch
formal
condition

```
summary(w_scen_slope)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: frequency ~ condition + sex + (1 | subject) + (1 + condition |
##     scenario)
##    Data: polite_data
##
## REML criterion at convergence: 783.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1567 -0.6021 -0.1097  0.5033  3.4023
##
## Random effects:
##  Groups   Name              Variance Std.Dev. Corr
##  scenario (Intercept)       278.2    16.680
##           conditioninformal  72.3     8.503   -0.51
##  subject  (Intercept)       615.6    24.812
##  Residual                   618.6    24.873
## Number of obs: 84, groups:  scenario, 7; subject, 6
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)        237.054     16.342  14.506
## conditioninformal   19.864      6.308   3.149
## sexM              -108.660     20.973  -5.181
##
## Correlation of Fixed Effects:
##             (Intr) cndtnn
## condtnnfrml -0.244
## sexM        -0.642  0.000
```

Does adding random slopes for the scenarios improve the model?

```
lmtest::lrtest(w_scen, w_scen_slope)
```

```
## Likelihood ratio test
##
## Model 1: frequency ~ condition + sex + (1 | subject) + (1 | scenario)
## Model 2: frequency ~ condition + sex + (1 | subject) + (1 + condition |
##     scenario)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -391.98
## 2   8 -391.79  2 0.3755     0.8288
```

Conclusion from this test: We have no evidence against the claim that the model with the simpler random effects is as good a fit to the data as the model with the more complicated random effect structure.

16.8 Interactions between random effects and fixed effects

Recall the honey data from the Statdew Valley interactive.

Your grandmother was a bit eccentric when it came to farm management. She swore that her bees produced more honey if she played them music in the morning. In fact, she even mentioned that she thought they particularly liked K-pop.

She left you some data that she thinks proves this. While the sample size isn't that large, you decide to take a look anyways. The data is called `honey` and records the honey produced each month for each of her 6 hives (in grams). Each hive listened to each of the song options (Dynamite by BTS, Adagio in B minor by Mozart and No music) for a month, three times. The order was randomized for each hive. It took 9 months to collect this data and the information your grandmother left you says honey production should have been fairly constant over this period and that all the hives experienced the same weather conditions, food access, etc.

```
glimpse(honey, width = 80)
```

```
## Rows: 54
## Columns: 3
## $ hive    <int> 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 1, 1, 1,~
## $ song    <chr> "No music", "No music", "No music", "No music", "No music", "N~
## $ honey_g <dbl> 196.56, 199.58, 200.71, 195.80, 199.58, 200.71, 226.80, 227.55~
```

16.8.1 Model formula

I said the model formula would look like this:

$$y_{ijk} = \mu + \alpha_i + b_j + +(\alpha b)_{ij} + \epsilon_{ijk}$$

where y_{ijk} is the amount of honey produced (in grams) in the k^{th} month by the j^{th} hive while listening to song i. Here, μ is the grand mean of honey production, α_i are the I fixed effects for `song` and b_j are the random effects for hive j. $(\alpha b)_{ij}$ are the IJ interaction terms for the interaction between the the hive and the song. What does this represent here? Well, maybe some hives respond to a certain song quite differently to other hives. $(\alpha b)_{ij}$ is a random effect because any term involving a random effect must also be a random effect. $(\alpha b)_{ij} \sim N(0, \sigma^2_{\alpha b})$, $b_k \sim N(0, \sigma^2_b)$ and $\epsilon_{ijk} \sim N(0, \sigma^2)$. All the random effects are mutually independent random variables.

16.8.2 How would we fit this with `lmer`?

```
library(lme4) # library from which we load the lmer function
```

This is how we would fit a model with *only* a hive effect:

```
mod0 <- lmer(honey_g ~ song + (1|hive), data = honey)
```

This is how we would fit our **full** model, with a hive effect AND a hive and song interaction:

```
mod1 <- lmer(honey_g ~ song + (1|hive) + (1|hive:song), data = honey)
```

16.8.2.1 Do we need the interaction?

```
lmtest::lrtest(mod1, mod0)
```

```
## Likelihood ratio test
##
## Model 1: honey_g ~ song + (1 | hive) + (1 | hive:song)
## Model 2: honey_g ~ song + (1 | hive)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -175.66
## 2   5 -211.25 -1 71.199  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: We get a really small p-value, so we have strong evidence against the hypothesis that the simper model, with no song/hive interaction is as good as our full model.

16.9 More fixed vs random effects practice

Consider the following two examples from Roback & Legler (2021), section 7.10. **Identify which variable is the response and which are the fixed effects and the random effects.**

> Cockroaches! For a study of cockroach infestation, traps were set up in the kitchen, bathroom, and bedroom in a random sample of 100 New York City apartments. The goal is to estimate cockroach infestation levels given tenant income and age of the building.

Variables: room type, tenant income,apartment, building age, building, count of cockroaches

> Radon in Minnesota. Radon is a carcinogen – a naturally occurring radioactive gas whose decay products are also radioactive – known to cause lung cancer in high concentrations. The EPA sampled more than 80,000 homes across the U.S. Each house came from a randomly selected county and measurements were made on each level of each home. Uranium measurements at the county level were included to improve the radon estimates.

Variables: radon measurement, home, level of home, uranium, county

16.9.1 Answers

16.9.1.1 Cockroaches

- response: count of cockroaches
- fixed effects: room type (kitchen, bathroom, bedroom), tenant income, building age
- random effects: apartment, building

16.9.1.2 Radon

- response: radon measurement
- fixed effects: level of home (e.g. upper or lower level), uranium (county level variable)
- random effects: home, county

16.10 Advice for polishing your writing

The U of T writing advice site has some great short articles with tips about common writing mistakes and tips for improving your writing. The full list from the 'Revising' section is below. Choose one or two that seem relevant to you and read them. If you don't know where to start, go for the 'Hit Parade'.

- Revising And Editing
- Hit Parade Of Errors In Grammar, Punctuation, and Style
- Using the Computer to Improve Your Writing
- Wordiness: Danger Signals and Ways to React
- Unbiased Language
- Punctuation
- Fixing Comma Splices
- Faulty Parallelism
- Passive Voice: When to Use It and When to Avoid It
- Fixing Dangling Modifiers
- Some Tools and Rules to Improve Your Spelling
- Plurals
- Possessives
- Subject-Verb Agreement
- Sentence Fragments

And How Not to Plaigarize is always worth reviewing!

17

Module 4

Materials for February 28–March 11, 2022.

17.1 Learning checklist

- Distributions
  - Recognize a form of the probability density function for Bernoulli, binomial, Poisson, exponential, and gamma.
  - Identify how changing values for a parameter affects the characteristics of the probability distribution.
  - Identify the mean and variance for each distribution.
  - Match the response for a study to a plausible random variable and provide reasons for ruling out other random variables.
  - Match a histogram of sample data to plausible distributions.
- Odds, risks and logistic regression
  - Create tables and calculate joint, marginal and conditional probabilities with them.
  - Calculate odds, risks, odds ratios (OR), and risk ratios (RR).
  - Understand why ORs and RRs are similar for rare outcomes.
  - Interpret logistic regression output.
  - Differentiate between logistic regression models with binary and binomial responses.
  - Use the residual deviance to compare models, and to test for lack-of-fit when appropriate.
- GLM properties and unifying theory
  - State the assumptions of GLMs.
  - Identify appropriate modeling approaches to start with from a description of a study/data.
  - Determine if a probability distribution can be expressed in one-parameter exponential family form.
  - Identify canonical links for distributions of one-parameter exponential family form.
- Poisson regression
  - Conduct Poisson regression and interpret the coefficients.
  - Use and describes offsets, as appropriate.
  - Check for overdispersion.
  - Consider negative binomial, quasi-likelihoods, and zero-inflated extensions to Poisson regression.

17.2 Introduction

In this module, we will draw on knowledge of common distributions from your previous statistics courses, and then moving into logistics regression, and then broadening our horizons further with the family of models that come under the title of generalized linear models.

This big picture: A key assumption of linear regression you've worked through so far is that is that the residuals should be normal. With our current tools, we can't well account for binary data, probabilities (bounded [0, 1]), skewed distributions or even counts and in many cases.

17.3 Distributions (recap)

Much of this section should be recap of things you've learned in second-year statistics courses.

Reading: Chapter 3 (§ 3.3.1, 3.3.2, 3.3.4, 3.3.6, 3.4.2, 3.5) of Roback, P. & Legler, J. Beyond Multiple Linear Regression. (2021). https://bookdown.org/roback/bookdown-BeyondMLR/.

17.3.1 Sections to preview and skim

- 3.3.1 Binary Random Variable
- 3.3.2 Binomial Random Variable
- 3.3.6 Poisson Random Variable
- 3.4.1 Exponential Random Variable
- 3.4.2 Gamma Random Variable

17.3.1.1 Recommended revision

- 3.5 Distributions Used in Testing
- Honestly, all of Chapter 3 is a great resource for revising these distributions.

17.3.2 Reading guide

Try to answer the following for the selected distributions:

- What is the probability distribution function?
- What is/are the parameter(s)?
- How do changes to the the parameter(s) effect the response?
- What are the **mean** and **variance**?
- What values can your response variable take?
- When might you use this distribution? Come up with an example.
- What R code can you use to explore the density of this distribution?
- Can you simulate the distribution? Play with the parameters for yourself.

17.3.3 Cheat sheet template

I've made a template for a 'cheat sheet' on which you can take these notes/play with code. You can pull it on the JupyterHub with this link. From your Home directory, you will need to navigate to `sta303-w22-activities -> m4`. There are several Rmd you can play with in this file. The one for this activity is: `1A dist-cheat-sheet.Rmd`.

Table 17.1: COVID-19 hospitalizations in Ontario, by age group

Age	Hospitalizations	Population
0 to 09	98	1518527
10 to 19	92	1617937
20 to 29	402	2100175
30 to 39	668	2056056
40 to 49	1010	1876585
50 to 59	1932	2060937
60 to 69	2714	1795047
70 to 79	3151	1159898
80 to 89	3354	539715
90+	1645	139551

17.4 Ontario COVID hospitalizations

Suppose you're interested in hospitalizations by age group in Ontario due to COVID. This table also shows the estimated population in each age group.

This data is old, it was retrieved from https://www.publichealthontario.ca/en/data-and-analysis/infectious-disease/covid-19-data-surveillance/covid-19-data-tool?tab=ageSex on 2021-02-28.

17.4.1 Creating tables in R

Suppose our raw data has an observation per row.

```
# Read teh data in
ontario <- readRDS("data/m4/ontario_full.RDS")
head(ontario)
```

```
Rows: 14,864,428
Columns: 2
$ Age    <fct> 0 to 09, 0 to 09, 0 to 09, 0 to 09, 0 to 09, 0 to 09, 0 to 09, 0 …
$ Status <chr> "Hospitalized", "Hospitalized", "Hospitalized", "Hospitalized", "…
```

17.4.2 Creating tables

You can create a table with the `table()` function in R. (The `xtabs` function is also useful for creating cross (x) tables.)

```
hosp_table <- table(ontario$Age, ontario$Status)
hosp_table
```

17.4.3 Calculations with tables

From this kind of table there are three types of proportions that we can calculate.

- **Joint**
  - Joint proportions reflect the proportion total observation for which given levels of your categorical variables co-occur. I.e., what proportion of people were over 90 and hospitalized?

  - General calculation: Cell value over the grand total.

- **Marginal**

  - Marginal proportions sum across rows or columns. I.e., what is the proportion the Ontario population that has been hospitalized? We'd need to add up all the hospitalized and then divide that by the sum of all the cells.
  - General calculation: Row or columns sums over the grand total

- **Conditional**

  - Conditional proportions hold one variable level as given, it is a bit like zooming in to only one row or one column.
  - General calculation: Cell value over a row or column sum.

The `margin.table()` and `prop.table()` functions will be very helpful to us!

17.4.4 Joint probabilities

Calculate joint proportions:

```
hosp_table/sum(hosp_table)
```

```
##
##            Hospitalized Not hospitalized
##   0 to 09  6.592921e-06     1.021519e-01
##   10 to 19 6.189273e-06     1.088400e-01
##   20 to 29 2.704443e-05     1.412616e-01
##   30 to 39 4.493950e-05     1.382756e-01
##   40 to 49 6.794745e-05     1.261788e-01
##   50 to 59 1.299747e-04     1.385190e-01
##   60 to 69 1.825835e-04     1.205787e-01
##   70 to 79 2.119826e-04     7.781981e-02
##   80 to 89 2.256394e-04     3.608353e-02
##   90+      1.106669e-04     9.277585e-03
```

$$\frac{n_{ij}}{\Sigma n_{ij}}$$

17.4.5 Marginal probabilities

Proportion of people in each age group:

```
round(margin.table(hosp_table, margin = 1)/sum(margin.table(hosp_table, margin = 1)), 3)
```

```
##
##  0 to 09 10 to 19 20 to 29 30 to 39 40 to 49 50 to 59 60 to 69 70 to 79
##    0.102    0.109    0.141    0.138    0.126    0.139    0.121    0.078
## 80 to 89      90+
##    0.036    0.009
```

Proportion of people in each hospitalization status group:

```
round(margin.table(hosp_table, margin = 2)/sum(margin.table(hosp_table, margin = 2)), 3)
```

```
##
##     Hospitalized Not hospitalized
##            0.001            0.999
```

17.4.6 Conditional probabilities

Conditional on each row:

```
knitr::kable(prop.table(hosp_table, margin = 1))
```

	Hospitalized	Not hospitalized
0 to 09	0.0000645	0.9999355
10 to 19	0.0000569	0.9999431
20 to 29	0.0001914	0.9998086
30 to 39	0.0003249	0.9996751
40 to 49	0.0005382	0.9994618
50 to 59	0.0009374	0.9990626
60 to 69	0.0015119	0.9984881
70 to 79	0.0027166	0.9972834
80 to 89	0.0062144	0.9937856
90+	0.0117878	0.9882122

Conditional on each column:

```
knitr::kable(round(prop.table(hosp_table, margin = 2), 3))
```

	Hospitalized	Not hospitalized
0 to 09	0.007	0.102
10 to 19	0.006	0.109
20 to 29	0.027	0.141
30 to 39	0.044	0.138
40 to 49	0.067	0.126
50 to 59	0.128	0.139
60 to 69	0.180	0.121
70 to 79	0.209	0.078
80 to 89	0.223	0.036
90+	0.109	0.009

17.4.7 Risk and odds

"Risk" refers to the probability of the occurrence of an event or outcome. Statistically, risk = chance of the ou

From Common pitfalls in statistical analysis: Odds versus risk

17.4.8 Hospitalization risk and odds

Let's focus on folks aged 80 to 89.

```
          Hospitalized Not hospitalized
80 to 89         3354           536361
```

This **risk** of being hospitalized for this group is $\frac{3354}{3354+536361} = 0.0062$.

The **odds** of being hospitalized are $\frac{3354}{536361} = 0.0063$.

These values look fairly similar. Odds and risks **will** be similar when the outcome of interest is rare. This can be seen by the fact that the only difference between the two calculations is whether the count of the outcome is included in the denominator or not. As a rule of thumb, an outcome is 'rare' if it occurs less than 10% of the time.

17.4.9 Odds ratio and risk ratios

Risk ratios are also called 'relative' risks. Risk ratios and odds ratios are…ratios of risks and odds respectively.

They are used to make comparisons between groups. Let's for example, compare 80 to 89 year olds with 10 to 19 year olds.

```
          Hospitalized Not hospitalized
10 to 19           92          1617845
80 to 89         3354           536361
```

$$OR = \frac{3354/536361}{92/1617845} = 110$$

$$RR = \frac{3354/(3354+536361)}{92/(92+1617845)} = 109$$

Once again, these values are similar because being hospitalized is (thankfully!) rare.

17.4.10 When do we use RR vs OR?

Calculation of *risk* requires as to know how many people are 'at risk.' As we'll see next week, in case-control studies, where such totals are not available to us, we cannot calculate a relative risk. BUT, we can calculate odds ratios and make a comment on the strength of association between our exposure and the outcome.

In cohort studies, where we do have the number number exposed, we can calculate either/both.

Logistic regression, which we'll be seeing more of in the next few weeks, calculates adjusted ORs and not RRs and so being able to interpret them is going to be important to us.

17.5 Generalized linear models (GLMs)

Generalized linear models are a **flexible** class of models that let us *generalize* from the linear model to include more types of response variables, such as *count, binary, and proportion data.*

Image description: A feminine person in active wear holding a yoga pose where they are standing on their right foot, leaning forward and extending their right arm and holding the left leg out behind their back with their left arm.

Let's get flexible, flexible…

17.5.1 Assumptions of the Generalized Linear Model

- The data $Y_1, Y_2, ..., Y_n$ are independently distributed, i.e., cases are independent.
  - Thus errors are independent… but NOT necessarily normally distributed.
- The dependent variable Y_i does NOT need to be normally distributed, but it assumes a distribution, typically from an exponential family (e.g. binomial, Poisson, gamma,…)
- GLM does NOT assume a linear relationship between the dependent variable and the independent variables, but **it does assume a linear relationship between the transformed response (in terms of the link function) and the explanatory variables**; e.g., for binary logistic regression $logit(p) = \beta_0 + \beta_1 X$.
- The homogeneity of variance does NOT need to be satisfied.
- It uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate the parameters, and thus relies on large-sample approximations.

17.5.2 Components of a Generalized Linear Model

Generalized linear models have three parts:

1. **random** component: the response and an associated probability distribution

2. **systematic** component: explanatory variables and relationships among them (e.g., interaction terms)
3. **link function**, which tell us about the relationship between the systematic component (or linear predictor) and the mean of the response

It is the **link function** that allows us to generalize the linear models for count, binomial and percent data. It ensures the linearity and constrains the predictions to be within a range of possible values.

17.5.3 Generalized Linear Models

$Y_i \sim G(\mu_i, \theta)$
$h(\mu_i) = X_i^T \beta$

- G is the distribution of the response variable
- μ_i is a location parameter for observation i
- θ are additional parameters for the density of G
- h is a link function
- X_i are covariates for observation i
- β is a vector of regression coefficients

17.5.4 Ordinary Least Squares again

17.5.4.1 GLM

$$\begin{aligned} Y_i \sim & G(\mu_i, \theta) \\ h(\mu_i) = & X_i^T \beta \end{aligned}$$

17.5.4.2 OLS

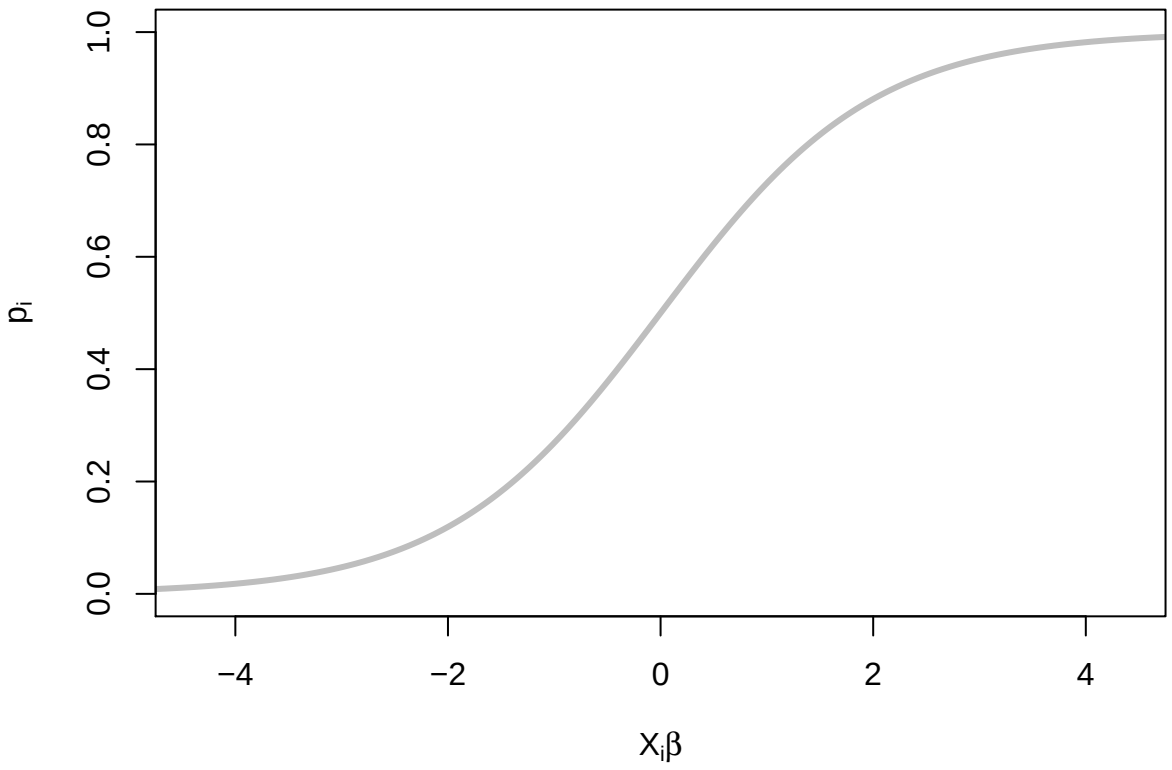
$$\begin{aligned} Y_i \sim & N(\mu_i, \sigma^2) \\ \mu_i = & X_i^T \beta \end{aligned}$$

OLS is just a flavour of GLM when:
- G is a Normal distribution
- θ is the variance parameter, denoted σ^2 - h is the identity function

17.6 Binomial (or logistic) regression

$$\begin{aligned} Y_i \sim & \text{Binomial}(N_i, p_i) \\ \log\left(\frac{p_i}{1-p_i}\right) = & X_i \beta \end{aligned}$$

- G is a Binomial distribution
- … or a Bernoulli if $N_i = 1$
- h is the logit link



- $X_i^T \beta$ can be negative
- p_i is between 0 and 1.

Let's look at an example...

17.6.1 Case study: Challenger disaster

The Challenger case study can be read and understood 'standalone,' but if you prefer having video commentary, this is the video I filmed last year. Not required to watch and please ignore any references to assessments from last year, etc.

Video from last year

17.6.1.1 Shuttle data

On January 28, 1986, the Space Shuttle Challenger broke apart 73 seconds into its flight, killing all seven crew members. The spacecraft **disintegrated** over the Atlantic Ocean. The disintegration of the vehicle began after a joint in its right rocket booster failed at liftoff. The failure was caused by the **failure of O-ring seals** used in the joint that were not designed to handle the unusually cold conditions that existed at this launch.

Image description: spiralling trail of white/grey smoke against a black background.

We will look at a data set about the number of rubber O-rings showing thermal distress for 23 flights of the space shuttle, with the ambient temperature and pressure at which tests on the putty next to the rings were performed.

17.6.1.2 Follow along with the case study

I've made a template for a 'cheat sheet' on which you can take these notes/play with code. You can pull it on the JupyterHub with this link. From your Home directory, you will need to navigate to `sta303-w22-activities -> m4`. There are several Rmd you can play with in this file. The one for this activity is: `1B challenger.Rmd`.

```
# This package is available in the JupyterHub, but you may need to install if working locally
# install.packages("SMPracticals")
data('shuttle', package='SMPracticals')
rownames(shuttle) = as.character(rownames(shuttle))
shuttle[1:4,]
```

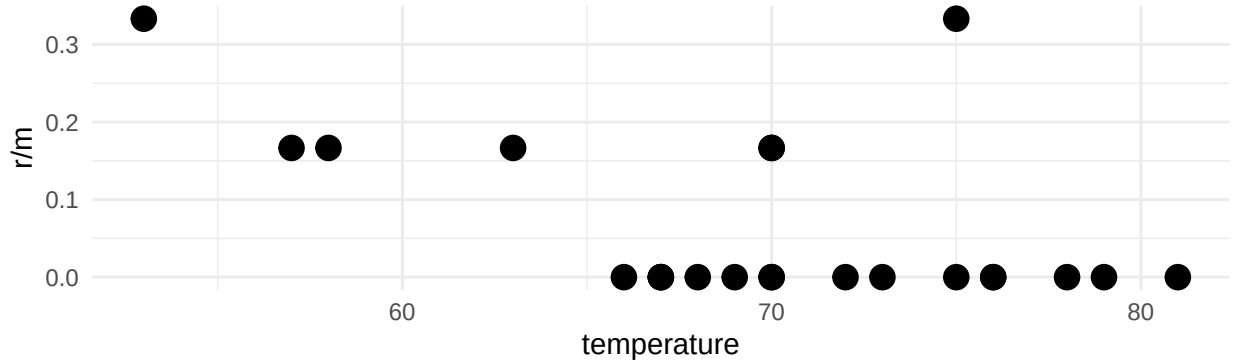
```
##   m r temperature pressure
## 1 6 0          66       50
## 2 6 1          70       50
## 3 6 0          69       50
## 4 6 0          68       50
```

- `m`: number of rings
- `r`: number of damaged rings

Thus we have a situation where we are interested in the number of successes out of a fixed number of trials. Hopefully your memories of the Binomial distribution are being triggered by that language.

```
# Base R plot
# plot(shuttle$temperature, shuttle$r/shuttle$m)

# ggplot
shuttle %>%
  ggplot(aes(x = temperature, y = r/m)) +
  geom_point(size = 4) +
  theme_minimal()
```



17.6.1.3 Are shuttle rings more likely to get damaged in cold weather?

We can think of **m** as the number of trials, and **r** as the number of "successes." (It feels weird to call damage a success, but it is our outcome of interest, so we treat it as such.)

$$\begin{aligned} Y_i \sim & \text{Binomial}(N_i, \mu_i) \\ \log\left(\frac{\mu_i}{1-\mu_i}\right) = & X_i\beta \end{aligned}$$

- m: number of rings, N_i
- r: number of damaged rings Y_i
- pressure, temperature: covariates X_i
- μ_i: probability of a ring becoming damaged given X_i
- $\beta_{\text{temperature}}$: parameter of interest

17.6.1.4 Inference: parameter estimation

$$\begin{aligned} Y_i \sim & G(\mu_i, \theta) \\ h(\mu_i) = & X_i\beta \\ \pi(Y_1 ... Y_N; \beta, \theta) = & \prod_{i=1}^{N} f_G(Y_i; \mu_i, \theta) \\ \log L(\beta, \theta; y_1 ... y_N) = & \sum_{i=1}^{N} \log f_G(y_i; \mu_i, \theta) \end{aligned}$$

- The Y_i are *independently distributed*
- **Joint density** π of random variables $(Y_1 ... Y_N)$ is the product of the marginal densities f_G.
- **Likelihood function** L given observed data $y_1 ... y_N$ is a function of the parameters.
- **Maximum Likelihood Estimation**:

$$\hat{\beta}, \hat{\theta} = \text{argmax}_{\beta,\theta} L(\beta, \theta; y_1 \dots y_N)$$

- The best parameters are those which are most likely to produce the observed data

17.6.1.5 Shuttle example in R

- `glm` works like `lm` with a `family` argument.
- Binomial models can take two types of inputs:

- If, as in this case, we have groups of trials, we need our response to be a matrix with two columns: `y` and `N-y`.
- If our `y` is a single 0/1 (or otherwise binary categorical variable) then we can set it up as usual, just a single column.

```
shuttle$notDamaged <- shuttle$m - shuttle$r
shuttle$y <- as.matrix(shuttle[,c('r','notDamaged')])
shuttleFit <- glm(y ~ temperature + pressure,
  family=binomial(link='logit'), data=shuttle)
shuttleFit$coef
```

```
##  (Intercept)  temperature     pressure
##  2.520194641 -0.098296750  0.008484021
```

17.6.1.6 Summary of fit

```
#Summarise the model
summary(shuttleFit)
```

```
##
## Call:
## glm(formula = y ~ temperature + pressure, family = binomial(link = "logit"),
##     data = shuttle)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0361  -0.6434  -0.5308  -0.1625   2.3418
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.520195   3.486784   0.723   0.4698
## temperature -0.098297   0.044890  -2.190   0.0285 *
## pressure     0.008484   0.007677   1.105   0.2691
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 16.546  on 20  degrees of freedom
## AIC: 36.106
##
## Number of Fisher Scoring iterations: 5
```

```
# Compute confidence intervals
confint(shuttleFit)
```

```
##                    2.5 %      97.5 %
## (Intercept) -4.322926283  9.77264497
## temperature -0.194071699 -0.01356289
## pressure    -0.004346403  0.02885221
```

There is no evidence that pressure is significantly associated with failure of O-rings…but how do we interpret these values?

17.6.1.7 Interpreting logistic models

$$
\begin{aligned}
Y_i \sim & \text{Binomial}(N_i, \mu_i) \\
\log\left(\frac{\mu_i}{1-\mu_i}\right) = & \sum_{p=1}^{P} X_{ip}\beta_p \\
\left(\frac{\mu_i}{1-\mu_i}\right) = & \prod_{p=1}^{P} \exp(\beta_p)^{X_{ip}}
\end{aligned}
$$

- μ_i is a probability
- $\log[\mu_i/(1-\mu_i)]$ is a log-odds
- $\mu_i/(1-\mu_i)$ is an odds
- If $\mu_i \approx 0$, then $\mu_i \approx \mu_i/(1-\mu_i)$

$$
\begin{aligned}
\beta_q = & \log\left(\frac{\mu_2}{1-\mu_2}\right) - \log\left(\frac{\mu_1}{1-\mu_1}\right) \\
\exp(\beta_q) = & \left(\frac{\mu_2}{1-\mu_2}\right) \Big/ \left(\frac{\mu_1}{1-\mu_1}\right)
\end{aligned}
$$

- β_q is the log-odds ratio
- $\exp(\beta_q)$ is the odds ratio
- exp(intercept) is the baseline odds, when $X_1 \ldots X_n = 0$.

17.6.1.8 Centring parameters

```
quantile(shuttle$temperature)
```

```
##   0%  25%  50%  75% 100%
##   53   67   70   75   81
```

```
quantile(shuttle$pressure)
```

```
##   0%  25%  50%  75% 100%
##   50   75  200  200  200
```

Currently the intercept is log-odds when temperature = 0 and pressure = 0. This isn't really a meaningful value for us to interpret, so we can center the covariate to a more meaningful baseline.

Let's centre the covariates so the intercept refers to:

- temperature = 70 (degrees Farenheit)

- pressure = 200 (pounds per square inch)

Table 17.2: MLEs of baseline odds and odds ratios, with 95intervals.

	est	2.5 %	97.5 %
Baseline	0.070	0.023	0.155
temperatureC	0.906	0.824	0.987
pressureC	1.009	0.996	1.029

```
shuttle$temperatureC <- shuttle$temperature  - 70
shuttle$pressureC <-  shuttle$pressure - 200
shuttleFit2 <-  glm(y ~ temperatureC + pressureC, family='binomial', data=shuttle)
```

17.6.1.9 Shuttle odds parameters

```
par_table = cbind(est = summary(
  shuttleFit2)$coef[,1],
  confint(shuttleFit2))
rownames(par_table)[1]= "Baseline"
```

```
round(exp(par_table),3) %>%
  knitr::kable(caption = "MLEs of baseline odds and odds ratios, with 95% confidence
intervals.")
```

17.6.1.10 Interpreting shuttle parameters

Because we centred our variables on meaningful values of temperature and pressure, our intercept is meaningful to interpret.

We can read from the table above that the odds of a ring being damaged when temperature = 70 and pressure = 200 is 0.07, which corresponds to a probability of:

```
round(exp(par_table[1,'est']) / (1+exp(par_table[1,'est'])), 3)
```

```
## [1] 0.065
```

We can only make this claim about probability for the intercept value!

$$P(\text{damage}|\text{temp} = 70, \text{pressure} = 200) = \frac{\beta_0}{1+\beta_0}$$

Each degree increase in temperature (in Fahrenheit) decreases the odds of damage by (in percent)

```
round(100*(1-exp(par_table[2,'est']) ), 3)
```

```
## [1] 9.362
```

17.7 GLMS: A unifying theory

17.7.1 Optional accompanying video

The following section can be read and understood be able to 'standalone,' but if you prefer having video commentary, this is the video I filmed last year. Not required to watch and please ignore any references to assessments from last year, etc. Might be most interesting/useful for the exponential family re-arrangements.

GLMS: A unifying theory

17.7.2 Readings

Reading: Chapter 5 of Roback, P. & Legler, J. Beyond Multiple Linear Regression. (2021). https://bookdown.org/roback/bookdown-BeyondMLR/.

Optional reading: Wood, S.N. Generalized Additive Models: An Introduction with R, 2nd Edition. 2017. https://ebookcentral-proquest-com.myaccess.library.utoronto.ca/lib/utoronto/detail.action?docID=4862399 (requires you to log in with your UTORid)

(The chapter in Wood covers more of the theory, but use what is covered in the slides as your guide to what is assessable.)

17.7.3 Recall

For a generalized model, we've said that:

$$Y_i \sim \text{some distribution}(\mu_i);\ g(\mu_i) = X_i^T\beta$$

17.7.4 Exponential family forms

The first model we met in the Generalized Linear Model family was logistic regression. But! there is a whole broader extended 'family' of these models that have a unifying trait:

their probability distribution function can be expressed in exponential form:

$$f_y(\theta) = e^{[\{y\theta - B(\theta)\}/A(\phi) + C(y,\phi)]}$$

where ϕ is an arbitrary 'scale' parameter and θ is know as the *canonical parameter* of the distribution. Functions A, B and C are arbitrary functions. You'll see lots of different letters and symbols used for these functions, I'm not sure which set to call the 'most' popular versions.

As you'll see in this module's reading, we can also set $\phi = 1$ (or it is by default 1 for Poisson and binomial) and talk about these as members of a 'one-parameter exponential family' where you can write the probability formula like this (θ is the 'one parameter'):

$$f(y;\theta) = e^{[a(y)b(\theta)+c(\theta)+d(y)]}$$

The canonical (also called "natural") parameter of a one-parameter exponential family is $\theta = g(\mu)$.

Exponentials with an extra parameter

The **Tweedie** and **negative binomial** distributions are also commonly employed. They can be expressed as members of the exponential family but with a single *additional* parameter. The negative binomial is often used when we have count data that is **overdispersed** relative to a Poisson distribution. I mostly only know Tweedie through its use by **actuaries** and have never used it in research myself.

17.7.4.1 Show that the normal distribution is in the exponential family

I chat through the steps in the video linked above.

We want to get from the Normal distribtion:

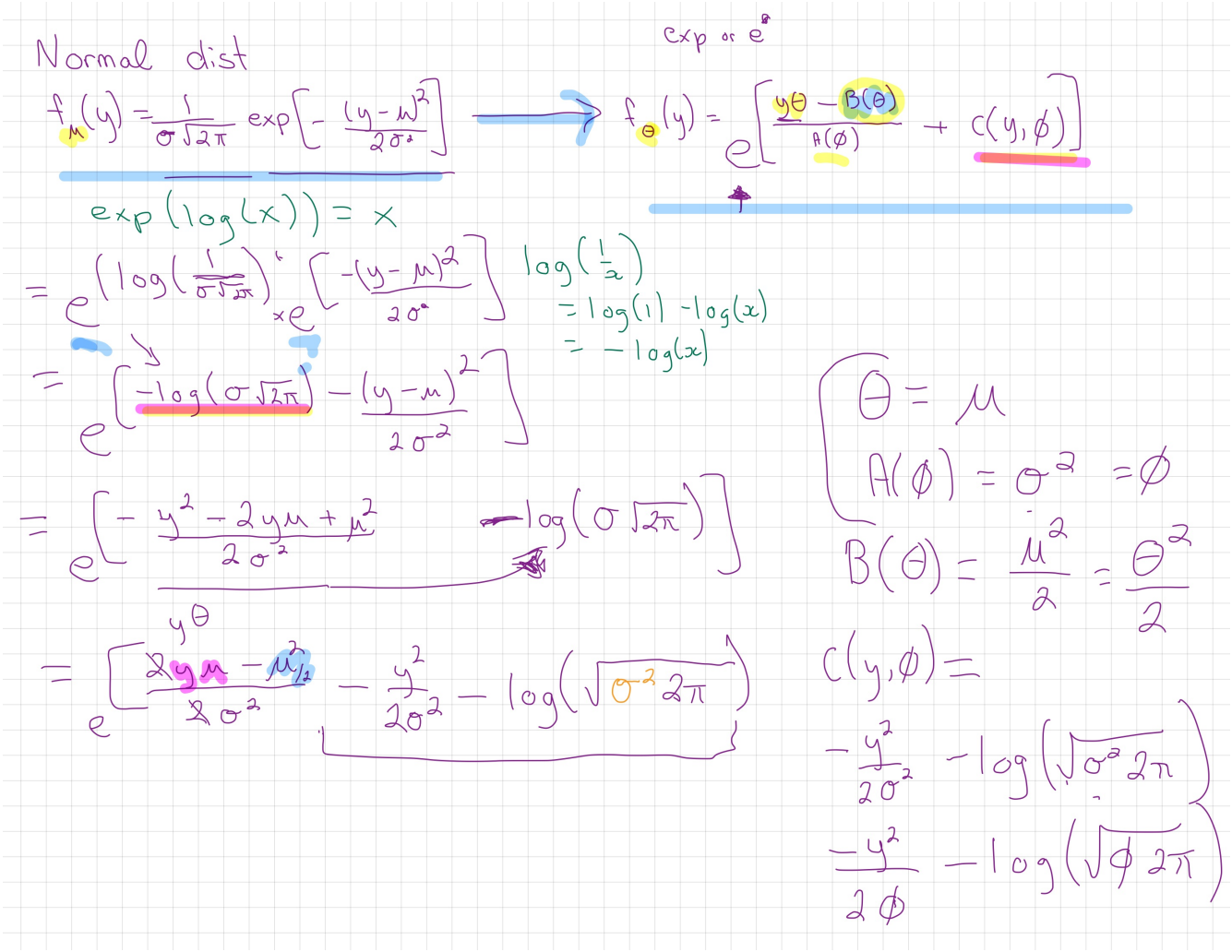
$$f_\mu(y) = \frac{1}{\sigma\sqrt{2\pi}}\exp[-\frac{(y-\mu)^2}{2\sigma^2}]$$

To our genreal exponential form:

$$f_\theta(y) = \exp[\frac{y\theta - B(\theta)}{A(\phi)} + c(y,\phi)]$$

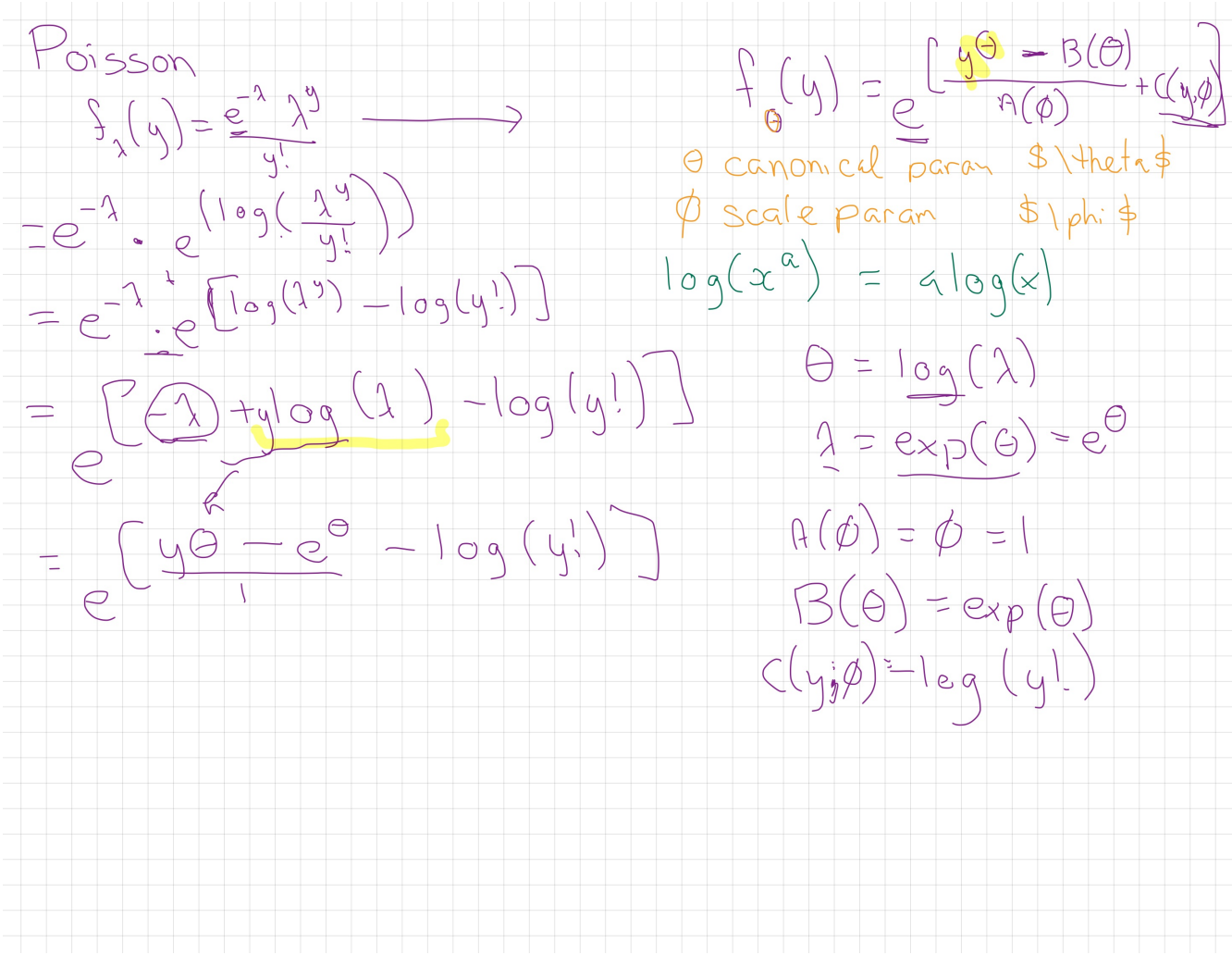
Note, this is the same as above, I've just tried to make it a little easier to read and focus on what is in the exponent.

Using the fact that $\exp(\log(x)) = x$, some algebraic rearrangements (and remembering our log rules!), and a knowledge of the final form we want to reach, we can show that the normal distribution is part of the exponential family.



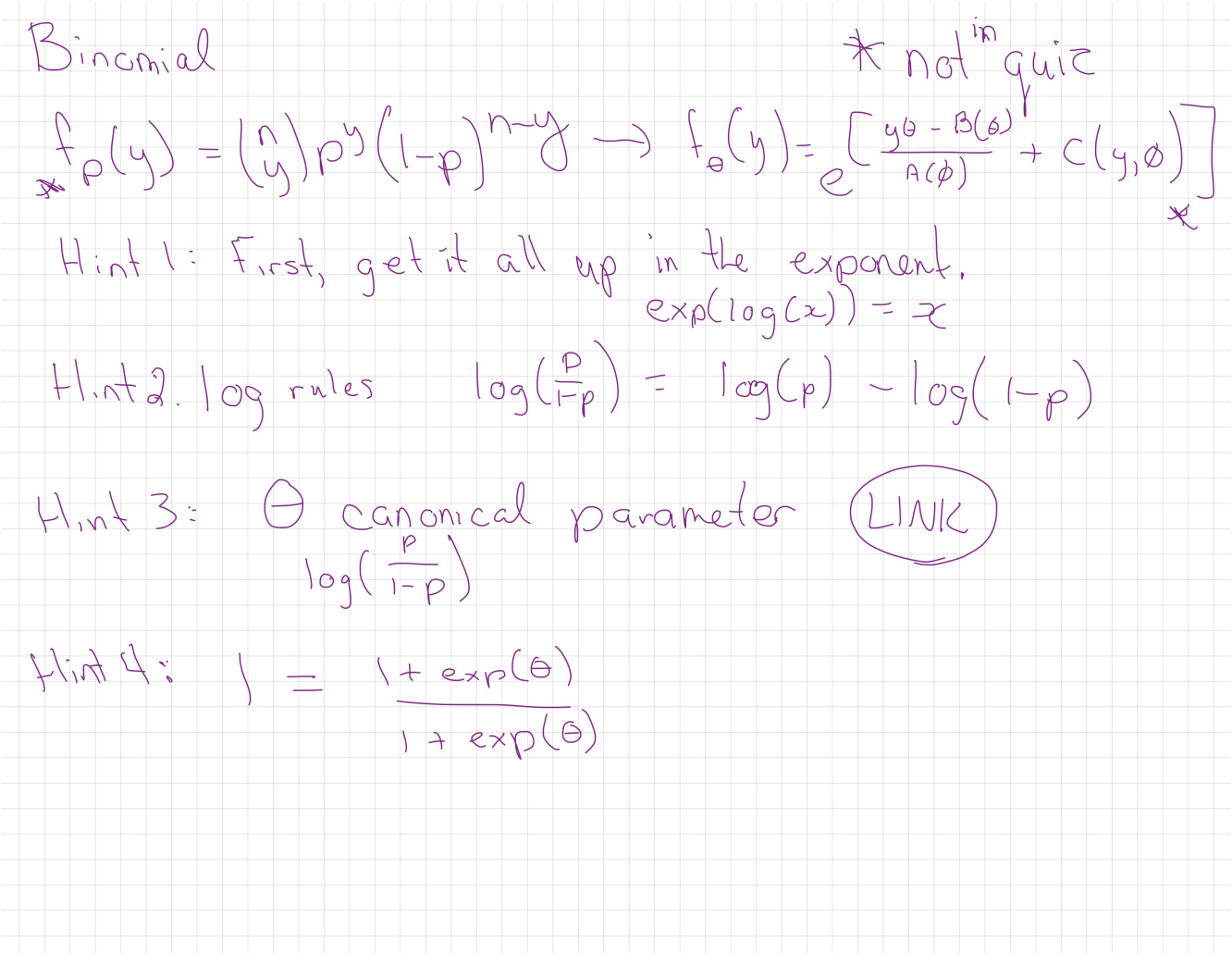
17.7.4.2 Show that the Poisson distribution is in the exponential family

I chat through the steps in the video linked above.



17.7.4.3 Your turn! Show that the binomial distribution is in the exponential family

I talk about some hints in the video linked above.



17.7.5 Canonical link functions

A link function, $g(\mu_i)$ is **canonical** if $g(\mu_i) = \theta_i$.

We've been using the canonical link for logistic regression, the logit link, where $g(\mu) = log(\frac{\mu}{n-\mu})$.

There are a few nice properties that come with using the canonical link:

- They ensure μ will be in the range the outcome variable can take (e.g. no probabilities greater than 1).
- Two methods for finding the MLE, Newton's method and the Fisher scoring method, will be identical when using the canonical link.
- Residuals will sum to 0, an 'observed unbiasedness.'
- The Hessian for the log-likelihood will be equal to its expected value.
- The Hessian is the matrix of second derivatives of the likelihood with respect to the parameters.

17.7.6 Exponential Family of Distributions

- Includes most of our familiar distributions:

  - normal, exponential, log-normal, gamma, chi-squared, beta, Bernoulli, Poisson, inverse Gaussian...and more!

- Who's *not* in the family?

  - Student's t-distribution, F-distribution, most distributions that are a mixture of others.

- Provides a unified theory for generalized linear models.

- A general, highly efficient method for finding MLEs numerically: Iterative weighted least squares (IRLS)

17.7.7 Iteratively re-weighted least squares algorithm (a sketch)

1. Start with an estimate of the mean, $\hat{\mu}_i = y_i + \delta_i$ (δ_i is usually 0, but sometimes it is a small constant to ensure that the $g(\mu_i)$ is finite).
2. Create some 'pseudodata,' z_i, using the values from the previous step and the original data and the associated weights (there are equations for all of this, of course, but they're not being covered in this class).
3. Find the value of $\hat{\beta}$ that minimizes the weighted least squares objective, $\Sigma_{i=1}^{n} w_i(z_i - \mathbf{X}_i\beta)^2$.
4. Update $\theta = X\hat{\beta}$ and $\hat{\mu}_i = g^{-1}(X_i\hat{\beta})$ and start at step 2 again.

17.7.7.1 How do decide when to stop?

We can judge that this algorithm has converged? One common way is checking the change in the negative log likelihood between each iteration and stopping once it is near zero (for a given value of 'nearness.')

17.7.8 A quick note on the large sample distribution of $\hat{\beta}$

From the properties of maximum likelihood, we have that: $\hat{\beta} \sim N(\beta, \mathcal{I}^{-1})$ or $\hat{\beta} \sim N(\beta, \hat{\mathcal{I}}^{-1})$

This isn't an exact distributional result, it relies on large sample approximations.

17.7.9 Deviance

It is nice to be able to make comments about GLMs in a similar way we are used to talking about the residual sum of squares in ordinary linear regression.

The **deviance** can take on this role for us. (In our R output it is called the *Residual deviance*.)

$$D = 2\{l(\hat{\beta}_{max}) - l(\hat{\beta})\}\phi$$

$l(\hat{\beta}_{max})$ is the maximum likelihood of the saturated model (super overfitting, one β per observation) and is the highest log likelihood we could possibly get with this data.

The **scaled deviance** is $D^* = D/\phi$. Once again relying on large sample results for likelihoods, we can say that if our model is true, $D^* \sim \chi^2_{n-p}$.

There is also the **null deviance**, D_0, which is the deviance between the null model (only an intercept) and the saturated model (perfect fit, one parameter per data point).

17.7.9.1 Comparing models

Suppose we're interested in testing: H_0: $g(\ ) = \mathbf{X}_{0\ 0}$ vs H_1: $g(\ ) = \mathbf{X}_{1\ 1}$

is the expected value of our response variable (and our observations are independent), and are from the same member of the exponential family of distributions and the models are nested. (I.e., $\mathbf{X}_0 \subset \mathbf{X}_1$.)

We can show, though won't here, that we can apply the likelihood ratio test as an expression of scaled model deviances, and that ϕ will cancel (nice!) so that this still works for us for GLMs where our scale parameter is not known.

A note on AIC from Simon Wood: *"In some quarters AIC has acquired cult status as the one true way to model selection, in a similar way to p-values, somewhat earlier. This is unfortunate nonsense. Like p-values, AIC is useful as a rough and ready quantitative aid to selecting between models, nothing more."*

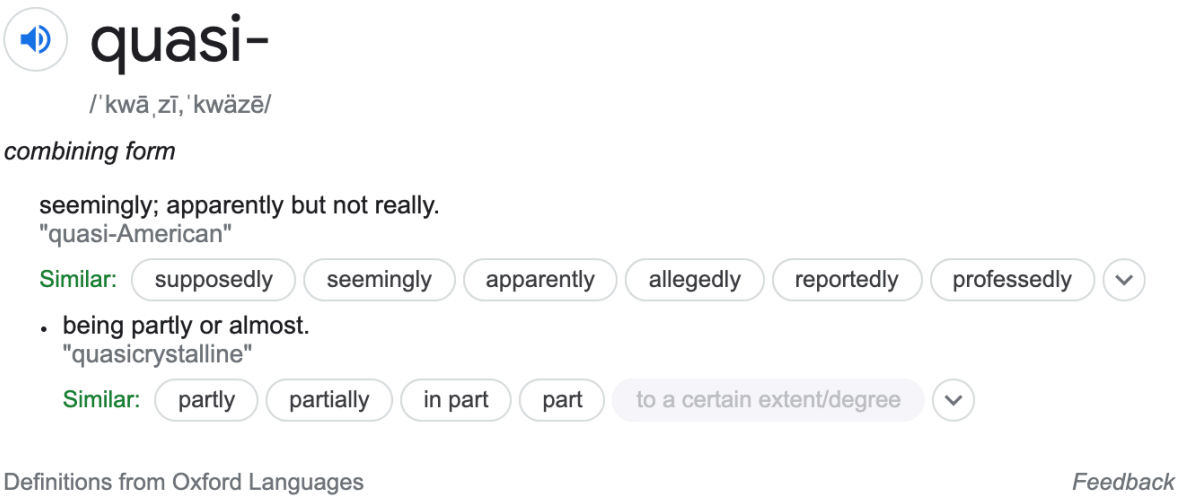
17.7.9.2 Quasi-likelihood

We've assumed so far that we know our response distribution to be a member of the exponential family, the whole unifying property of our GLM framework, so far.

BUT, sometimes we don't know the distribution quite so precisely. We may have a specific relationship between the mean and variance of the response in mind, but that's it.

Is all lost?

No! WE can actually get pretty far with the idea of **quasi-likelihood**.



Our main **practical interest** in this is that we can use it to model count data for which our variance is higher than we could get from the Poisson or binomial distributions.

Examples (family component of `glm`):
`family = quasibinomial(link = "logit")` `family = quasipoisson`

17.7.9.3 So what does this whole section mean for us in practice?

- We'll use the **canonical link** functions unless we have good reason to prefer something else (gamma is really the only example I tend to use a non-canonical link for).
- Under the hood, `glm` is making use of nice properties of our **exponential family** and using the iteratively re-weighted least squares algorithm.
- **Deviances** tell us about goodness-of-fit.
- In this course, we'll use **likelihood ratio tests** when comparing nested GLMs.
- We can (easily) use **quasi-likelihood** methods in situations where we want to 'loosen' up our distributional assumptions.

17.7.10 Reading: Logistic regression

Reading: Chapter 6 (§ 6.2, 6.4) of Roback, P. & Legler, J. Beyond Multiple Linear Regression. (2021). https://bookdown.org/roback/bookdown-BeyondMLR/.

This chapter should provide some good review of odds and logistic regression.

17.7.11 Logistic regression case study: Trying to lose weight

This case study is drawn from the content in Chapter 6 (§ 6.7) of Roback, P. & Legler, J. Beyond Multiple Linear Regression. (2021). https://bookdown.org/roback/bookdown-BeyondMLR/.

It differs from the **Challenger** example in that the response is binary (0/1) instead of a number of successes out of a given number of trials. This is a really common application of these methods.

17.7.11.1 Access the code for the case studies

You can pull the code on to the JupyterHub with this link. From your Home directory, you will need to navigate to `sta303-w22-activities -> m4`. There are several Rmd you can play with in this file. The ones for this reading `6.7 weightloss.Rmd` and `6.7 weightloss_ans.Rmd`.

17.8 Poisson regression key concepts + reading

17.8.1 Reading

You may wish to read the entire Poisson regression chapter, but I recommend you focus on these sections at least.

Reading: Chapter 4 (§ 4.2, 4.4, 4.6, 4.7, 4.8, 4.9, 4.10) of Roback, P. & Legler, J. Beyond Multiple Linear Regression. (2021). https://bookdown.org/roback/bookdown-BeyondMLR/.

17.8.2 Key concepts for Poisson regression

17.8.2.1 Use case

Counts, which can also extend to rates, i.e. counts per some unit of time or space.

17.8.2.2 Assumptions

- Poisson response
- Independence
- Mean = variance
- $log(\lambda)$ is a linear function of x

17.8.2.3 Model

$$Y_i \sim \text{Poisson}(\lambda_i)$$
$$\log(\lambda_i) = X_i\beta$$

Canonical link: log

- We interpret **coefficients** as **risk ratios**
  - Similar to logistic regression, we are interpreting the coefficients on the multiplicative scale after transforming with inverse of our link function.
- **Offsets**
  - Accounts for different denominators in rates, while allowing for counts to still be the response.
- Note: We're using 'log' here to mean the 'natural log,' i.e., base e. `log()` in R is also base e.
- **Overdispersion**
  - A key assumption of Poission regression is that the mean and variance of our response are equal. If the variance is greater than the mean we have overdispersion.
  - If we ignore overdispersion, our standard errors might be falsely small, meaning we'll probably have falsely small p-values, which might lead us to choose an more complicated model
- **Negative binomial regression**
  - This is another approach to dealing with overdispersion. As we've seen negative binomial isn't part of the one parameter exponential family, it requires a second parameter, but this gives us more flexibility AND assumes an explicit likelihood function (unlike quasipoisson).
- **Zero-inflated Poisson (ZIP)**
  - Some situations have two underlying processes, one that is Poisson and one that always produces zeros.

17.8.2.4 Zero-inflated Poisson situations


Suppose you're modelling the fish caught by people staying at a lakeside resort. The count of fish caught will always be zero for people who don't go fishing and sometimes zero for people who do.

Another common example arises with fertility data. If you have the count of live-births per person, for people who can give birth to a child, some will never have children, either by choice (e.g., through contraceptive use) or due to fertility/health status.]

17.8.3 Poisson regression case studies: (1) Household size in the Philippines and (2) Campus crime

17.8.3.1 Reading guides for the case studies

Read through these case studies, experiment with the code and pay particular attention to these questions.

17.8.3.1.1 Household size in the Philippines

- How do we interpret the coefficients?
- How do we compare models?
- What are the reasons why this model might not fit well?

You can ignore 4.4.5 and 4.4.8, if you wish.

17.8.3.1.2 Campus crime (includes 4.7, 4.8 and 4.9)

- Why would we use offsets in logistic regression?
- How do we include offsets in the model in R?
- What makes us concerned about our model assumptions?
- How does this example help us account for multiple testing?

17.8.4 Access the code for the case studies (optional)

You can pull the files on to the JupyterHub with this link. From your Home directory, you will need to navigate to `sta303-w22-activities -> m4`. There are several Rmd you can play with in this file. The ones for this reading `4.4 phillipines.Rmd` and `4.6 campus-crime.Rmd`.

17.9 Extra for the curious (NOT assessed)

The information on the following slides is **not assessed in this course**.

Consider taking a course like STA442 to go deeper! (Courses vary instructor to instructor, but the following is based on notes from Patrick Brown who has taught this course.)

17.9.1 Efficient maximization (for your reference only)

- Iteratively Reweighted Least Squares is the 'classic' algorithm when G is in the exponential family
- … but GLMs are easy for any density which is differentiable
- The derivatives with respect to β are easy to compute with the chain rule

$$\frac{\partial}{\partial \beta_p} \log L(\beta, \theta; y_1 \dots t_N) =$$

$$\sum_{i=1}^{N} \left[\frac{d}{d\mu} \log f_G(Y_i; \mu, \theta) \right]_{\mu = h^{-1}(X_i^T \beta)} \left[\frac{d}{d\eta} h^{-1}(\eta) \right]_{\eta = X_i^T \beta} \cdot X_{ip}$$

- Analytical expressions exist for the derivatives of $\log f_G$ and h^{-1}
- Second derivatives are also tractable
- Numerical maximization to find $\widehat{\beta}$ is fast when derivatives are available

17.9.2 Numerical maximizers (for your reference only)

- There are hundreds of them!
- `optim` is the standard `R` optimizer, which has 6 methods available.
  - some methods will use gradients if you provide them.
- `TrustOptim` uses derivatives and 'trust regions,' the method used in INLA.
- `ipopt` is probably the cutting edge.
- Statisticians don't make enough use of of-the-shelf optimizers.

17.9.3 Automatic differentiation (for your reference only)

$$\sum_{i=1}^{N} \left[\frac{d}{d\mu} \log f_G(Y_i; \mu, \theta) \right]_{\mu = h^{-1}(X_i^T \beta)} \left[\frac{d}{d\eta} h^{-1}(\eta) \right]_{\eta = X_i^T \beta} \cdot X_{ip}$$

- Overkill for most GLMs, but infinitely extensible.
- Computers evaluate logs, sines, and other functions through some Taylor-series-like polynomial thing… which are easy to differentiate!
- AD programs can take computer code and figure out how to differentiate it.
- Used in Neural Nets, Hamiltonian MCMC, optimization, and many more.

18

Module 5

Materials for March 21–April 1, 2022.

18.1 Introduction

What models have we looked at so far?

- recapped linear regression + model matrices
- linear **mixed** models (LMM)
  - for when our observations aren't independent due to some sort of grouping
- **generalized** linear models (GLM)
  - for situations where linear models don't work because of the response type

What's next?

- *generalized* linear mixed models (GLMM)
  - when we shouldn't use LMMs because of the response
  - and we shouldn't use GLMs because of grouping/repeated measures in our data
- *generalized additive models* (GAM)
  - when we want even more flexibility in the relationships between our predictors and response.

18.2 GLMMs

18.2.1 What were mixed effects, again?

We call these mixed effect models because they mix **fixed** effects (the way you're used to comparing the differences between treatments or effect of covariates) and **random** effects (generally experimental or observational blocks within which observations are grouped).

These types of models can be especially useful for people working with psychological, economic, social and ecological data.

18.2.2 Pros and cons of generalized linear mixed models

Pros

- Powerful class of models that combine the characteristics of generalized linear models (GLMs) and linear mixed models (LMMs).
- They can be used with a range of response distributions (e.g. Poisson, Binomial, Gamma).
- They can be used in a range of situations where observations are grouped in some way (not all independent).
- Fast and can be extended to handle somewhat more complex situations (e.g. zero-inflated Poisson).

Cons

- Some of the standard ways we've learned to test models don't apply.
- Greater risk of making 'sensible' models that are too complex for our data to support.

Note: GLMMs are still under **active** development by statisticians so not all the answers are know (even by the experts working on them).

18.2.3 Assumptions

18.2.3.1 Pause and think

Based on the assumptions for the previous models we've considered in this course, what do you think the assumptions for GLMMs will be?

18.2.3.2 The assumptions for GLMMs

- Our units/subjects are independent, even through observations within each subject are taken not to to be. (Replace subjects with 'groups,' same idea.)
- Random effects come from a normal distribution.
- The random effects errors and within-unit residual errors have constant variance. I.e., Are variances of data (transformed by the link function) homogeneous across categories?
- The chosen link function is appropriate / the model is correctly specified.

18.2.3.3 Pause and think

Here are the general model forms for the classes of models we've met so far:

Linear Model

$$Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$$
$$\mu_i = X_i\beta$$

Generalized Linear Model

$$Y_i \sim \mathrm{G}(\mu_i, \theta)$$
$$h(\mu_i) = X_i\beta$$

(Recall that MVN is multivariate normal.)

18.2.3.4 Questions

1. What are the similarities in each row?
2. What are the similarities in each column?
3. What do they all have in common?

18.2.4 Example: Bacteria in blood samples

```
# Data comes from the MASS package
data(bacteria, package='MASS')
head(bacteria)
```

```
##   y ap hilo week  ID     trt
## 1 y  p   hi    0 X01 placebo
## 2 y  p   hi    2 X01 placebo
## 3 y  p   hi    4 X01 placebo
## 4 y  p   hi   11 X01 placebo
## 5 y  a   hi    0 X02   drug+
## 6 y  a   hi    2 X02   drug+
```

- `y` is presence or absence of a certain bacteria in a blood sample
- `ID` is the subject the sample came from
- `ap` represents the treatment variable, with levels 'p' for placebo and 'a' for active.
- `week`, either 0, 2, 4, 6, or 11 indicates the number of weeks since the first test.

The question: does the treatment reduce the probability of bacteria in the sample?

The complication: data are not independent. Some people are more susceptible to bacteria/infection than others.

18.2.4.1 Bacteria model

$$\begin{aligned} Y_{it} \sim & \text{Bernoulli}(\rho_{it}) \\ \text{logit}(\rho_{it}) = & \mu + X_{it}\beta + U_i \\ U_i \sim & \text{N}(0, \sigma^2) \end{aligned}$$

- Y_{it} is presence of bacteria in individual i at time t
- X_{it} has indicator variables for week and treatment type.
- U_i is an individual-level *random effect*.
- $U_i > 0$ if i is more likely than the average to have the bacteria (allows for within-individual dependence)

This is a *Generalized Linear Mixed Model* (GLMM)

18.2.5 Inference for GLMMs

We know the probabilities of the Y_{it} conditional on the random effects:

$$pr(Y_{it}|U_i; \mu, \beta) = \frac{\exp(X_{it} + U_i)}{1 + \exp(X_{it} + U_i)}$$

If we knew the U_i, maximize the above to estimate $\hat{\beta}$

Integrate out the unknown U_i!

$$pr(Y; \mu, \beta, \sigma) = \int pr(Y|U; \mu, \beta) pr(U; \sigma) dU$$

writing $U = [U_1 \ldots U_N]$

Maximize the above to get $\hat{\mu}$, $\hat{\beta}$, $\hat{\theta}$

Plug-in parameter MLEs when predicting U_i

$$\text{E}(U_i|Y; \hat{\mu}, \hat{\beta}, \hat{\sigma})$$

18.2.6 Problems with Likelihood that affect our inference for GLMMs

~10 years ago

- We couldn't evaluate the likelihood, let alone maximize it.
- There were approximate methods which get around this:
  - Generalized Estimating Equations
  - Penalized Quasi-likelihood
  - Hierarchical Likelihood (Lee and Nelder)

- They could provide good parameter estimates and standard errors for the β but could be poor at finding $\text{var}(U_i|Y)$.
- Packages: `proc glmix` in SAS, `glmmPQL` in R.

Today

- Frequentist methods were later to arrive than Bayesian ones, but now they're here:
  - Laplace approximations,
  - importance sampling, and
  - automatic differentiation.
- Packages: `lme4`, `glmmTMB`, `prevmap`

18.2.7 Back to the bacteria

```
# How many children?
length(unique(bacteria$ID))
```

```
## [1] 50
```

```
# How many weeks?
table(bacteria$week)
```

```
##
##  0  2  4  6 11
## 50 44 42 40 44
```

```
# Were all children tested each week?
table(bacteria$week)/50
```

```
##
##    0    2    4    6   11
## 1.00 0.88 0.84 0.80 0.88
```

```
# How many in each group?
bacteria %>%
  group_by(ID) %>%
  slice(1) %>%
  group_by(ap) %>%
  summarise(n = n())
```

```
## # A tibble: 2 x 2
##   ap        n
##   <fct> <int>
## 1 a        29
## 2 p        21
```

18.2.7.1 Fitting a model with lme4

Create a new binary variable y:

```
bacteria$newy = as.integer(bacteria$y=='y')
```

Set treatment variable to placebo for everyone at week zero:

```
bacteria$ap[bacteria$week == 0] = 'p'
```

Set placebo to be the baseline:

```
bacteria$ap = fct_relevel(bacteria$ap, "p")
```

Run the model:

```
#install.packages("lme4")
bRes = lme4::glmer(newy ~ factor(week) + ap + (1 | ID),
        family='binomial', data=bacteria)
```

18.2.7.2 Parameter estimates

```
bRes = lme4::glmer(newy ~ factor(week) + ap + (1 | ID),
        family='binomial', data=bacteria)
summary(bRes)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: newy ~ factor(week) + ap + (1 | ID)
##    Data: bacteria
##
##      AIC      BIC   logLik deviance df.resid
##    206.1    229.8    -96.0    192.1      213
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0862  0.1328  0.2736  0.4342  1.2792
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  ID     (Intercept) 1.656    1.287
## Number of obs: 220, groups:  ID, 50
##
## Fixed effects:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.7988     0.6265   4.468 7.91e-06 ***
## factor(week)2    0.9561     0.8634   1.107   0.2682
```

```
## factor(week)4   -0.6407     0.7586  -0.845   0.3983
## factor(week)6   -0.7644     0.7648  -0.999   0.3176
## factor(week)11  -0.7834     0.7453  -1.051   0.2932
## apa             -1.2271     0.6041  -2.031   0.0422 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) fct()2 fct()4 fct()6 fc()11
## factor(wk)2 -0.410
## factor(wk)4 -0.578  0.610
## factor(wk)6 -0.591  0.604  0.706
## factr(wk)11 -0.605  0.612  0.711  0.713
## apa         -0.100 -0.471 -0.478 -0.469 -0.462
```

```
# VarrCorr lets us easily pull out the SD of the random effect
lme4::VarCorr(bRes)
```

```
##  Groups Name        Std.Dev.
##  ID     (Intercept) 1.2867
```

```
# Let's grad the fixed effects
round(summary(bRes)$coef, digits=2)
```

```
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)        2.80       0.63    4.47     0.00
## factor(week)2      0.96       0.86    1.11     0.27
## factor(week)4     -0.64       0.76   -0.84     0.40
## factor(week)6     -0.76       0.76   -1.00     0.32
## factor(week)11    -0.78       0.75   -1.05     0.29
## apa               -1.23       0.60   -2.03     0.04
```

18.2.7.3 Confidence intervals

- This takes a long time to run
- ...it is computing profile likelihoods so that we can estimate appropriate confidence intervals for our random (`.sig01`) and fixed effects.

```
(bConfint = confint(bRes))
```

```
## Computing profile confidence intervals ...
```

```
##                    2.5 %      97.5 %
## .sig01          0.5447208  2.27618243
## (Intercept)     1.7222831  4.23051462
## factor(week)2  -0.7149525  2.73645615
## factor(week)4  -2.1956116  0.83274632
## factor(week)6  -2.3382217  0.71406560
## factor(week)11 -2.3220296  0.65311207
## apa            -2.5352226 -0.07351393
```

Notice that were are reporting the standard deviation of the random effects, NOT the individual random effect for each person.

18.2.7.4 Aside: Plotting random effects

We focused a lot on the variance and standard deviation of our random effects, but these statistics are are OF something, namely the random effect values for each level of the random effect. We can see the random intercept associated with each subject in the plot below. Sometimes we might be interested in interpreting these values, but usually not as much as our fixed effect.

```
# ranef() allows us to pull out the random effects
lattice::dotplot(lme4::ranef(bRes, condVar=TRUE))
```

```
## $ID
```

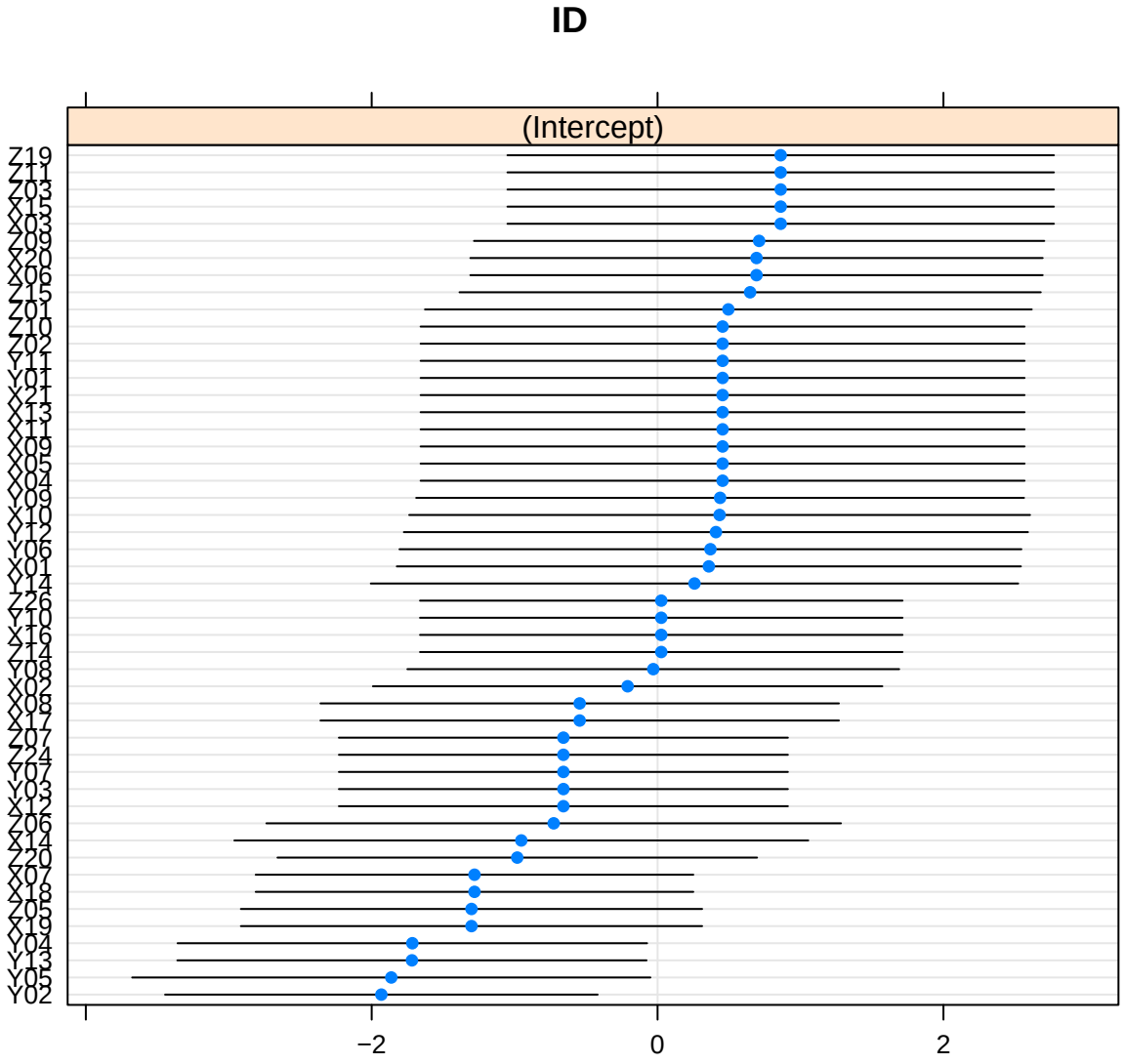


Figure 18.1: Value of the random intercepts for each subject in the bacteria study

18.2.7.5 Prepare a table for reporting

```
# Get the estimates transformed
ests <- format(round(exp(summary(bRes)$coeff)[,1], 2), nsmall = 2)
ests
```

```
##    (Intercept)  factor(week)2  factor(week)4  factor(week)6 factor(week)11
##        "16.43"        " 2.60"        " 0.53"        " 0.47"        " 0.46"
##            apa
##        " 0.29"
```

```
# Get your confidence intervals
cis <- format(round(exp(bConfint),2)[-1,], nsmall = 2)
cis
```

```
##                2.5 %   97.5 %
## (Intercept)    " 5.60" "68.75"
## factor(week)2  " 0.49" "15.43"
## factor(week)4  " 0.11" " 2.30"
## factor(week)6  " 0.10" " 2.04"
## factor(week)11 " 0.10" " 1.92"
## apa            " 0.08" " 0.93"
```

```
## But make it even prettier
cis_pretty <- str_c("(", trimws(cis[,1]), ", ", cis[,2], ")")
cis_pretty
```

```
## [1] "(5.60, 68.75)" "(0.49, 15.43)" "(0.11,  2.30)" "(0.10,  2.04)"
## [5] "(0.10,  1.92)" "(0.08,  0.93)"
```

```
# What are the nice names for the rows and columns?
rownames_for_table <- c("Baseline odds", "Week 2", "Week 4", "Week 6", "Week 11", "Active treatment")
colnames_for_table <- c("Estimate", "95% CI")

my_pretty_table <- cbind(ests, cis_pretty)
rownames(my_pretty_table) <- rownames_for_table
colnames(my_pretty_table) <- colnames_for_table
```

```
knitr::kable(my_pretty_table, align = c("r", "r"))
```

	Estimate	95% CI
Baseline odds	16.43	(5.60, 68.75)
Week 2	2.60	(0.49, 15.43)
Week 4	0.53	(0.11, 2.30)
Week 6	0.47	(0.10, 2.04)
Week 11	0.46	(0.10, 1.92)
Active treatment	0.29	(0.08, 0.93)

18.2.7.6 Pulling it all together: A short description of the methods (model written out previously)

This study investigates the association between an active treatment and the presence or absence of *H. influenzae* in children with otitis media in the Northern Territory of Australia. Children were randomly assigned to take either a placebo or drug after the baseline tests at week 0. 50 children were involved in the study (29 in treatment group, 21 in the placebo group) with blood tests conducted at 0, 2, 4, 6 and 11 weeks of treatment. Not all children were tested at all checks, but at least 80% were tested at any given check.

Our outcome of interest was the presence or absence of the bacteria and our predictors were week, treated as a factor, and whether the child was in that active treatment or placebo group. As there were repeated measures for each child, a generalized linear mixed model with logit link was used and random intercepts were estimated for each child.

Menzies School of Health Research 1999–2000 Annual Report. p.20. http://www.menzies.edu.au/icms_docs/172302_2000_Annual_report.pdf.

18.2.7.7 Pulling it all together: Results

The odds of there being bacteria in a blood sample at week 0 are roughly 16:1. Recall that all children are represented in the baseline as everyone was set as receiving the 'placebo' at week 0, i.e., no active treatment *before* treatment started.

Children receiving the treatment had 71% lower odds (95% confidence interval from 7 to 92% lower odds) of having bacteria present in the subsequent tests. There was no significant change in odds of finding bacteria in a child's blood test from week to week.

	Estimate	95% CI
Baseline odds	16.43	(5.60, 68.75)
Week 2	2.60	(0.49, 15.43)
Week 4	0.53	(0.11, 2.30)
Week 6	0.47	(0.10, 2.04)
Week 11	0.46	(0.10, 1.92)
Active treatment	0.29	(0.08, 0.93)

18.2.8 Generalized Linear Mixed Models, more generally

$$
\begin{aligned}
Y_i &\sim \pi(\lambda_i; \theta) \\
\lambda_i &= h(\eta_i) \\
\eta_i &= \mu + W_i\beta + U_i \\
U &\sim \text{MVN}[0, \Sigma(\theta)]
\end{aligned}
$$

- The bacteria model has
  - $\pi(\eta_i; \theta) = \text{Bernoulli}(\lambda_i)$
  - $\theta = \sigma$
  - $\Sigma(\theta) = \sigma^2 I$
  - $h(x) = \log(x)$
- The dimension of U and sometimes β is very large,
- Whereas typically the number of elements in θ is small.

(MVN is multivariate normal)

18.2.9 Some key conclusions

- GLMMs are easy to fit (thanks to Taylor series/Gaussian approximations, which we're not covering here)
- Interpreting GLMMs is similar to interpreting GLMs; we need to be thoughtful as link function makes interpretation subtle at times.
- Personal preference dictates whether to use `glmmTMB` or `lme4::glmer`.
- `glmmTMB` is faster on large datasets, but I find `lme4` approachable for beginners.
- Be Bayesian if you wish, but Frequentist GLMMs are a feasible and convenient option.

18.2.10 GLMM Reading

Reading: Chapter 11 of Roback, P. & Legler, J. Beyond Multiple Linear Regression. (2021). https://bookdown.org/roback/bookdown-BeyondMLR/.

This is a really interesting study that looks to update evidence from 2004/2005 that found that referees appeared to be 'evening out' foul calls across NCAA men's college basketball games. It also expands upon the original logistic regression that was applied to attempt to account for additional correlations in the data.

18.3 Case control studies and conditional logistic regression (OPTIONAL)

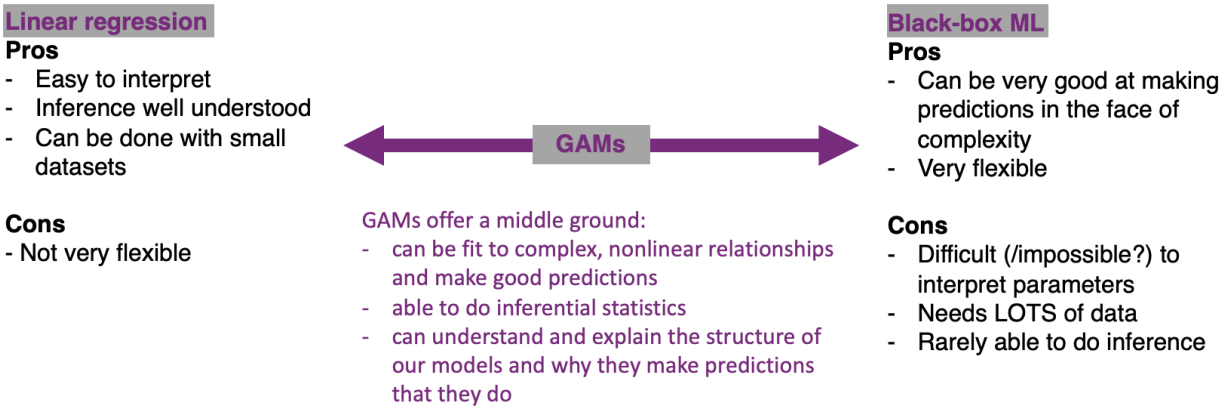
Based on what we're focusing on this semester, I am making the content about case control studies and conditional logistic regression optional.

If you're interested in this, you can check out the 28 minute video from last year, but it won't be assessed.

18.4 GAMs

The models you're probably most familiar with at this point in your statistics education, namely those from courses like STA302 and 303, are on one of an interpretability continuum. We can (with some statistics training) interpret what the coefficients for linear regression mean about our data and make inferences in well understood ways. They aren't the most flexible of models though. On the other side, there are our 'black-box' machine learning methods. These can be incredibly flexible and valuable tools for prediction, but in many cases difficult, if not impossible to interpret.

Generalized additive models (or GAMs) are a form of non-parametric modelling that allow us some of the flexibility of machine learning and some of the interpretability of our classic statistical methods.



Adapted from: Noam Ross https://noamross.github.io/gams-in-r-course/chapter1

You might also see the following discussed in relation to this class of models:

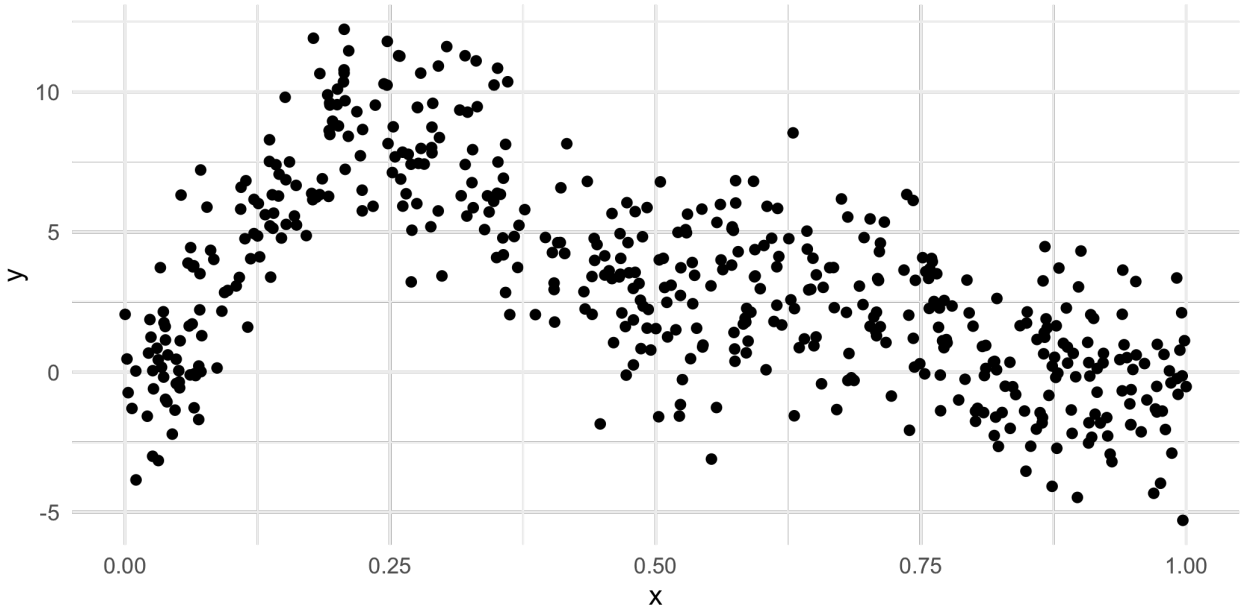
- Penalized likelihood
- Smoothing
- Fitting wiggly lines through points
- Semi-parametric models
- Splines

18.4.1 A one tweet GAM lesson

https://twitter.com/ucfagls/status/842444686513991680?ref_src=twsrc%5Etfw

18.4.2 Some fake data

It is very wiggly.

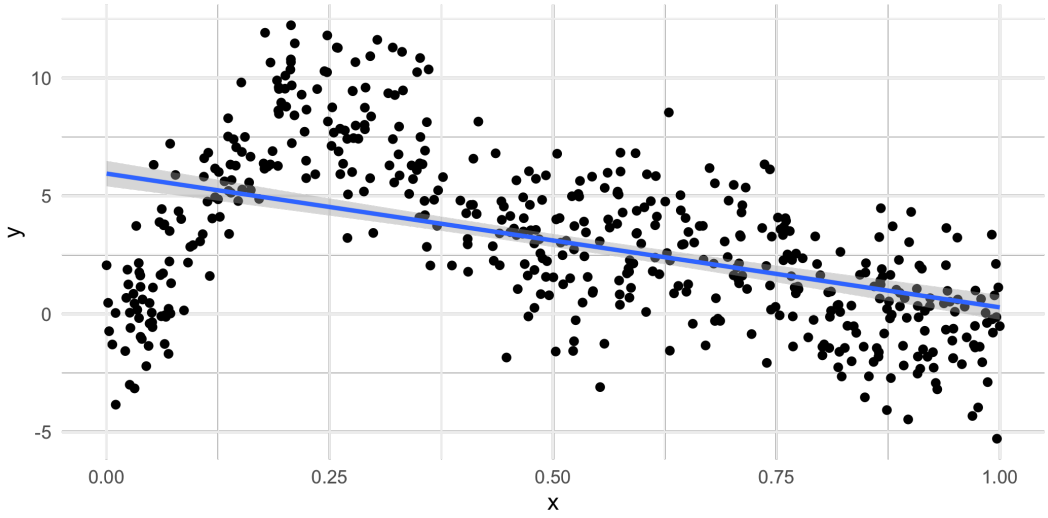


18.4.3 Linear model?

```
p + geom_smooth(method="lm")
```

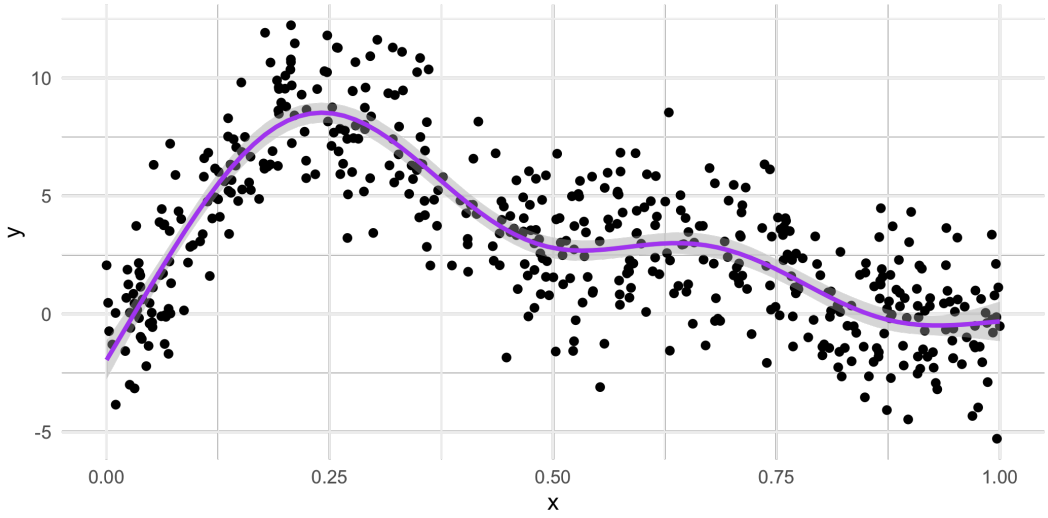
```
## `geom_smooth()` using formula 'y ~ x'
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Well, that seems bad…

18.4.3.1 Wouldn't something like this be much nicer?



But, HOW do we get the wiggles?

That is our goal for the rest of this module.

18.4.4 How do we get the wiggles?

Answer: Splines!

If you've looked at interference of waves in physics, you'll love this. If you haven't….you'll also love this!

GAMs are both *smooth* and *flexible* thanks to actually being made up of multiple not-as-flexible functions. *Imagine the Power Rangers robot teaching you a yoga class.*

- Each smooth is the sum of a number of **basis functions**
- Each basis function is multiplied by a coefficient
- Each of those coefficients is a parameter of our model

18.4.5 Splines

- Splines are *functions* composed of simpler functions
- Our simpler functions are *basis functions* & the set of basis functions is a *basis*
- When we model using splines, each basis function b_k has a coefficient β_k
- The resulting spline is the sum of these weighted basis functions, evaluated at the values of x

$$s(x) = \sum_{k=1}^{K} \beta_k b_k(x)$$

18.4.6 Picturing basis functions

- **Plot a** shows the basis functions of a GAM where all the coefficients are the same.

- **Plot b** shows the same basis functions *after* model-fitting, where each has a coefficient fit to the data.
- Basis functions add up to create the overall smooth shape.

Describing this one, nonlinear relationship (one response and one explanatory variable) requires several parameters, plus an intercept.

Image created by Noam Ross.

18.4.7 Wiggle, wiggle, wiggle

There is a GIF in the web version. Link: https://github.com/gavinsimpson/intro-gam-webinar-2020/blob/master/resources/basis-fun-anim.gif?raw=true.

GIF by Gavin Simpson

18.4.7.1 Taking a peak at our coefficients

```
library(mgcv) # you will need to install this
gam_mod <- gam(y ~ s(x, k=7), data=my_data, method="REML")
coef(gam_mod)
```

```
## (Intercept)      s(x).1      s(x).2      s(x).3      s(x).4      s(x).5
##    3.074352  -10.512816   14.800850   -5.780337   -3.959907   18.597941
##      s(x).6
##   10.047384
```

This is just some meaningless fake data, we'll work through a Case Study more fully.

18.4.7.2 How many basis functions do we want?

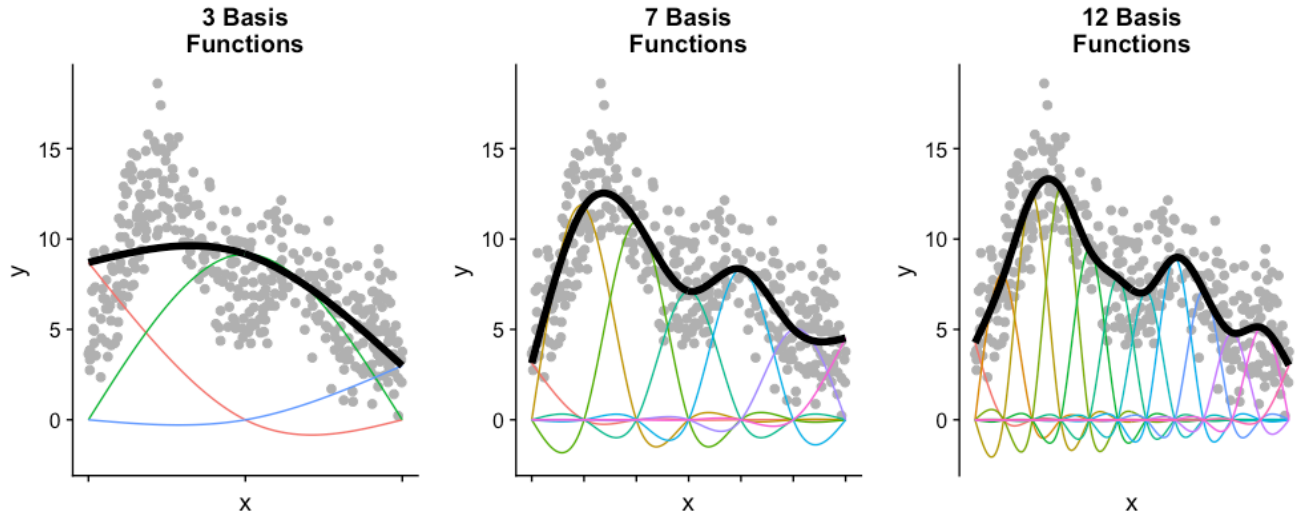


Image created by Noam Ross.

18.4.7.3 Smoothing

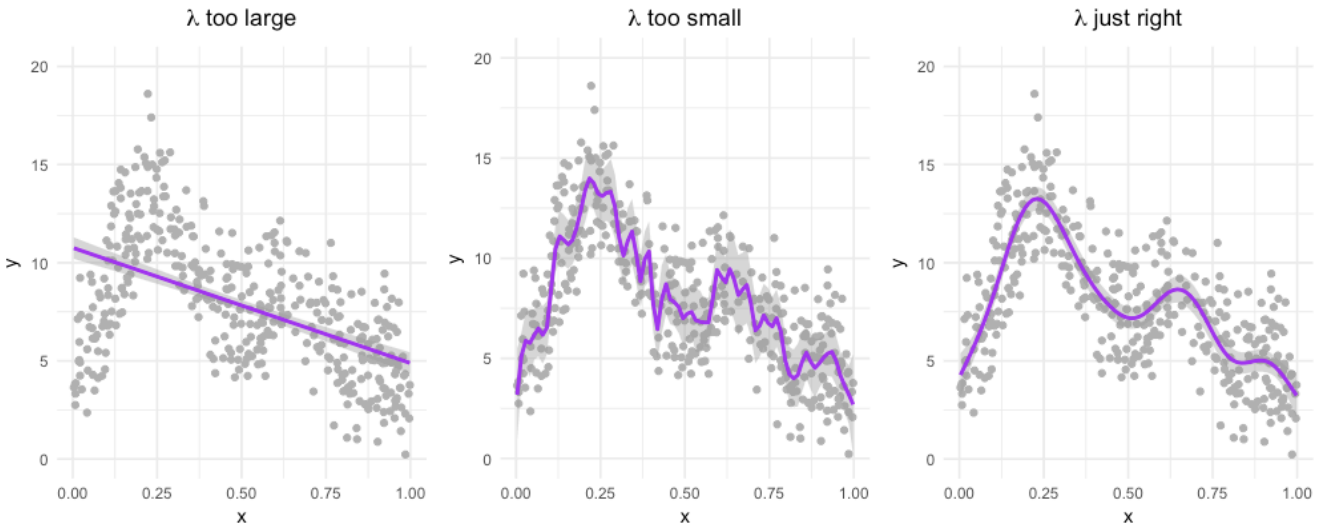


Image created by Noam Ross.

18.4.7.4 Smoothing

You can think of our fit as being:

$$\text{penalized log lik} = \log(\text{Likelihood}) - \lambda \cdot \text{Wiggliness}$$

where λ is a smoothing parameter.

We can set lambda with the `sp` (smoothing parameter) option in `gam()` BUT it is recommended that we let R find the best one for us using restricted maximum likelihood ("REML").

```r
# Sets parameter for the whole model
gam_mod <- gam(y ~ s(x), data=my_data, sp=0.1)

# Set the parameter for one specific term
gam_mod <- gam(y ~ s(x1, sp=0.1) + s(x2), data=my_data)

# Let R do it for you - the recommended way
gam_mod <- gam(y ~ s(x), data=my_data, method = "REML")
```

18.4.8 Choices to make

18.4.8.1 Wiggliness

There are LOTS of ways to pick your wiggle: AIC, generalized cross-validation (GCV, part of the name of `mgcv` and the default for that package), ML and REML.

18.4.8.2 Basis complexity

We can set a maximum wiggliness by setting a 'size,' k, that indicates a maximum number of small functions that could be used to build the model. If we set it bigger than the data, we'll get an error, and if we set it much bigger than needed, it is computationally costly.

Our effective degrees of freedom (edf) will always be less than k. We can check if we've been sensible in our choice of k with `gam.check()`.

18.4.8.3 Basis expansions

In the polynomial models we used a polynomial basis expansion of x

- $x^0 = 1$ — the model constant term
- $x^1 = x$ — linear term
- x^2
- x^3
- ...

So! If the **effective degrees of freedom** we need for a term is approximately 1, then we're really just smoothing it down to a linear term, the way a covariate would usually enter a model as a fixed effect in our previous models. We may choose to just put it in the model as such, so that we can interpret the coefficient it receives.

18.4.9 Generalized additive (mixed) models

We can combine everything we've done in this course so far into generalized additive models (including adding random effects).

$$\begin{aligned} Y_i \sim & G(\mu_i, \theta) \\ g(\mu_i) = & X_i\beta + Z_iU + f(W_i) \end{aligned}$$

- Y_i are responses
- G is the response distribution
- X_i, Z_i and W_i are covariates

- U are our random effects
- $f(w)$ is some sort of wiggly line
- If we put no restrictions or assumptions on f, the estimate $\hat{f}(w)$ will interpolate the data perfectly (which isn't very interesting)

Random effects

When fitted with REML or ML, *smooths* can be viewed as just fancy *random effects*. AND, excitingly, random effects can be viewed as smooths!

If your random effects are fairly simple, you can fit those in `mgcv::gam()` without needing the more complex GAMM functions, like `gamm4::gamm4()`

These two models are equivalent:

```
# library(mgcv)
# You don't need to know anything about this data,
# just consider how that variables enter the models
m_nlme <- lme4::lmer(travel ~ 1 + (1 | Rail), data = Rail, REML = TRUE)
m_gam  <- gam(travel ~ s(Rail, bs = "re"), data = Rail, method = "REML")
```

18.4.10 Random effects

The random effect basis, `bs = 're'`, is *not* as computationally efficient as *lme4* if we have complex random effects terms or even if we just have random effects with many levels (which isn't really that unusual with random effects).

Instead we could use `gamm()` or `gamm4::gamm4()`:

- `gamm()` fits using `lme()`
- `gamm4::gamm4()` fits using `lmer()` or `glmer()`

I.e., you're wanting a response with a conditional distribution that isn't *normal*, use `gamm4::gamm4()`

18.4.11 Case studies: Cherry trees and Portugese larks

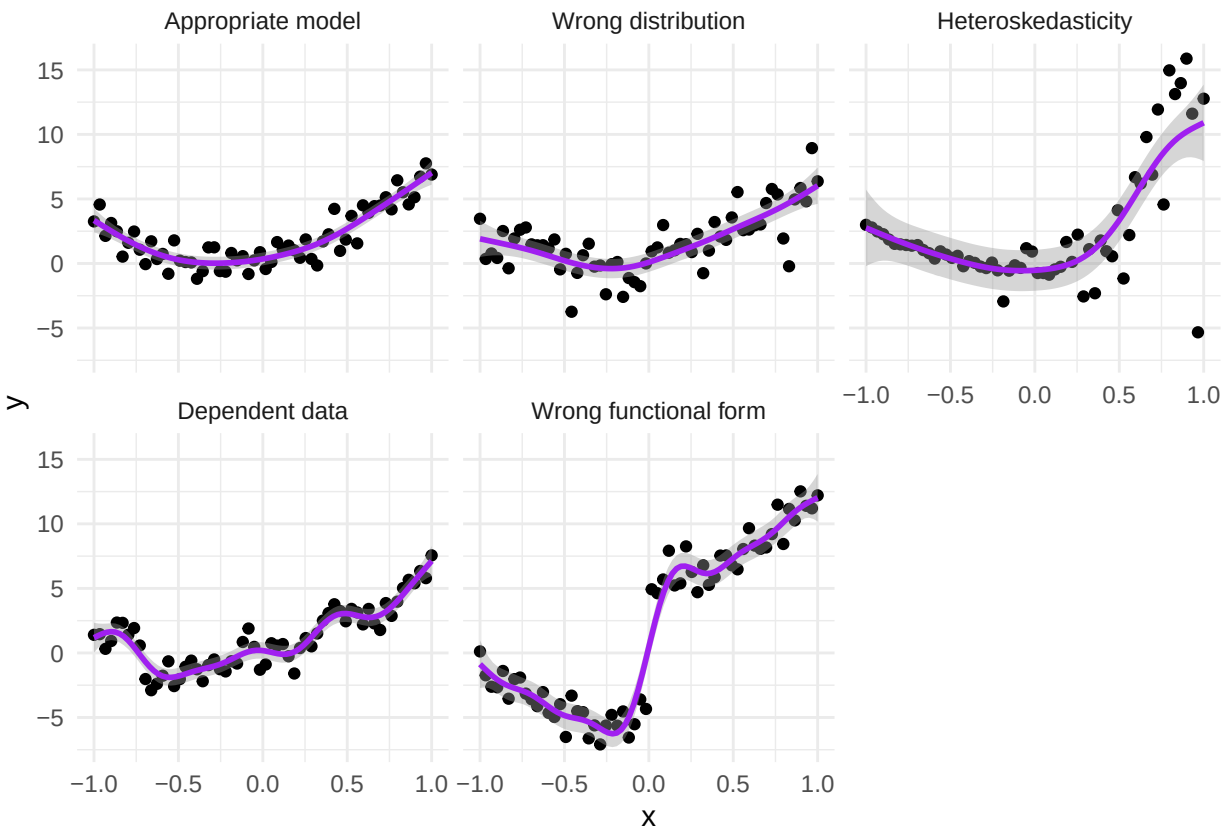
18.4.11.1 Access the code for the case studies

You can pull the code on to the JupyterHub with this link. From your Home directory, you will need to navigate to `sta303-w22-activities -> m5`.

18.4.12 Further comments on GAMs

18.4.12.1 GAMs are models too

How accurate your predictions are depends on how good the model is, as always. (Credit: Eric Pedersen & Gavin Simpson)



18.4.12.2 Variable selection

Unmodified smoothness selection by GCV, AIC, REML etc. will not usually remove a smooth from a model (not set edf to 0). Most smoothing penalties view the null-space of a smooth as 'completely smooth' and so further penalization does not change it.

With `select = TRUE` we add an extra penalty to the null-space part (the part of the spline that is perfectly smooth).

If you don't have this, smoothness selection can usually only penalize a smooth back to a linear function

(because the penalty that's doing smoothness selection only works on the non-smooth (the wiggly) parts of the basis). To perform selection we need to be able to penalize the null space (the smooth parts of the basis) as well.

18.4.12.3 Smoothness selection

The `method` argument to gam selects the smoothness selection criterion. For many practitioners, 'ML' or 'REML' are their default choice, though not the default in `gam()`. Using a likelihood based approach essentially treats the smooth components as random effects.

18.4.12.4 More conditional distributions than you can shake a squiggly line at

A GAM is just a fancy GLM! So we can fit any of the models we've learned (Poisson, Logistic, Gamma as we saw today in the cherry example). The creators of the `mgcv` package (Simon Wood & colleagues (2016)) have extended the methods to some non-exponential family distributions that are also very helpful, of which we've seen Negative Binomial and Zero-inflated Poisson.

- `binomial()`

- `poisson()`
- `Gamma()`
- `inverse.gaussian()`
- `nb()`
- `tw()`
- `mvn()`
- `multinom()`
- `betar()`
- `scat()`
- `gaulss()`
- `ziplss()`
- `twlss()`
- `cox.ph()`
- `gamals()`
- `ocat()`

18.4.12.5 A symphony of smoothers

The type of smoother is controlled by the `bs` argument (think *bas*i*s*)

The default is a low-rank thin plate spline `bs = 'tp'`

Many others available (thanks Gavin Simpson for making this list):

- Cubic splines `bs = 'cr'`
- P splines `bs = 'ps'`
- Cyclic splines `bs = 'cc'` or `bs = 'cp'`
- Adaptive splines `bs = 'ad'`
- Random effect `bs = 're'`
- Factor smooths `bs = 'fs'`
- Duchon splines `bs = 'ds'`
- Spline on the sphere `bs = 'sos'`
- MRFs `bs = 'mrf'`
- Soap-film smooth `bs = 'so'`
- Gaussian process `bs = 'gp'`

18.4.12.6 How do we talk about GAMs?

Presenting results from GAMs is similar to presenting results from other models we've learned except that for smoothed terms we have no single coefficient you can make inference from (i.e. negative, positive, effect size etc.).

For smoothed variables, we rely a lot on visual methods (e.g. `plot(gam_model)`) for describing our results and we can also make inference from predicted values.

For parametric variables, we can make inferences like we normally would.

GAMs are especially useful for accounting for a non-linear phenomenon that may not be the main thing you are interested in. This is similar to how we have already used random effects to account for correlation in our data that is not the main thing of interest but shouldn't be ignored.

18.4.13 Conclusions

- GAMs are GLMMs
- Anything you can do with LMs/LMMs/GLMs/GLMMs you can do with GAMs
  - All the different ways we learned to deal with response functions for GLM and GLMM apply here too
- use ML to estimate parameters
- use a lot of knots

18.4.13.1 Other GAMs things we won't get to look at in detail in this course

- changing basis functions (see a list of options by running `?smooth.terms`)

References

Appendix

19

Resources

19.1 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.

19.1.1 Admin

Logo	Description
QUERCUS UNIVERSITY OF TORONTO	Quercus will be used for timed assessments, some submissions and announcements.
zoom	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
F	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.

Logo	Description
	Markus will be used for the submission of code and analysis documentation for several assessments.
	peerScholar will be used for the Create, Asses and Reflect phases of the module writing and peer review tasks. You can access the peerScholar links through the associated Quercus assessments.

19.2 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-jupyerhub-uoft.netlify.app.

19.3 Team Up!

Link to student-facing Quercus Support resources.

Please only attend one of these sessions if you are prepared to...

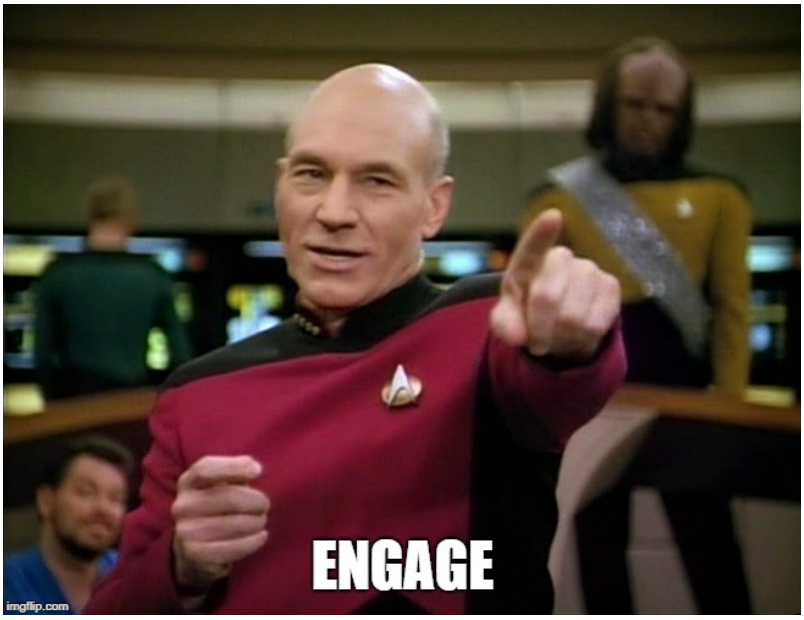


Image description: Captain Picard (of the Starship Enterprise) in uniform, pointing forwards. Text reads: ENGAGE

19.3.1 To participate in these activities, you need:

- A U of T LICENSED Zoom account to access the call. Some troubleshooting tips in this announcement: Make sure your Zoom is licensed!.
- A reasonable **internet connection**.
- A **working microphone.**
- To be WILLING TO ACTIVELY PARTICIPATE. Please don't ruin this for others who are interested in engaging in meaningful group problem-solving.
  - There is **no reason to come and lurk**. These sessions are the activity and nothing else.
  - You can easily get marks for your knowledge basket in SOOO many other ways AND I will make an asynchronous practice version available so you don't miss out on practice.

19.3.2 Before the activity you should:

1. Remember that **you don't have to participate in Team Up! activities**, there are plenty of other ways to earn your knowledge basket marks.
   - Only participate if you really are going to participate (see above).
2. Check you can access **Shiny interactives** on the JupyterHub: Module 1 interactive.
   - If you get a 'Not Found' error this is most likely due to one of two issues:
     - Solution 1: Ensure you don't have any other interactives open. (Close any tabs with the orange-arcs JupyterHub logo).
     - Solution 2: Go to https://jupyter.utoronto.ca, open a new RStudio session by going New -> RStudio (see slide 3 of these instructions) and move any .Rmd files in your Home directory (see slide 10 of these instructions) to a new folder (or delete them if you don't need them). Remember: No naked Rmds in YOUR home!

3. Decide if you want to be **randomized into a group** or **work with a group of people you already know in the course** (section/tut enrolment doesn't matter).
    - If **randomizing**, join the Zoom link for the tutorial time you are attending. 'Arrive' on time as latecomers may not be placed in a group.
    - If **working with your own group**, you don't need to join the call (but can if you get stuck with the tech!). Decide at which time you want to do the activity and set up a personal Zoom/Teams/WeChat/Instagram/carrier pigeon/semaphore chat. Zoom has nothing to do with Team Up!.
4. Read over the Team Up! instructions below.

19.3.3 Some rules/logistics

- **Speak up!** Practice talking about statistics. For these activities, I'd honestly rather you be wrong out loud than right in silence.
- **Be kind** and constructive. You are in a team, practice being a good team member.
- Team Up! does not have a great back end. If you don't submit correctly or don't join a group correctly, I CANNOT go in and check or move you around. **FOLLOW INSTRUCTIONS**.
    - No accommodations will be made for Team Up! issues. Team Up! marks will not be updated by hand (e.g., if you end up in a solo group/tech issues, etc.).
- You must do this as a group activity, during a tutorial time, if you wish to receive a knowledge basket grade.
    - **A 'group' is 2–4 people.**
    - All members of groups of 1 (solos) and 5+ will receive a zero. Group size is based on what is **recorded by Team Up!** through your registration with your group codes.
- You cannot do a Team Up! more than once. You cannot attend both sessions.
- If you have a ghost in your room (randomized in the class organized Zoom call), please 'Ask for help' (there is a button). 'Raise hand' *doesn't* alert me if you're in a breakout room.

Timing note: Assessment pages will become available at 15 minutes past the hour and remain available for 45 minutes. Ensure sure your team submits by the end of the session.

19.3.4 Team Up! Instructions

Participating in Team Up! activities can earn you knowledge basket points. They will be based on how well your team does, so work together and give it a good effort.

You will complete the following 5 steps: .midi[1. **Introductions** (first/personal name + a one-word description of how your week is going so far). Use this time to confirm everyone understands the instructions. 2. **Driver** sets up Team Up! group 3. **Members** join Team Up! group 4. Complete the activity 5. **Submit** the activity

19.3.4.1 Step 1: Decide on roles

You will need:

- An **Instructions Master** who is responsible for helping the team follow these instructions carefully.
- A **Screensharer** who is responsible for sharing the activity on their screen.

- A **Driver** for the Team Up! side of the activity. The Driver acts as team leader, responsible for navigating the quiz, confirming the team's final agreed upon answer to a question, and submitting the completed quiz score to Quercus.The Driver has a steering wheel next to their ID and is differently coloured than other Team Members.

Team Members (other than the Driver) participate and choose vote on answers, so all team mates can see what they have chosen, but they cannot submit their choices to Team Up!.

DECIDE THIS BEFORE CLICKING ANYTHING TO DO WITH TEAM UP!

19.3.4.2 Step 2: Driver sets up group

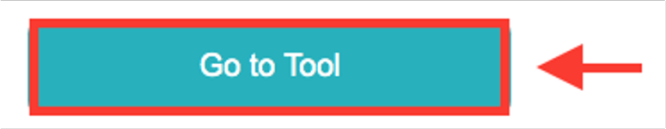
The person in your group whose first/personal name comes last alphabetically will be the Leader (Driver) for the session.

The Driver must do the below steps FIRST to get the group code. Then they will share the group code with all the members of the group so they can also join.

1 Log-in to Quercus from your device (phone, tablet, computer).

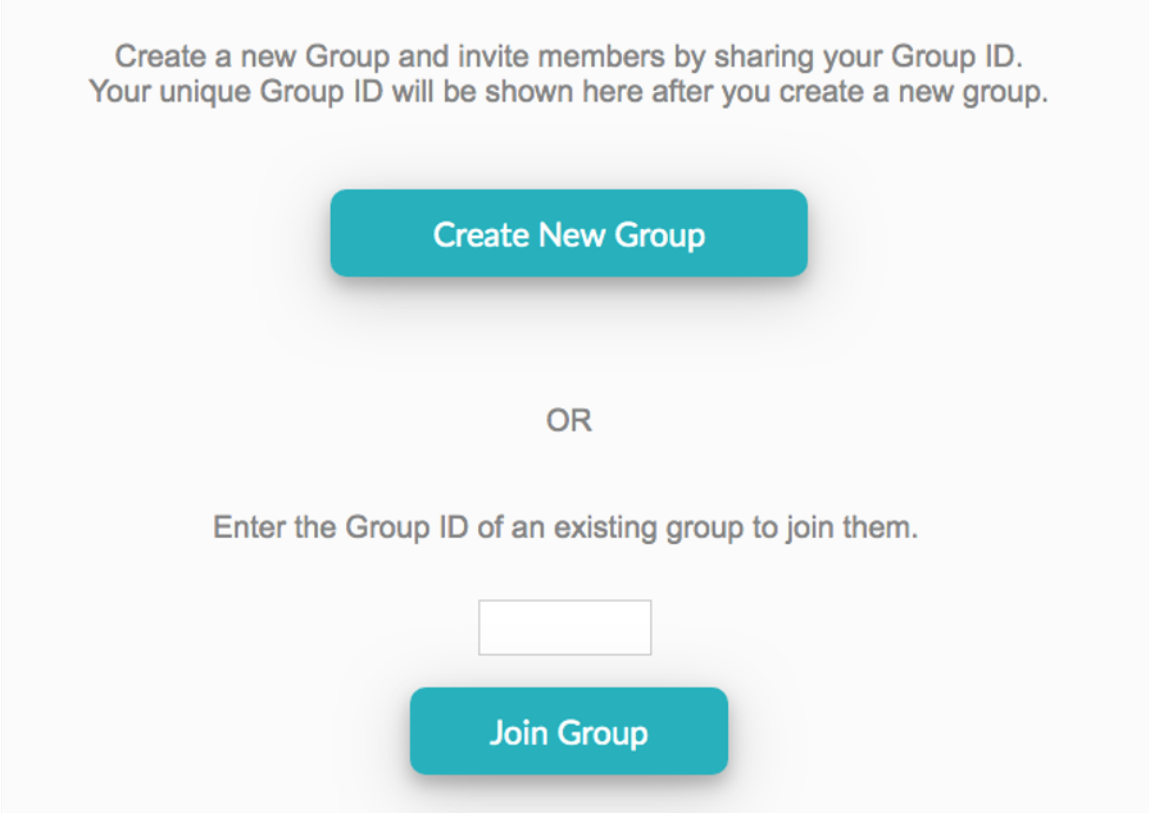
1 Click on the Team Up! assignment from the Course Overview page.

2 Click 'Go to Tool.' If it is not appearing, refresh your page or try a different browser.

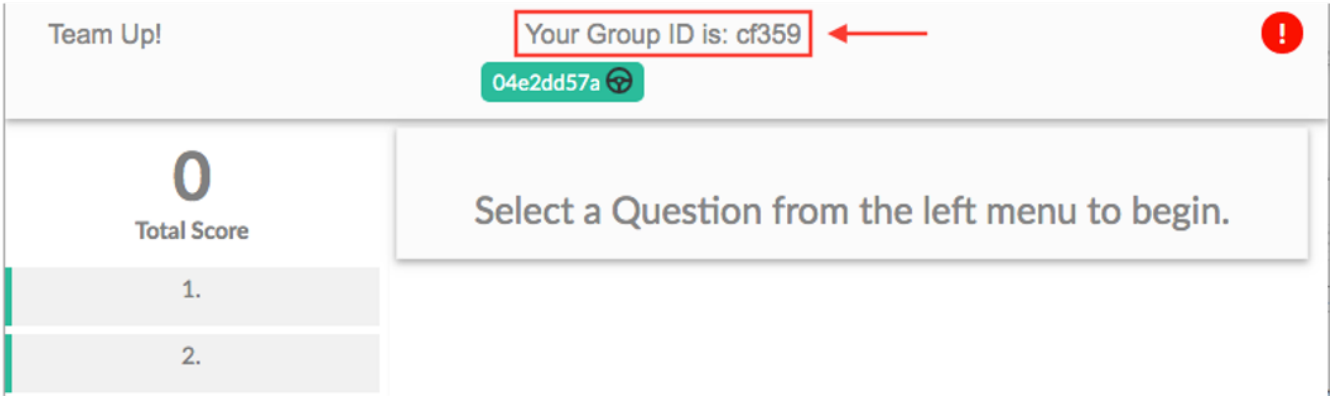


NON-DRIVERS! DO NOT TOUCH ANYTHING!

3 Choose "Create New Group"



4 Share group ID in Zoom

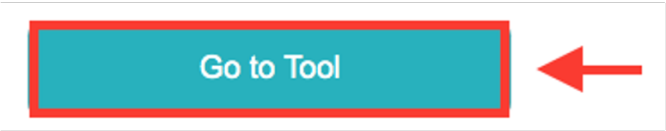


19.3.4.3 Step 3: Members join

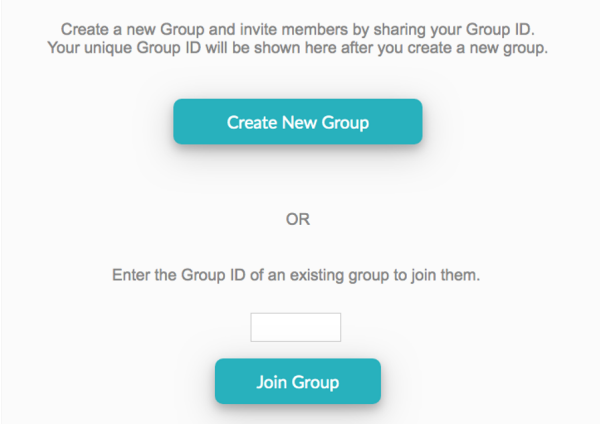
Once the Driver has shared the group ID, AND ONLY THEN do the following:

1 Click on the Team Up! assignment from the Course Overview page.

2 Click 'Go to Tool.' If it is not appearing, refresh your page or try a different browser.



3 Input the code your driver provided and click 'Join Group.'



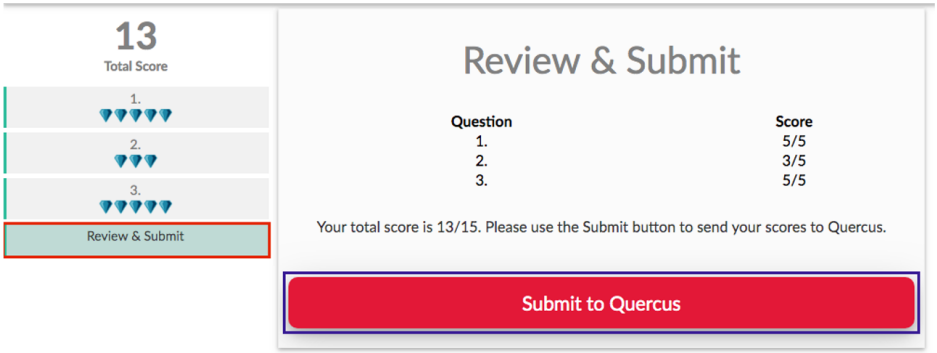
19.3.4.4 Step 4: Complete the activity

1. You will need to have the **interactive activity (question document)** associated with this task open on to see the answer options and update any code. **Be prepared to switch between Team Up! and the JupyterHub tab.**
2. The Driver chooses a question, and the team can participate by choosing their own answers and discussing their choices with each other. Once the team decides on a final answer, the driver chooses it and clicks the Check Answer button. **Note**: The Driver can check an answer despite what the team members choose, so be careful that you have consensus before clicking the Check Answer button. You cannot undo.

3. If the answer is **correct**, you will be awarded the full score or "diamonds" for that question. You will be able to proceed onward by clicking the Next button that will appear under the Check Answer button or by clicking on a question in the menu on the left. If your team's answer is **not correct**, discuss the question and choose again. Repeat until you get the correct answer. All teammates can see whether responses are correct/incorrect as well as their team's overall score as they progress through the Team Up! session.

19.3.4.5 Step 5: The Driver MUST click Submit to Quercus for your answers to be processed.

If this is not done, no one on the team will get a grade!!!!



19.4 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

19.4.1 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.

19.4.2 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 Biden[1] in class)

19.4.3 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.

[1] Yes, these are real images students have used.

19.4.4 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.
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.

19.4.5 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.

19.4.6 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.

19.4.7 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).

19.5 Student support services and resources

19.5.1 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/

19.5.2 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.

19.5.3 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.

19.5.4 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.

20

FAQs and Errata

20.1 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.

20.1.1 Course admin

20.1.1.1 STA303 pre-requisites

20.1.1.1.1 ***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.

20.1.1.2 Lecture and tutorial times

20.1.1.2.1 ***How exactly do the tutorials and lectures work?*** **Wednesdays**: 1st part is (recorded) content/demo and the second is office hour (unrecorded). Aiming for half-half, but will vary.

Thursdays: The 1st Thursday of a module will have a Team Up! or other practice activity. This will also be made available asynchronously, but the knowledge basket points for a Team Up! can only be earned by **synchronous, group participation**. The 2nd Thursday will be drop-in TA office hours to get help on content and assessments.

Thursday sessions are not recorded, but the Team Up! or other activities content will be made available to everyone. There are several other ways to earn your knowledge basket points if this doesn't suit your tech setup (e.g. no mic) or time zone.

20.1.1.3 Recorded lectures

20.1.1.3.1 ***"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.

20.1.1.4 Sections

20.1.1.4.1 ***"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."

20.1.1.5 Attending synchronous class

20.1.1.5.1 ***"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.

20.1.1.5.2 ***"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. More info and troubleshooting advice here.

20.1.1.5.3 ***"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.

20.1.1.5.4 ***"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.

20.1.2 Assessments FAQ

20.1.2.1 Knowledge basket FAQ

20.1.2.1.1 ***"I don't really understand the knowledge basket, how will that work? Can we earn more than 5 marks?"*** The knowledge basket maxes out at 5%, otherwise I get in trouble for giving away too many "free" points. (I know, I wish.)

20.1.2.1.2 ***"Can we do as many knowledge basket activities as we want in order to gain the full 5%?"*** Yes! But do try to plan ahead, as I won't be adding a bunch of make-up opportunities at the end…part of the goal is to incentivize personalized, regular participation in the course to create the best learning outcomes for you.

20.1.2.1.3 ***"How fast will we receive grades back for the knowledge basket activities?"*** It will depend, but many should be fairly quick—a few days if autograded/based on completion. 1–2 weeks for professional development proposal and evidence & reflection.

20.1.2.1.4 ***"I am not too sure what 'pre-knowledge quiz: 80% or workshop' is supposed to mean."*** If you score 80% or more on the pre-knowledge check, you automatically get 0.5% for doing it + 0.5% for a high score. If you do the check but get less than 80% you get 0.5% for completion, and can still get that additional 0.5% if you come participate in the workshop.

Workshop open to all. And you don't have to go if you have other plan for filling your basket. It is meant to be helpful to you, though.

And no, you can't get the workshop AND 80%+ points. One or the other for a max of 1%.

20.1.2.1.5 ***"What will additional opportunities like speaker series reflections be like?"*** May vary depending on the speaker/workshop, but for next week (grad panel) it will likely be attendance + a short Quercus survey. Just confirming a few things still so, TBA. (Watch for more details in announcements.)

20.1.2.1.6 ***"Are the three phases of each writing assignment graded on completion?"*** Yes, but see caveats here in the general instructions for the writing tasks.

20.1.2.1.7 ***"No questions, although seems to be a hectic course with a lot of minor deadlines."*** If you don't want lots of little deadlines, you don't have to have lots of little deadlines! Do the professional development tasks + a few Team Ups! or the professional development task + punctuation art. Throw in the pre-knowledge check and you're looking great!

20.1.2.2 Portfolios FAQ

20.1.2.2.1 ***"I am curious about the format of portfolio and assessment and how they will be conducted."*** 2022-01-12: **More info over the next two weeks**, but these will be a lot like a traditional assignment, but with a writing task (the module writing tasks are good indicators and practice) and some additional narrative about what you're learning/demonstrating. You'll submit commented code in an Rmd and the associated PDF.

20.1.2.3 Mixed assessments FAQ

20.1.2.3.1 ***"How does the mixed assessment work? Is it like a midterm?"*** Mostly, except there will also be a small amount of pre-work. More information closer too, but the overall goal is to make it more real-world. When you prep for a big meeting in the real world, you might calculate some summary stats, have a few basic plots ready, have some notes, etc. Sure, there are some things where you'll have to think it through in the moment (in the mixed assessment and in the real world), but I hope having the untimed pre-work helps lessen some of the test anxiety folks can feel. Less mind reading, more targeted prep.

20.1.2.3.2 ***"What's the difference between mini and mixed assessments? Are they both term tests?"*** One is bigger and one is smaller. ;) But yeah, the mini has one 50 minute timed component while the main mixed assessment will have more pre-work and two 50 minute timed parts.

20.1.2.3.3 ***Will we get some sample practice problems to work on before the mini-mixed and mixed assessments?*** The 'mini' is effectively practice problems for the main, but Team Up!s will also be useful practice for the mixed assessments. If there isn't a Team Up! (which you've be able to review asynchronously for study purposes) for a module I'll offer a practice quiz (no points) option or other practice activity.

20.1.2.3.4 ***"What exactly will we do in assessment weeks?"*** There is **no new content or writing module**. They are a time to focus your 6–8 hours of weekly STA303 time on the assessment. **Mixed:** doing the pre-work and the 100 minute (2 x 50) timed component. **Project week:** working with your group/self, asking questions in office hours. The expectation is that you'll start the project well before then, but I know the end of the semester can be stressful and busy, so I'm making things as simple and focused for you as I can.

20.1.2.4 Final project FAQ

20.1.2.4.1 ***"What kind of activities will the final project entail?"*** More details before reading week, but you will receive data and a client brief and be asked to create a written report that includes an executive summary as well as a more technical report.

20.1.2.5 Assessments: General FAQ

20.1.2.5.1 ***"How we can choose path A or B for the marking scheme– would this be on Quercus?"*** As long as you're okay with getting the higher of the two marks, I'll calculate your grade both ways and give you the higher one.

20.1.2.5.2 ***"So are all assessment due on Thursday 3:03?"*** Almost all the big ones! The mixed assessments have a window, but professional development proposal, professional development evidence & reflection, mini-portfolio, portfolio and project are all due on a Thursday at 3:03 p.m.

20.1.2.5.3 ***"Are we allowed to answer assignments with programming languages other than R?"*** This is an R-based course, so only coding assessments using R will be accepted. (Note: For the professional development task ONLY you could use another language in a blog post/write-up etc., as we are grading your goal-setting and reflection on a communication/collaboration task, not your code directly.)

20.1.2.5.4 ***"What are the minimum requirements to pass the course?"*** 50% overall, no special conditions.

20.1.3 Team Up!

20.1.3.1 Troubleshooting & FAQ advice from U of T CTSI

Link to student facing Quercus Support resources.

20.1.3.1.1 ***"My screen seems to be lagging compared to those of my group members (e.g., 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!

20.1.3.1.2 ***"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.

20.1.3.1.3 ***"How do I send my completed quiz results?"*** Click the large red "Submit to Quercus" button at the end of the Team Up! quiz.

20.1.3.1.4 ***"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.

20.1.3.1.5 ***"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.

20.1.3.1.6 ***"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.

20.2 Other

20.2.1 References

20.2.1.0.1 ***"Can you (Prof. Bolton) write me a reference?"*** 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.

20.2.1.0.2 ***"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

20.3 Errata

ID	Location	Note
1	Module 1 Team Up!	See corrected asynch answers. Marks were corrected for all groups.
2	HTML version of Syllabus in course guide	Mention of a different % for the knowledge basket for Path A. Both paths have a 5% cap.
3	Mini-portfolio instructions	There should be 6 columns, not 5 in the tibble `ci_vals`. `ci_vals`, tibble with ~~5~~ 6 columns (after all steps completed)
4	Portfolio instructions	There are 5 tasks total (some have sub parts). In Task 5: Simulating p values, there are only 3 simulated datasets to make, `sim1`, `sim2`, `sim3`. A previous Re-pulling the template through the template link will update your instruction document. Note: The mention of `sim4` in the `sim_desc` dataset IS intentional. Notice the instruction to "Make any alterations necessary to `all_sim` so that you can join on the dataset `sim_description` to provided better labels for each simulation (but without including any irrelevant labels)." Piazza: https://piazza.com/class/kx47tj4fmy65dg?cid=212

ID	Location	Note
5	Portfolio instructions	Task 2a) Missing fragments: "Use `kable` to display the table nicely and add an appropriate caption." "`agg_mod`, a main effects model where `yield_avg_int` is the response and `patch` and `treatment` are the predictors. Use appropriate data aggregated across both `patch` and `treatment`." `var_ab`, variance in yield explained by the interaction between patch and treatment, after accounting for the fixed effects and other sources. Piazza: https://piazza.com/class/kx47tj4fmy65dg?cid=216
6	Portfolio instructions	Task 3a) Test 3: In the original version the 3rd test was shown as `interpret_ci(10, 20, 95, 99)` in the template. The instructions say `interpret_ci(10, 20, -1, tibble(stat = 3))` and is the version that should be used. Piazza: https://piazza.com/class/kx47tj4fmy65dg?cid=238
7	Statdew Valley	In the Summary tables, the model formula should read $\beta_1 \text{song}_i$, not β_2. Piazza: https://piazza.com/class/kx47tj4fmy65dg?cid=241 At the end of **16.4.5.6.5 Summary of all our calculated error variances:** Added a little more information AND some example code. Piazza: https://piazza.com/class/kx47tj4fmy65dg?cid=240
8	Portfolio instructions	Task 2a) ~~`my_patch`, your 'patch' of strawberries, grown based on a `set.seed` that is the last THREE digits of your numeric student ID.~~ `my_patch`, your 'patch' of strawberries, grown based on using your `last3digplus` as the seed (created in the first part of the task).
9	Portfolio instructions	Task 5: Q-Q plots should be 3x1 not 2x1. Piazza: https://piazza.com/class/kx47tj4fmy65dg?cid=285

21

Bits and pieces

21.1 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)
```

21.2 M1 supporting information on matrices (not assessed)

21.2.1 Background

21.2.1.1 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 umber 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 $(X^TX)^{-1}X^TY$. The elements of β can't be estimated if X^TX (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 out 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 satisfies linear independence. IF we ditch the intercept, we can have k dummies, but this is only usually useful in the specific case of ANOVA.

21.2.2 Example

```
library(tidyverse)
library(palmerpenguins)
```

```
# function to replace NAs with 0 and text with 1
dummify <- function(x){
  if_else(is.na(x), 0, 1)
}

# create a smaller toy version of the penguin dataset (just for diplay purposes)
set.seed(24601)
pengwings <- penguins %>%
  group_by(species) %>%
  sample_n(4) %>%
  select(body_mass_g, species)

pengwings
```

```
## # 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
```

```
# creating a version of the data where there is a dummy column for each level of species instead of one sp
wider <- pengwings %>%
  pivot_wider(id_cols = everything(), names_from = species, values_from = species, names_prefix = "species
  mutate_at(vars(starts_with("species.")), dummify)

wider
```

```
## # 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
```

21.2.2.1 Classic regression, k-1 dummies (intercept)

```
mod1 <- lm(body_mass_g ~ species, data = pengwings)
summary(mod1)
```

```
##
## 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
```

```
head(model.matrix(mod1))
```

```
##   (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?

```
qr(model.matrix(mod1))$rank == ncol(model.matrix(mod1))
```

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

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

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

```
mod2 <- lm(body_mass_g ~ species.Adelie + species.Gentoo + species.Chinstrap, data=wider)
summary(mod2)
```

```
##
## 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
```

```
model.matrix(mod2)
```

```
##    (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
```

```
qr(model.matrix(mod2))$rank
```

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

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

```
qr(model.matrix(mod2))$rank == ncol(model.matrix(mod2))
```

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

Not linearly independent. Why?

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

```
m <- model.matrix(mod2)

m[,1] == m[,2] + m[,3] + m[,4]
```

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

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

```
mod3 <- lm(body_mass_g ~ 0 + species.Adelie + species.Gentoo + species.Chinstrap, data=wider)
summary(mod3)
```

```
##
## 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
```

```
model.matrix(mod3)
```

```
##    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
```

```
qr(model.matrix(mod3))$rank
```

```
## [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?

```
qr(model.matrix(mod3))$rank == ncol(model.matrix(mod3))
```

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

Great, back to being linearly independent.

21.2.4 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/

21.3 *p* values (recap)

Remember: "Small but mighty!" The *smaller* the p value, the *stronger* the evidence against the null hypothesis.

- We should never be making claims in in favour/against the alternative hypothesis. Our *statistical* claims are always about the null. This may inform practical decisions in less strict ways, depending on practical significance, risks, etc., but a correct statistical claim for a p value should ALWAYS be about the null. Absolutes are rare in teaching statistics.

- A common threshold for rejecting or failing to reject the null hypothesis is 0.05. This is mostly from habit/convention, rather than some truly meaningful cosmic value. "Yeah, 1 in 20? Unlikely, sure." Difference disciplines/sub-disciplines develop norms appropriate to their context.
- Many statisticians—especially in light of the reproducibility crisis and poor public, and even sometimes researcher, understanding of p values—prefer to make statements about the **strength** of evidence, not just reject/fail to reject.

p value range	Strength comment
> 0.1	No evidence against the null
$0.05 < p$ value < 0.1	Weak evidence against the null hypothesis
$0.01 < p$ value < 0.05	Moderate/some evidence against the null hypothesis
$0.001 < p$ value < 0.01	Strong evidence against the null hypothesis
< 0.001	Very strong evidence against the null hypothesis

21.3.1 What if you get a value that is exactly one of these thresholds?

It would be very rare to get a p value of exactly one of these thresholds. If you were to, you could decide based on your judgment of the context. Should you be more or less conservative in your interpretation based on the needs of the project/client?

So, you'll almost always have something, that while it may round to 0.001 (or another threshold), is really something like 0.001345789 or 0.00098963. What judgment would you make in each case, with this in mind?

But remember, these are **fuzzy boundaries based on easy-to-grasp probabilities for humans**, particularly humans used to thinking in base 10. 1 in 20, 1 in 100, 1 in 1000, etc. Easy to think about but not universally meaningful thresholds. Honestly, being a little over or under really doesn't make that much of a difference. p values are continuous random variables. We impose discrete categories for our own ease. Used too rigidly, these guidelines have the same problem as rejecting/failing to reject, but just with a few more levels.

22

Announcements summary

This is mostly to aid searching, as you can easily search the course guide, but not always as easily the full text of announcements on Quercus.

Title (and link)	Date	Key topics
Welcome to STA303/1002!	2022-01-10 3:03 p.m. ET	• Welcome • Course structure and delivery • How the course is managing online/in-person • Communication policy • To dos • Volunteer notetaker role • Volunteer student representative role
STA303 L0101 class starts online in 20 minutes (10:10 a.m.)	2022-01-12 9:50 a.m.	For L0101 • Reminder that class starts at 10:10 • Link to Zoom (and reminder to be licensed) • Link to slides on the course overview page
Make sure your Zoom is licensed!	2022-01-12 1:41 p.m.	How to ensure authorization/license on your U of T Zoom account
STA303 L0201 class starts online in 20 minutes (3:10 p.m.)	2022-01-12 2:50 p.m.	For L0201 • Reminder that class starts at 3:10 • Link to Zoom (and reminder to be licensed) • Link to slides on the course overview page • Note that the first part of the session will probably run long
Team Up!: Important information for tutorial sessions today	2022-01-13 10:15 a.m.	• What you need to participate in Team Up! activities • What to do before the activity • Rules and logistics

Title (and link)	Date	Key topics
Reminders before the 5 p.m. team up	2022-01-13 4:31 p.m.	• Reminder to read previous announcement carefully • Key points
'Getting to know you' survey grades posted	2022-01-13 7:56 p.m.	
An end of week recap	2022-01-14 2:30 p.m.	• Slides and recordings from class as well as a version of the **Team Up! questions with answers** are all available on the Course overview page. • Join Piazza, our course discussion board. • **Frequently Asked Questions** are answered here in the course guide. (Save yourself and others time: Please check this and the syllabus *before* asking on Piazza.) • Module 1 writing reminder.
Module Team Up! grades posted	2022-01-17 10:05 a.m.	See the FAQs in this announcement.
Knowledge basket (overall grade return) grade *updated*	2022-01-17 10:07 a.m.	See the FAQs in this announcement.
Week 2 overview	2022-01-17 10:10 a.m.	• Knowledge basket grades have been updated – If your Team Up! grade is missing it is because your team didn't submit or you ended up in a solo group ID • Grad school info session and panel on Wednesday (in class) + Prof office hours • TA office hours on Thursday • Get a head start on Module 2 with these readings and case study • Welcome to those who have just joined off the waitlist (don't worry about any missed assessments)
Second Wednesday speaker series: grad school info session and panel + office hour today	2022-01-19 9:30 a.m.	FAQ for the grad school info session and panel.
M1: Updates, reminders and resources	2022-01-20 12:00 p.m. ET	• What does the announcement about **in-person/online** mean for STA303/1002? (Not much.) • **To-dos** (Things due today, tomorrow, or that you can work on now) • **TA office hours** in tutorial times this week (12 p.m. and 5 p.m.) • **Resources** mentioned in Wednesday's grad school information session and panel

Title (and link)	Date	Key topics
Module 1 graduate school info session and panel	2022-01-24 10:34 a.m. ET	
Pre-knowledge check (final grade return)	2022-01-24 10:34 a.m. ET	Will be updated again after the workshop
Module 1 check-in	2022-01-24 10:34 a.m. ET	
Knowledge basket (overall grade return)	2022-01-24 10:34 a.m. ET	Updated with new assessments
'Getting to know you' survey	2022-01-24 10:35 a.m. ET	See previous release, this was for any remaining hidden grades from folks who had recently joined the course.
Pre-knowledge check (quiz)	2022-01-24 10:36 a.m. ET	
Module 1 writing	2022-01-24 10:36 a.m. ET	
Module 2: Launch + Knowledge basket update	2022-01-24 10:45 a.m. ET	• **Module 2: Professional skills for data analysis** is available now – **Mini-portfolio instructions** will be up by tomorrow and we will go over it in class on Wednesday • Knowledge basket **grades** have been **updated** (if you have a question about your grade, please make sure you read this section carefully before posting) • What happens if you/someone you know were **removed** from the course (**UCheck**)? • What to do if you've joined off the **waitlist** (don't worry about missed knowledge basket assessments) • Statistics **job opportunities** + student rep opportunities

Title (and link)	Date	Key topics
M2: Mid-week megaphone	2022-01-26 9:00 a.m. ET	• **Mini-portfolio instructions** now available and we will go over the assessment in class today • Deadlines this week – Team Up! tomorrow • Opportunities: Toronto Data Workshop & funded training in teaching technical skills • *No points* practice version of the pre-knowledge check available now
Module 2 Team Up! grades posted	2022-01-31 5:28 p.m.	See the FAQs in this announcement.
M2: Second week overview + TA OH in 30 minutes	2022-01-31 5:30 p.m. ET	• New College Statistics Aid Center office hours run by one of our course TAs (Lei) 6:00 to 9:00 p.m. ET TODAY — all welcome! • Bonus mini-portfolio office hour with TA Rob TOMORROW (Feb 1, 2:10 to 4:00 p.m. ET) • Second Wednesday speaker series: **Pre-knowledge workshop** with TA Dawn (0.5 pts) • Knowledge basket **grades** have been **updated** (if you have a question about your grade, please make sure you read this section carefully before posting) • **Student representatives:** Yutong Chen & Zhelan Li (reach out with feedback!) • [Optional] Are your **R coding skills** feeling rusty, crusty, or not really there? • [Optional; get ahead] Readings for **Module 3** • [Optional; get ahead] Writing sample prompt for the **portfolio assessment** (full instructions coming soon)

Title (and link)	Date	Key topics
Mini-portfolio office hour starts NOW	2022-02-01 2:10 p.m. ET	**Timing:** Tuesday, February 1, 2:10 to 4:00 p.m. ET **Zoom link**: https://utoronto.zoom.us/j/81630651947 (Links to an external site.) Have you: • Watched this part of the **recording** from last Wednesday's afternoon class that shows you how to get started? 32:00–end (approx) • **Knit** the template? • Tried uploading to **MarkUs**? (You can upload as many times as your like before the deadline, we only mark the last upload.) • Created a **checklist** of all the things you need to do? – Cover page – Introduction – Statistical skills sample – Writing sample – Reflection – Code is **commented** and **displayed** in the PDF • Caught up with all the tips and hints on **Piazza**? – Caption vs figure caption? (Links to an external site.) – Which geom to pick for task 2? (Links to an external site.) – Why can't R find my data? (Links to an external site.) – Wait, shouldn't there be 6 columns? (Links to an external site.) (and added to the Errata (Links to an external site.)) – Population mean or simulation mean? (Links to an external site.) – What is the 2 in N(10,2)? – How to get the histograms on top of each other? (Links to an external site.) – …and so much more! (Links to an external site.) • Made a plan to come to **office hours** today/tomorrow/Thursday? • Remembered the course policy around **extension requests** and **declarations of illness/emergency**? (Links to an external site.) • Remembered that you don't actually have to do this assessment if you want to take Assessment Path B (Links to an external site.)? • Seen that you could start on the main portfolio writing sample already?

Title (and link)	Date	Key topics
Help notes pre mini-portfolio submission	2022-02-02 2:38 p.m. ET	My **last office hour before the submission of the mini-portfolio is today, 4:10 to 5:00 p.m. ET** (after the workshop in the first half, same link). All are welcome, regardless of section enrolment or attendance at the first part. Just so you're aware, I *can't* stay past 5:00 p.m. ET today (though some weeks I can/do) and will be minimally available *tomorrow* due to my other teaching and meeting commitments. Please plan accordingly. There will also be a **TA office hour tomorrow from noon to 1:00 p.m.** (and one Please make sure you're **knitting early and often** and have checked out some of the I'm sure you know this already, but just in case: not having your Piazza question answered, knitting errors, or not knowing where to submit are not reasons I will grant you an extension.
Module 3: Portfolio + R resources + module 2 check-in responses	2022-02-08 8:00 a.m. ET	• **Module 3: Linear Mixed Models** is available now – **Portfolio instructions** are available and we will go over them in class on Wednesday • Knowledge basket **grades** have been **updated** (if you have a question about your grade, please make sure you read this section carefully before posting) • **Team Up! activities** will remain **online** (but this doesn't mean you can't meet in person!) • **R resources** • **Reminders** – Need to request an **extension** or declare an illness/personal emergency? **Please use the form!** – **UTORid** student id number. In almost all admin situations (like managing extensions) I need your UTORid which will be based on your name and may or may not have a number.
Module 2 check-in grades posted	2022-02-08 9:15 a.m. ET	
Pre-knowledge check (final grade return) grades posted	2022-02-08 9:58 a.m. ET	Max score is 1 which could be earned by completing the pre-knowledge check quiz and scoring 80%+ OR by completing the quiz and attending the workshop (75% attendance + post-event survey). Partial marks also available. (0.5 for quiz alone, 0.5 for workshop alone).
Module 2 writing grades posted	2022-02-08 9:59 a.m. ET	
Knowledge basket (overall grade return) grades posted	2022-02-08 10:34 a.m. ET	Updated with new assessments

Title (and link)	Date	Key topics
Mini-portfolio grades released	2022-02-11 6:44 p.m. ET	Announcement of grades with sample solutions and feedback.
Module 3 check-in grades released	2022-02-23 4:29 p.m. ET	

Updates discontinued, page just kept to avoid broken links.

The volunteer notetaking program is one of the most commonly used resources provided to students registered with Accessibility Services. Here at the University of Toronto, the accessibility office is seeking dependable volunteer note-takers to assist students living with a disability to achieve academic success.

Volunteers report that by giving back to the U of T community, their class attendance and notetaking skills improve.

To become a volunteer notetaker, all you must do is attend classes regularly & submit/upload your notes consistently/weekly. Volunteers can receive co-curricular credit(s) and/or a certificate of appreciation and/or be eligible for year-end prize draws.

To become a volunteer note-taker, please follow these 4 steps:

1. Register Online as a Volunteer Note-Taker at: https://clockwork.studentlife.utoronto.ca/custom/misc/home.aspx
2. Click on Volunteer Notetakers, and sign in using your UTORid.
3. Select the course(s) you wish to take notes for. Please note: you do NOT need to upload sample notes or be selected as a volunteer to begin uploading your notes.
4. Start uploading notes.

Email us at as.notetaking@utoronto.ca if you have questions or require any assistance with uploading notes. If you are no longer able to upload notes for a course, please also let us know immediately. For more information about the Accessibility Services Peer Notetaking program, please visit https://studentlife.utoronto.ca/program/volunteer-note-taking/]. Thank you for your support and for making notes more accessible for our students.

AS Note-taking Team

22.0.1 Looking for student representatives for STA303

Our course is part of a small pilot of student representative pilot for the Department of Statistical Sciences and I am looking for two undergraduate student representatives to work with me to represent our course. (STA1002 students, there isn't a grad version of this, but you can also provide feedback to these reps and I encourage you to also reach out to me directly!)

Student representatives work with their instructor to solicit and respond to student feedback throughout the semester. Representatives talk with their instructors about how the course is going, provide feedback to the department about student needs more generally and learn about department resources so that they can help connect students to these, as relevant.

You can learn more about the experience from previous reps here.

22.0.1.1 The role of a student representative

Once a rep is selected, they will:

- Attend a mandatory training session.
- Have their email made available to the class on Quercus so students can contact them. They may also introduce themselves in a synchronous class, or write a small blurb for an announcement or create a pre-recorded video. This is up to the student/instructor team.
- Reach out to students to elicit feedback. Additional tasks (e.g., helping with optional student pulse surveys or organising study groups or class events, etc.) would be left to the discretion of the student representative and instructor (they are *not* a required part of the program).
- Keep in regular contact with the instructor as issues or concerns arise.

22.0.1.2 Benefits of being a student representative

- Being part of creating a great learning experience for themselves and others.
- Networking with other students and members of the Department.
- Opportunities to work closely with an instructor that may mean they could be a strong referee for you and/or lead to taking a reading course with them etc.
- A chance to demonstrate and improve communication, leadership, and organisational skills.
- Something to put on a CV.

22.0.1.3 How to become a rep

If you are eligible (see below), you can put yourself forward via this form. Please make sure you choose the correct section, where applicable! Student representatives for each section will be selected via lottery from the eligible students that have expressed interest by Monday, January 24 at 6:00 p.m. ET.

22.0.1.4 Eligibility criteria

- You will need to be available to attend an online training, 11:10 a.m.–12:00 p.m. ET on Friday, 28 January 2022.
- You must be enrolled in the section you put yourself forward for.
- You *don't* have to be enrolled in a DoSS program to put yourself forward for the role, though it may be of more interest/use to you if you are.

PDF version of below ad here.

Student voice

Join *phase 2* of a DoSS pilot: The Undergraduate Consultative Committee

Student representatives work with their instructor to solicit and respond to **student feedback** throughout the semester.

Read what past reps had to say about this program in the Department News.

How are representatives selected?
Interested students **register** at **forms.office.com/r/3ZjP0TBpQd**. The reps will be chosen by lottery from those who register their interest by 6:00 p.m. ET on Monday, Jan 24. Students will be informed of the outcome on Jan 25 or 26.

Required training:
Fri, Jan 28 at 11:10 a.m.
(via Zoom)

Time commitment:
2–3 hours of meetings, 1–5 hours of emails & other engagement.

SIGN UP NOW

Register here (https://forms.office.com/r/3ZjP0TBpQd)

Register by 6:00 p.m. ET on Jan 24.

Leadership • Empathy • Community • Transparency