Course Info + Intro

STA304

Week 1 – Synchronous Lecture

Plan for this session

- Introductions
- Syllabus Info
- Survey Design
- Selection Bias
- Measurement Error
- Sampling and Non-Sampling Errors

Prof. Samantha-Jo Caetano (Sam)

- Born in Toronto, moved to suburbs (Mississauga) as a teenager, also lived in Waterloo and Hamilton.
- Undergrad in Math & Applied Stats at UTM
- Masters and PhD in Statistics from McMaster
- **Current hobbies:** cooking, reading, video games

Research interests

- Statistics Education
- Biostatistics (Survival & Cancer Data)
- Model Validation

Two Truths and One Lie

1. I have one heart in my body.

2. I have two dogs that collectively weigh over 300 lbs.

3. I can ride a unicycle.

Go to **pollev.com/sta**

Course Information

TA Team

Mercedes

Bushra

Maksim

Justin

Stella

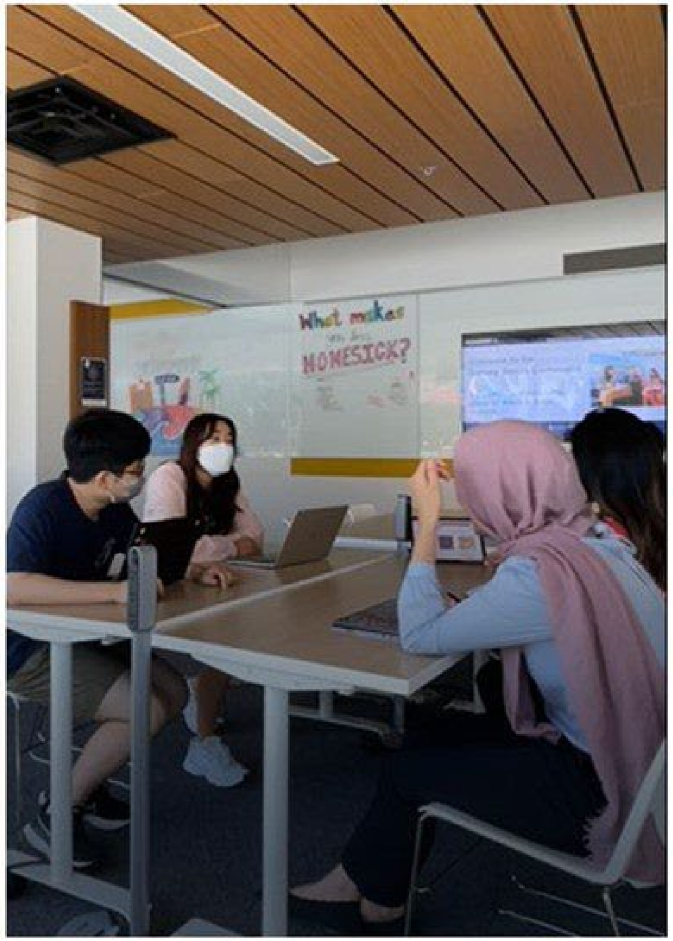
Mark

Fatema

Nnenna

See them during office hours, on Piazza, in class, through email & marking.
Check out the “Get to know the TA team” page.

UNIVERSITY OF TORONTO
FACULTY OF ARTS & SCIENCE



Join or Lead an RSG

- Meet weekly with up to 8 classmates online
- Review and discuss course material
- Prepare for tests and exams
- Get student advice from upper year mentors

Last year, over 3000 students joined a Recognized Study Group (RSG) where they met friends and reached their study goals.

Plan for success this term by joining your RSG today.

Join an RSG today: **uoft.me/recognizedstudygroups**

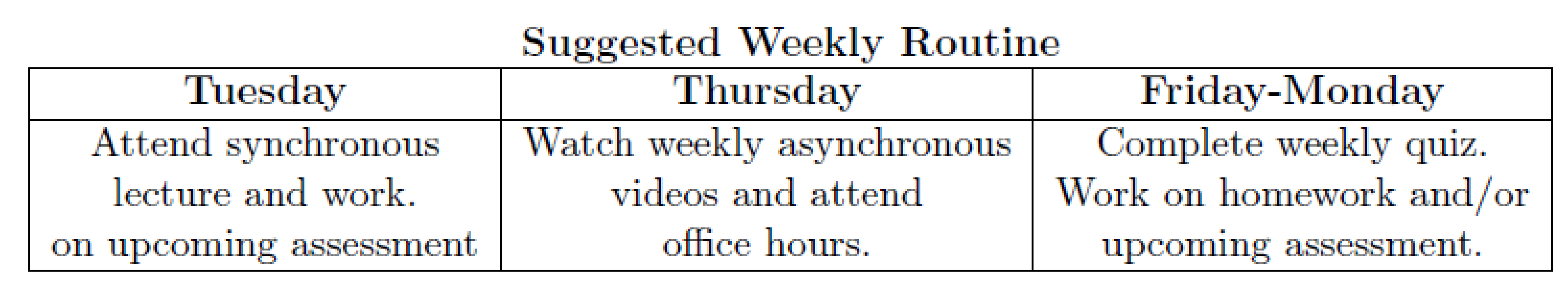
SIDNEY SMITH COMMONS

@sidneysmithcommons

CCR
CO-CURRICULAR RECORD

Schedule

IN-PERSON



Suggested Weekly Routine

Tuesday	Thursday	Friday-Monday
Attend synchronous lecture and work. on upcoming assessment	Watch weekly asynchronous videos and attend office hours.	Complete weekly quiz. Work on homework and/or upcoming assessment.

VIRTUAL

Assessments

- 10 quick quizzes or surveys
  - (Best 6 of 10)
- 10 weeks of in-class attendance
  - (Best 6 of 10)
- 2 Assignments
- 1 Midterm
- 1 Final Exam

Assessment	Due Date
Quiz/Survey (‘Best 6 of 10’)	Mondays at 11:59pm ET
Attendance (‘Best 6 of 10’)	Tuesdays during lecture
Assignment 1	Thursday January 30 at 8pm ET
Assignment 2	Thursday March 13 at 8pm ET
Midterm (in person)	Tuesday February 11(during lecture)
Final Exam (in person)	TBA by Faculty

Assessments

- 1 week grace period on Assignments
- But if you hand it in late the grading may be completed later
- Turnitin on Assignments

- If you do better on the final exam then the weight of the midterm will go onto the final
- If you miss the midterm the weight automatically goes onto the final.

Assessment	Due Date
Quiz/Survey ('Best 6 of 10')	Mondays at 11:59pm ET
Attendance ('Best 6 of 10')	Tuesdays during lecture
Assignment 1	Thursday January 30 at 8pm ET
Assignment 2	Thursday March 13 at 8pm ET
Midterm (in person)	Tuesday February 11(during lecture)
Final Exam (in person)	TBA by Faculty

Course Communication

- Email: sta304@utoronto.ca
  - Use this for personal inquiries.
  - Use this for self-declared absence form.
- Piazza
  - For general course related inquiries
  - Sharing ideas/promoting professional, academic and social events.
  - Netiquette: keep Piazza professional and overall positive.
    - Any concerns regarding grading, accommodations, personal issues, etc. should be sent to the course email.
- Office Hours
  - Instructor office hours are on Thursdays during the “lecture” times.
  - TA office hours TBA on Quercus.

Course Communication

- Missed Midterm
  - No need to notify us of a missed test, the weight goes automatically onto the final exam.
- Missed Attendance
  - No need to notify us of a missed attendance, as top 6 (out of 10) count toward the final grade.
- Missed Quiz
  - No need to notify us of a missed quiz, as top 6 (out of 10) count toward the final grade.
- Regrade Requests
  - Form to be posted on Quercus when grades are released.
  - Must complete the form 24h after grade release but within 1 week of grade release.

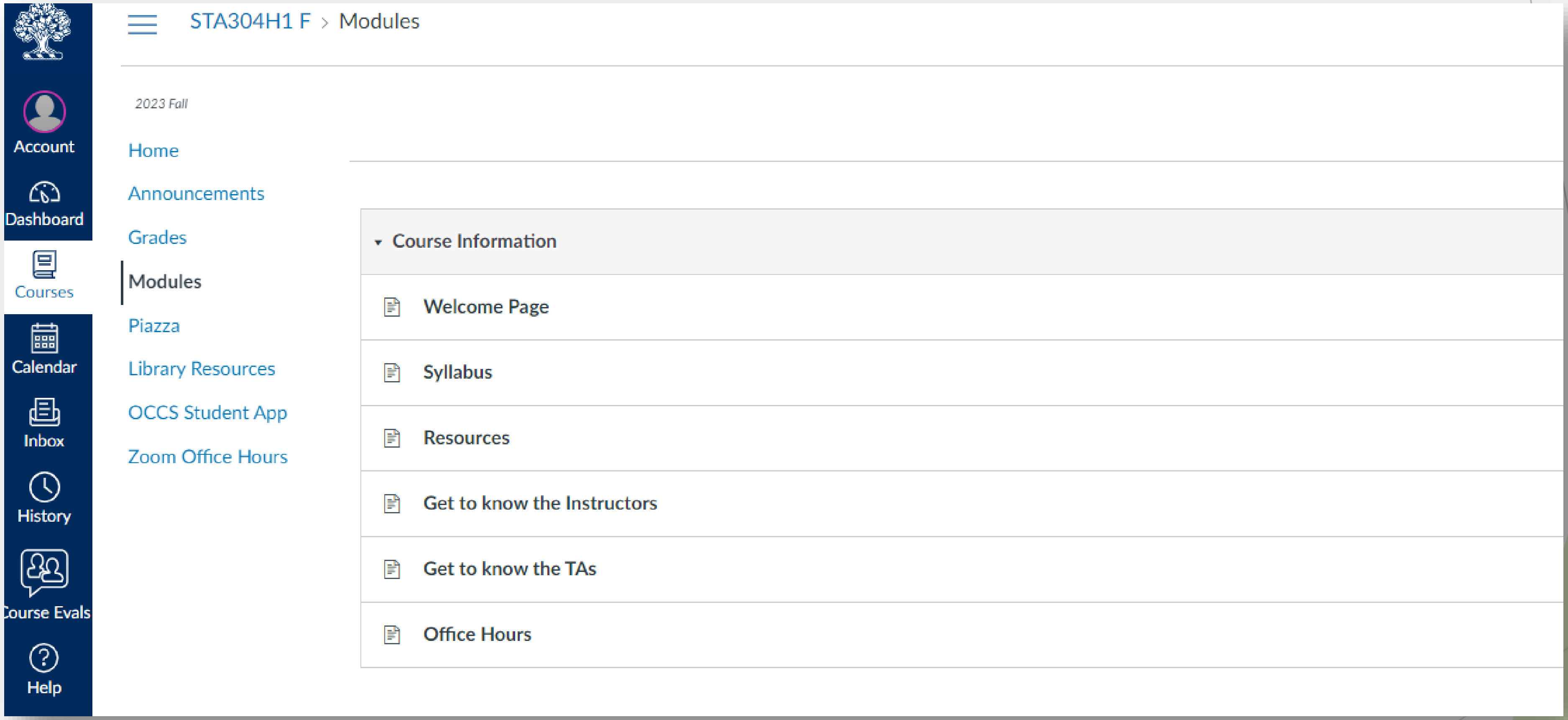
Textbook (Not Required)

- Wu, Changbao and Mary E. Thompson, 2020, Sampling Theory and Practice, Springer.This is the primary reference for the course.
  - Available for free as in pdf through the UofT library.

Programming

- R
- Rstudio
- Jupyterhub
- R Markdown
- Tidyverse

Course Quercus Page



STA304H1 F > Modules

2023 Fall

- Home
- Announcements
- Grades
- Modules
- Piazza
- Library Resources
- OCCS Student App
- Zoom Office Hours

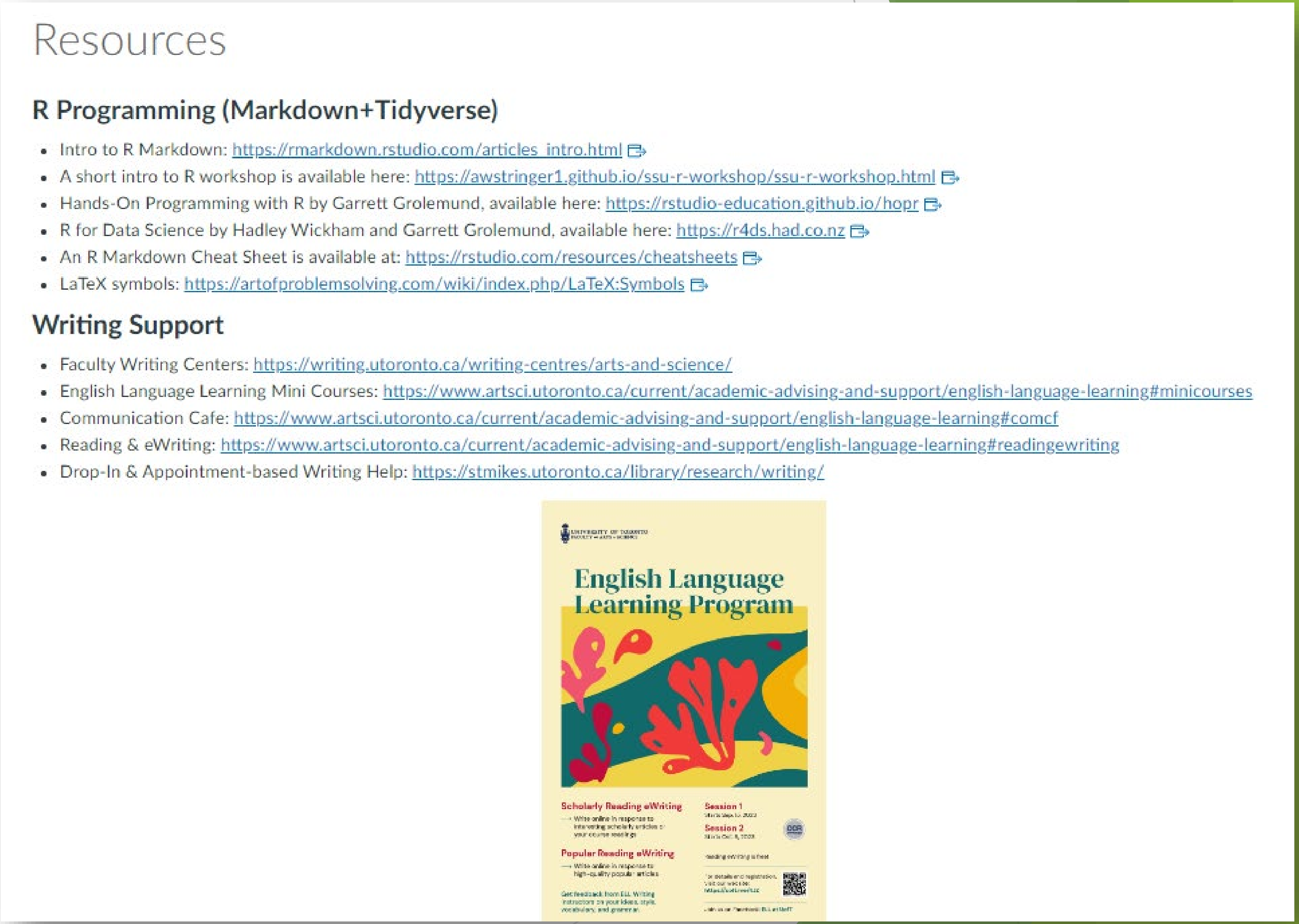
▾ Course Information

- Welcome Page
- Syllabus
- Resources
- Get to know the Instructors
- Get to know the TAs
- Office Hours

- Please take a look at the Resources and Office Hours page later this week.

Writing in STA304

- Reading Instructions
- Rubrics for Assignments
- Starting Assignments early & ask questions early!



Resources

R Programming (Markdown+Tidyverse)

- Intro to R Markdown: https://rmarkdown.rstudio.com/articles_intro.html
- A short intro to R workshop is available here: https://awstringer1.github.io/ssu-r-workshop/ssu-r-workshop.html
- Hands-On Programming with R by Garrett Grolemund, available here: https://rstudio-education.github.io/hopr
- R for Data Science by Hadley Wickham and Garrett Grolemund, available here: https://r4ds.had.co.nz
- An R Markdown Cheat Sheet is available at: https://rstudio.com/resources/cheatsheets
- LaTeX symbols: https://artofproblemsolving.com/wiki/index.php/LaTeX:Symbols

Writing Support

- Faculty Writing Centers: https://writing.utoronto.ca/writing-centres/arts-and-science/
- English Language Learning Mini Courses: https://www.artsci.utoronto.ca/current/academic-advising-and-support/english-language-learning#minicourses
- Communication Cafe: https://www.artsci.utoronto.ca/current/academic-advising-and-support/english-language-learning#comcf
- Reading & eWriting: https://www.artsci.utoronto.ca/current/academic-advising-and-support/english-language-learning#readingewriting
- Drop-In & Appointment-based Writing Help: https://stmikes.utoronto.ca/library/research/writing/

Academic Integrity

- Working independently on timed assessments.
- Not sharing code.
- Citing any external resources you use.
- Paraphrasing, not copying.
- Turnitin.

Generative AI

- You are not permitted to use generative AI (e.g., ChatGPT, Bard, etc.) on closed book assessments (i.e., midterm, final exam)

- You are permitted to use generative AI on assignments.
  - There will be specific instructions on how to cite and report your usage of it.
  - We are recommending (for those of you that decide to use generative AI) to use it as a “tool” and not as a “crutch”
  - Note: usage of generative AI is not required in this course! Use at your own capacity/risk.
    - E.g., if the generative AI software you use is “down” that will not affect due dates in this class.

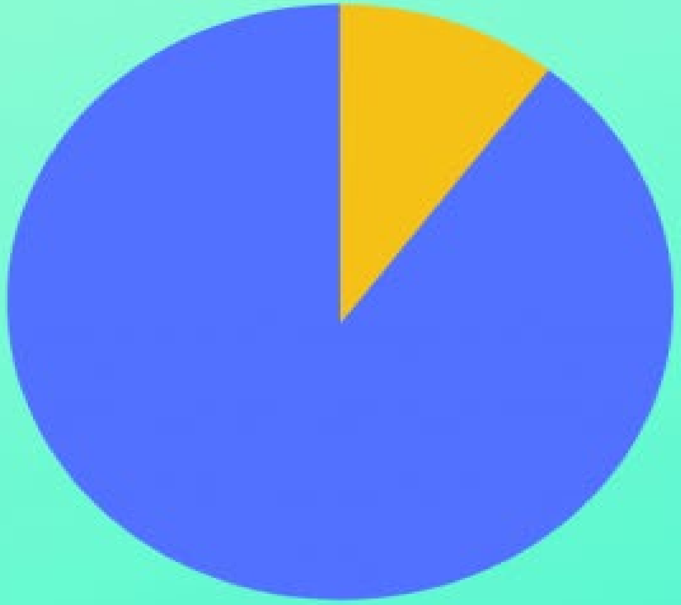
Paraphrasing

- Paraphrasing: express the meaning of (the writer or speaker or something written or spoken) using your own words

- Best practice is to read something. Take some time to not look at it and then try to re-express what you remember.

Questions?

Surveys & Sampling



Surveys

Survey Critique

Fill Out (or Read through) this Survey (4 min):

https://www.surveymonkey.com/r/6GX9GMT

Disclaimer: This is just a random survey I found online…

Think (2 min):

Think about at least 1 pro and at least 1 con of the survey questions, data collection process, etc.

Share (2 min):

We will discuss as a class what your pros and cons were. If we don't have time to get to your group, please let us know at pollev.com/sta.

Survey Design

- Avoid leading questions.
  - Leading questions push/guilt the survey taker to answer a certain way.
  - E.g., “How bad do you think the new policies are?”

- Avoid double barrel questions.
  - Double barrel questions contain multiple parts/questions within one “question”.
  - E.g., “How do you feel about the policy and how did you learn about it?”

- Avoid using niche jargon.

- Only ask what you need to know about.

- Try to keep questions neutral.

Testing your Survey

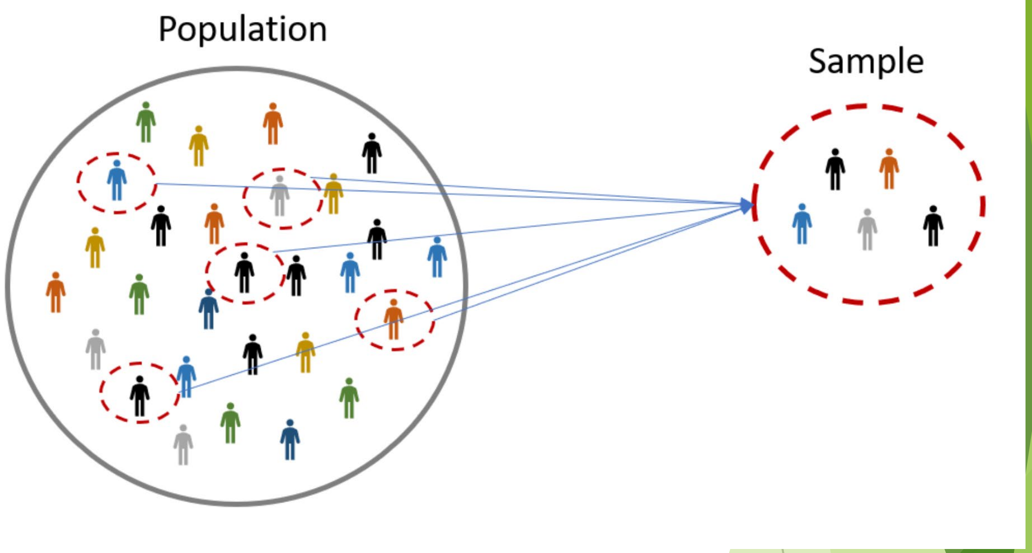
- When you design a survey have a small group of people (friends, colleagues, etc.) try the survey and let you know their thoughts.
  - Helps to find typos/ambiguities.
  - Help refine the survey and make it more clear.
  - Check that the answers you get from respondents match your expectations.
    - E.g., if you have an “Other” category that is selected a lot, then perhaps you need to expand your options.
    - E.g., if you have a text question perhaps it can help you refine the question style.

Some Definitions

Populations vs Sample

- Population: The entire group we wish to study/make inference on.

- Sample: The group which we collect data on.



Populations vs Sample

- Population: The entire group we wish to study/make inference on.

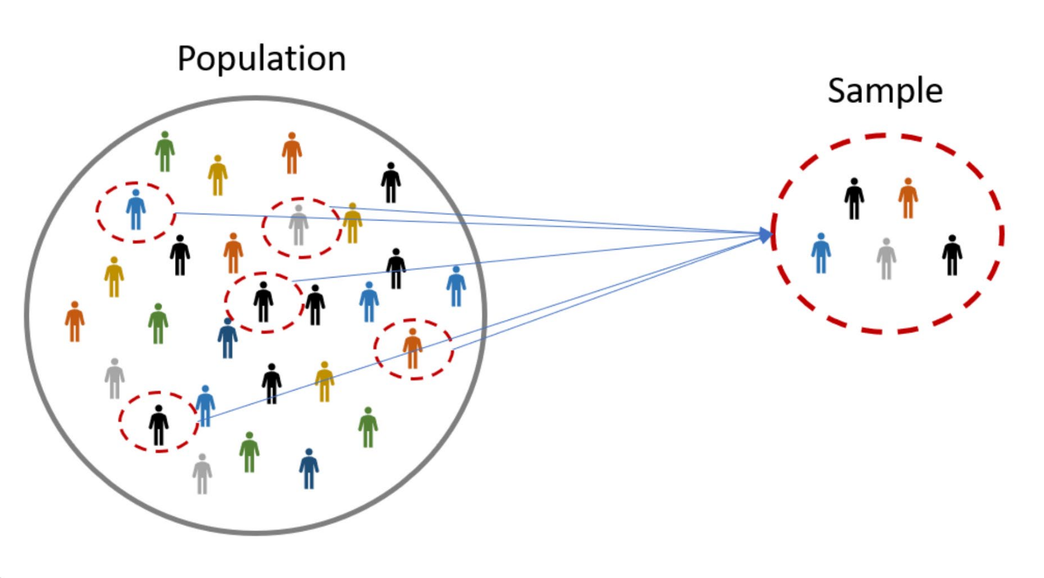
- Sample: The group which we collect data on.

Note: There are 3 populations to be mindful of:

Target population

Frame population

Sample population



Target Population

- **Definition**: The **target population** is the entire group of individuals or units that a researcher intends to study or make inferences about. This is the broader group that you want your findings to apply to.

- **Example**: If you’re studying the effectiveness of a new drug, your target population could be all people with a specific medical condition who are eligible for the treatment.

- **Key Point**: The target population is often too large or inaccessible to study in its entirety, so researchers focus on a subset (the sample) to make inferences about the whole group.

Frame Population

- **Definition:** The **sampling frame** is the list or database from which the sample is actually drawn. It is a practical representation of the target population, but it might not fully capture everyone in the target population due to limitations in access or records.

- **Example:** For the same drug study, the frame population might consist of patients registered in a particular hospital's database or a list of people who have volunteered for clinical trials. It may not include everyone with the condition, but it represents a segment of the target population.

- **Key Point:** The sampling frame may introduce biases if it doesn’t closely align with the target population (e.g., if the list is incomplete or selective).

Sample Population

- **Definition:** The **sample population** is the actual group of individuals or units that are selected from the frame population to participate in the study. This is the smaller subset of the frame population that the researcher collects data from.

- **Example:** Continuing with the drug study, the sample population could be the 200 people selected from the list of eligible patients at the hospital who have agreed to participate in the trial.

- **Key Point:** The sample population is intended to be representative of the frame population, and by extension, the target population, but it’s typically much smaller in size.

How these relate to each other

- The **target population** is the ideal group you want to make conclusions about.
- The **frame population** is the group from which you can practically select participants, but it may not include everyone in the target population.
- The **sample population** is the actual group selected from the frame population to participate in the study.

Example in Context

- (5 minutes) Let’s say I am interested in the sleep habits of my STA304 students, and have capacity to survey 25 students. Get into small groups of 2-4 people and try to define a target, frame and sample population.

- **Target Pop:**

- **Frame Pop:**

- **Sample Pop:**

Example in Context

- (3 minutes) Let’s say I am interested in the sleep habits of my STA304 students. Get back into your small groups and try to identify at least 3 groups who I may be missing.

Example in Context

- (3 minutes) Let’s say I am interested in the sleep habits of my STA304 students. Get back into your small groups and try to identify at least 3 groups who I may be missing.

Go to **pollev.com/sta**

Selection Bias

- Occurs when the individuals included in a study or analysis are not representative of the general population or the population intended for study.

- For example, in a medical study, if only individuals who are already healthy are included, the findings would not accurately reflect how the treatment works in a broader, more diverse group of people, such as those with underlying conditions.

Selection Bias

- Occurs when the individuals included in a study or analysis are not representative of the general population or the population intended for study.

- For example, in a medical study, if only individuals who are already healthy are included, the findings would not accurately reflect how the treatment works in a broader, more diverse group of people, such as those with underlying conditions.

Selection Bias

- **Non-random sampling:** Choosing participants based on certain characteristics (e.g., selecting only volunteers for a study).
- **Loss to follow-up**: In longitudinal studies, if certain groups drop out of the study at higher rates than others (e.g., sicker patients), it could lead to biased results.
- **Exclusion of certain groups**: If certain groups are systematically excluded from the study (e.g., people from specific regions or demographics), the findings might not reflect the broader population.

Measurement Error

- The difference between the true value of a quantity and the value that is actually measured.
- It arises because all measurements are subject to various factors that introduce inaccuracies, such as instrument limitations, human errors, environmental influences, or sample variability.
- Can be classified into two main types:

  1. Systematic error/bias (or measurement bias)
  2. Random error (precision error)

Measurement Bias

- This type of error consistently skews the measurements in one direction, either too high or too low, and tends to be reproducible.
- It occurs due to factors like faulty equipment calibration, a biased measurement process, or incorrect measurement techniques.
- Measurement bias errors can often be identified and corrected.

Measurement Bias

Scenario: Measuring the length of a table with a ruler.

- **Error**: Suppose the ruler you are using is slightly stretched or warped, so all measurements are consistently a few millimeters longer than the true length.
- **Cause**: This error is due to a defect in the measuring instrument (the ruler), which causes all measurements to be biased in one direction (too long).
- **Effect**: If you repeatedly measure the table's length using this ruler, you will always get the same inaccurate result, and the measurements will be consistently wrong by a fixed amount.
- **Correction**: This error could be corrected by using a properly calibrated ruler or adjusting for the known stretch in the ruler.

Reminder: measurement bias results in consistent, repeatable inaccuracies in one direction, often due to equipment or procedural issues.

Precision Error

- This error occurs unpredictably and varies in magnitude and direction with each measurement.
- It is caused by uncontrollable factors like slight variations in the environment, small fluctuations in instruments, or minor inconsistencies in technique.
- Random errors are inherent to most measurement processes and can be reduced by averaging multiple measurements.

Precision Error

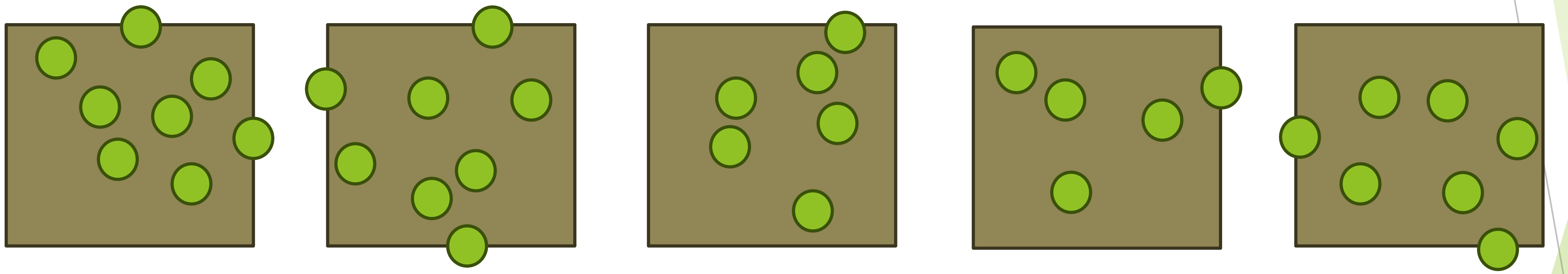
Scenario: Measuring the temperature of water with a thermometer.

- **Error**: When you dip the thermometer into the water, you get slightly different readings each time, even though the temperature is the same. For instance, one measurement reads 24.1°C, the next reads 24.3°C, and the next one is 24.2°C.
- **Cause**: This error is due to small, uncontrollable fluctuations such as slight air drafts, minor variations in the thermometer's sensitivity, or slight differences in how the thermometer is placed in the water.
- **Effect**: The error varies randomly in different trials, and no consistent pattern emerges in the direction of the error. It can be minimized by taking multiple measurements and averaging the results.
- **Correction**: Reducing random error involves techniques like increasing the precision of the instruments or taking repeated measurements and calculating the average.

Reminder: Precision error results in variability from measurement to measurement, due to uncontrollable or unpredictable factors, and averages out over time if multiple measurements are taken.

Example:

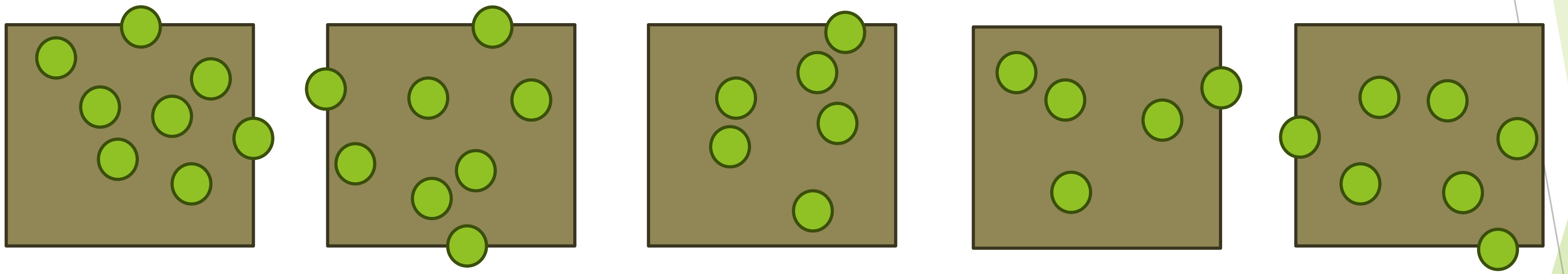
In surveys of vegetation areas are divided into smaller plots. A sample of the plots are selected and the number of plants in each section are recorded.



What is a potential issue in measuring these plots? How can we accommodate for it mathematically/in our study design?

Example:

In surveys of vegetation areas are divided into smaller plots. A sample of the plots are selected and the number of plants in each section are recorded.



What is a potential issue in measuring these plots? How can we accommodate for it mathematically/in our study design?

Go to **pollev.com/sta**

Sampling and Non-Sampling Errors

Sampling Errors:

- These errors arise because a sample is only a subset of the entire population. Since you are only studying a part of the population, there is always some level of variability between the sample and the population.

Non-sampling Errors:

- These errors occur for reasons other than the sampling process itself and can distort survey or research results even if a perfect sample is selected. They are more related to problems during data collection, measurement, or analysis.

Sampling Errors

Sampling Errors:

- occur because the sample is not the population.
- are random and arise due to chance. Different samples drawn from the same population could yield different results.
- can be reduced by increasing the sample size. (but this is usually infeasible)

Example: If you're surveying UofT STA304 students (N=550) about their sleeping habits and you randomly select 25 of them, the results (e.g., average bedtime) will likely not match the results (average bedtime) of the population (all 550 students), because we are missing 525 student responses.

Non-Sampling Errors

Non-Sampling Errors:

- arise due to issues such as improper data collection methods, mismeasurement, respondent bias, or data processing mistakes.
- are often systematic (biasing results in one direction) rather than random.
- can be minimized through careful survey design, ensuring accuracy in data collection, and controlling biases during analysis.
- more difficult to correct than sampling bias.

Non-Sampling Errors

Examples:

- **Non-response error:** Some individuals in the sample population do not respond, leading to biased results if the non-respondents are different from the respondents.
- **Measurement error:** Mistakes in how data is collected or recorded, such as faulty instruments, ambiguous survey questions, or incorrect responses from participants.
- **Processing error:** Errors during data entry, coding, or analysis.
- **Response bias:** Participants might not answer truthfully or could misunderstand the question.

Case Study (Mock Weekly Attendance)

On the next slide will include a Case Study to reflect on the topics covered in class today.

In small groups discuss the case study prompt and please enter your reflections in the Quercus Weekly Attendance (Not for grades).

Case Study

A university decides to conduct a survey to understand the study habits of its students and determine how they manage their time for studying. The survey asks questions about study frequency, average hours spent studying per week, and preferred study environments.

Survey Method:

- The university's research team decides to collect data from 200 students. A random sample is selected from the students who attend the university's **library during peak hours (10 AM to 4 PM)**. The survey is distributed via email, and students are asked to complete it voluntarily.

Findings:

- Of the 200 students surveyed, 140 respond, and 60 do not respond to the survey. Many of the students who responded mention that they prefer studying in quiet, distraction-free environments, and they spend about 15-20 hours per week studying.

In small groups discuss any selection, sampling, non-sampling and measurement biases. Please enter your answers in the Quercus "Attendance (Not for grades)"

Homework

Homework

- Week 1 – Quick Quiz will be released Friday and is due Monday.
  - This quiz will be about the materials released on Thursday.
  - Topics: Syllabus, materials covered in class today and materials released on Thurs
- Assignment 1 will be posted on or by Monday January 20
- Weekly lecture videos to be posted on Thursday.
- I have office hours on Zoom on Thursday (11am-12pm & 4pm-5pm).
- TA Office Hours start January 13