Course Info + Intro

STA304

Week 1 – Synchronous Lecture

Plan for this session

- Introductions
- Syllabus Info
- Survey
- Coding exercise
- Intro to Simulation

Logistics

Volunteer note-taker needed

To become a volunteer note-taker, please follow these steps:

1.Register Online as a Volunteer Note-Taker at: https://clockwork.studentlife.utoronto.ca/custom/misc/home.aspx

2.For a step-to-step guide please follow this link to the Volunteer Notetaking Portal Guide

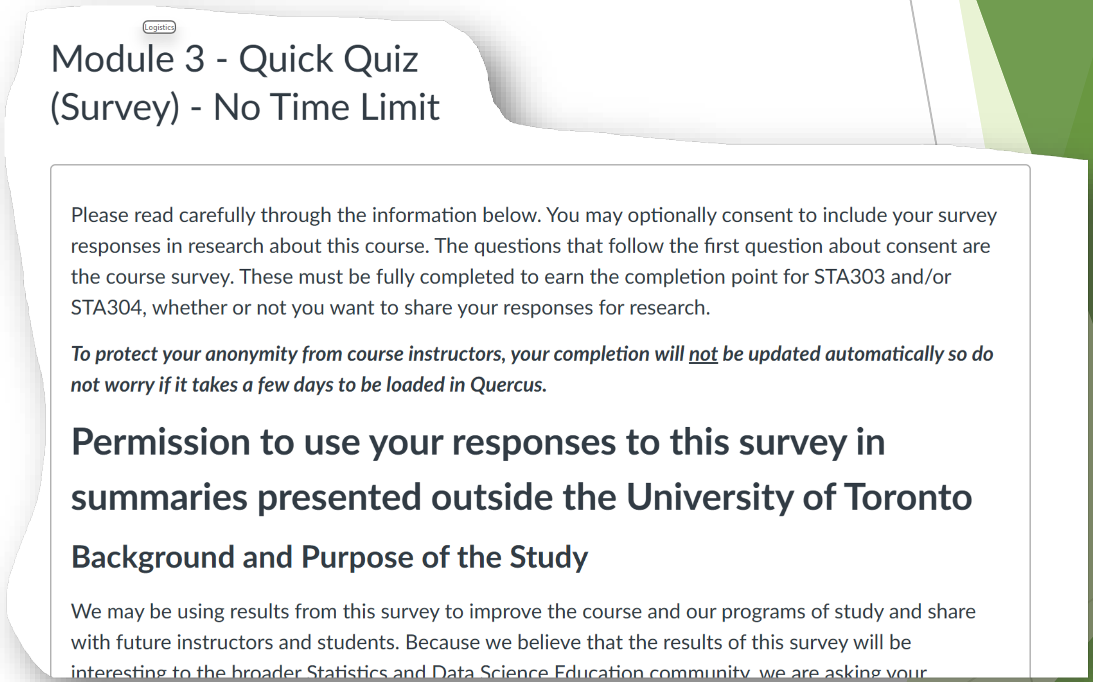
3.Click on *Volunteer Notetakers*, and sign in using your UTORid

4.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.

5.Start uploading notes.

Survey this weekend

- In lieu of a quiz, there will be a survey this weekend.
- No time limit.
- Due Monday at 11:59pm ET.
- Complete the survey to get the 2% of your grade.
- Will ask about consenting to sharing your submissions. If you do not, that is okay, and you will still receive 2% so long as you complete the rest of the survey.
- Taken in MS Forms, so there will be a delay in the grades being released.



Module 3 - Quick Quiz (Survey) - No Time Limit

Please read carefully through the information below. You may optionally consent to include your survey responses in research about this course. The questions that follow the first question about consent are the course survey. These must be fully completed to earn the completion point for STA303 and/or STA304, whether or not you want to share your responses for research.

To protect your anonymity from course instructors, your completion will not be updated automatically so do not worry if it takes a few days to be loaded in Quercus.

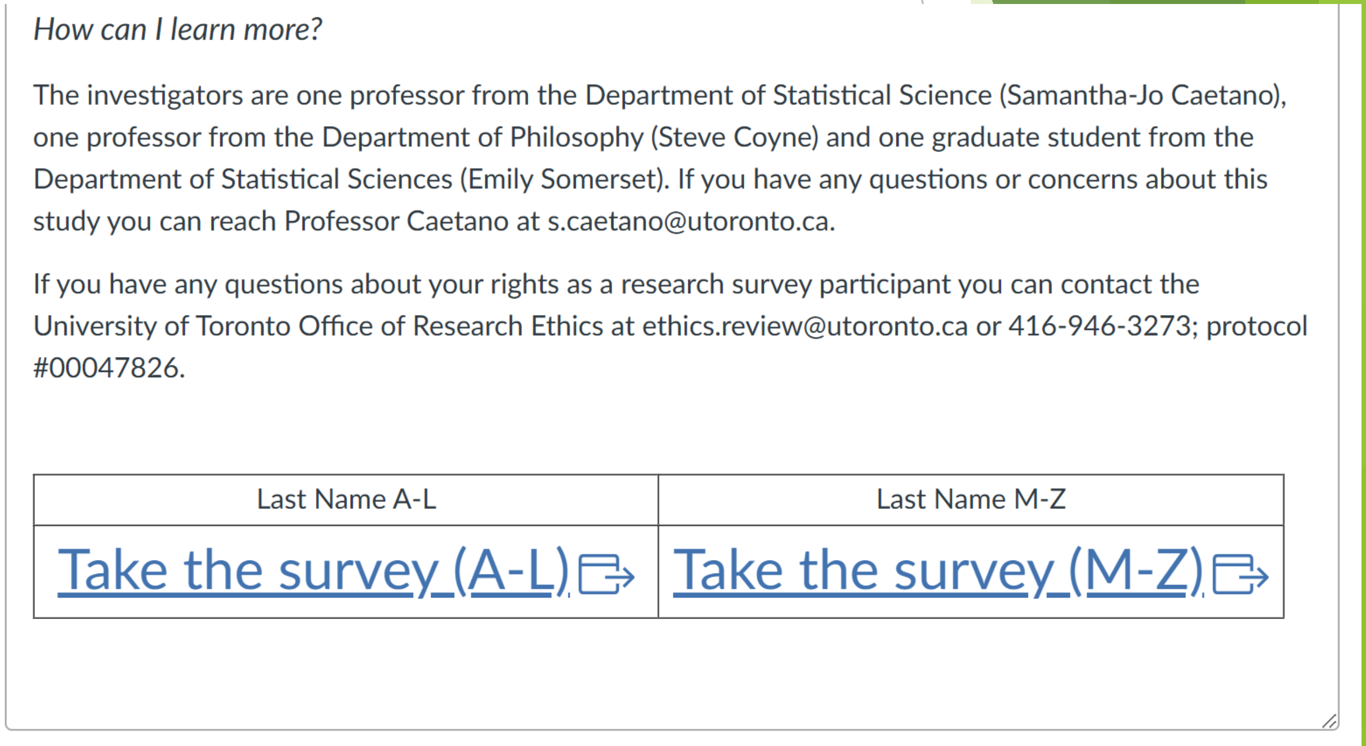
Permission to use your responses to this survey in summaries presented outside the University of Toronto

Background and Purpose of the Study

We may be using results from this survey to improve the course and our programs of study and share with future instructors and students. Because we believe that the results of this survey will be interesting to the broader Statistics and Data Science Education community, we are asking your

Survey this weekend

- There are two versions
- Be sure to take the correct version, based on your last name.



How can I learn more?

The investigators are one professor from the Department of Statistical Science (Samantha-Jo Caetano), one professor from the Department of Philosophy (Steve Coyne) and one graduate student from the Department of Statistical Sciences (Emily Somerset). If you have any questions or concerns about this study you can reach Professor Caetano at s.caetano@utoronto.ca.

If you have any questions about your rights as a research survey participant you can contact the University of Toronto Office of Research Ethics at ethics.review@utoronto.ca or 416-946-3273; protocol #00047826.

Last Name A-L	Last Name M-Z
Take the survey (A-L)	Take the survey (M-Z)

Next Week (Tues Jan 21)

- We will have a guest lecturer next week.
  - Steve Coyne from Philosophy to talk about Ethics in Survey Design

- I will still be here. We will do some coding demo and some statistical case studies.

- The in-class attendance will be pen and paper. **Please bring a writing utensil next class.**

Coding Demo

Let’s do some coding:

https://r.datatools.utoronto.ca/hub/user-redirect/git-pull?repo=https%3A%2F%2Fgithub.com%2FSamanthaJoCaetano%2FSTA304-W25-Module2.git&urlpath=rstudio%2F&branch=main

Sampling

What is sampling?

Sampling is the process of selecting a subset (or sample) from a larger population or data set to make inferences or draw conclusions about that population. The key idea is that it's often impractical or impossible to collect data from every individual in a large population, so instead, a smaller, manageable group is chosen.

Sampling is used in a variety of fields like statistics, research, and data analysis, and it helps to save time and resources while still providing meaningful insights.

What is sampling?

Sampling is the process of selecting a subset (or sample) from a larger population or data set to make inferences or draw conclusions about that population. The key idea is that it's often impractical or impossible to collect data from every individual in a large population, so instead, a smaller, manageable group is chosen.

Sampling is used in a variety of fields like statistics, research, and data analysis, and it helps to save time and resources while still providing meaningful insights.

There are two categories of sampling techniques we will discuss:

- Non-Probability based
- Probability based

Non-Probability-based sampling

In non-probability-based sampling, the selection process is not random, and some individuals in the population may have no chance of being selected. This means that the sample is not necessarily representative of the population, which can lead to bias. While it’s often quicker and easier to conduct, non-probability sampling is less reliable for making generalizations to the entire population.

Common types of non-probability-based sampling include:

- **Convenience Sampling**: Samples are selected based on ease of access or availability.
- **Judgmental or Purposive Sampling**: The researcher selects individuals based on their judgment or knowledge of the population.
- **Quota Sampling**: The sample is selected to meet specific quotas based on certain characteristics (e.g., gender, age), but without random selection.
- **Snowball Sampling**: Initial subjects are chosen, and then they help identify others, often used in studies with hard-to-reach populations.

Probability-based sampling

In probability-based sampling, every individual in the population has a known, non-zero chance of being selected. This approach uses randomization, which helps ensure that the sample is representative of the population. Because of this randomness, probability-based sampling tends to produce more reliable, unbiased results and allows for generalizations to be made from the sample to the larger population.

Common types of probability-based sampling include:

- **Simple Random Sampling**: Every individual has an equal chance of being selected.
- **Systematic Sampling**: A sample is selected by choosing every "kth" individual from a list.
- **Stratified Sampling**: The population is divided into subgroups (strata), and individuals are randomly selected from each subgroup.
- **Cluster Sampling**: The population is divided into clusters, and entire clusters are randomly selected.

Notes

Key Differences between: Non-Probability based and Probability based sampling:

- **Randomization**: Probability sampling relies on random selection, while non-probability sampling does not.
- **Representativeness**: Probability-based sampling tends to produce more representative samples, allowing for broader generalizations, whereas non-probability sampling may introduce bias.
- **Inference**: In probability-based sampling, it's possible to make statistical inferences about the population, but in non-probability sampling, this is typically not possible because the sample may not represent the population well.

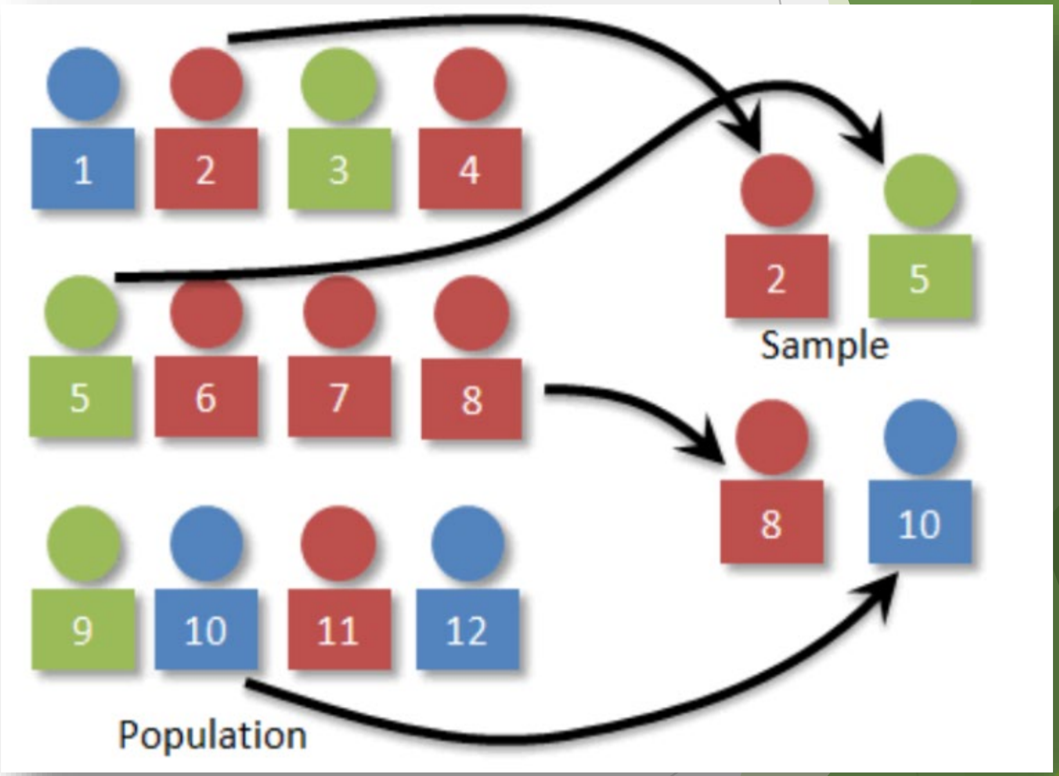
Simple Random Sampling (SRSs)

- Select units from the population at random. To ensure each unit has equal probability of being selected.

Examples:

Randomly selecting 10 patients in a hospital waiting room.

Randomly select 25 employees (out of 200) at a company, by selecting names from a hat.



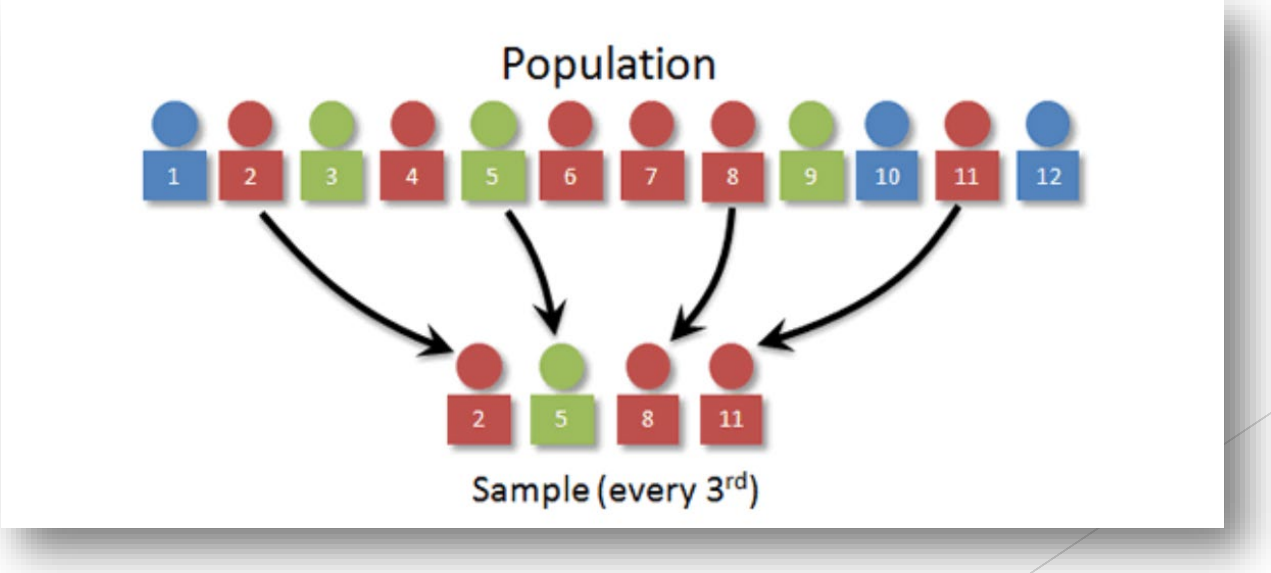
Systematic Sampling

- Sometimes we draw a sample by selecting individuals systematically (usually from a list/line). To make it random (i.e., probability-based sampling), you must still start the systematic selection from a randomly selected individual and then choose the unit in every N/n position on the list.

Examples:

• You have 100 people and want to select 5, so you randomly select the 14th name and then select the following names: 14, 34, 54, 74 and 94.

• Interview every 5th customer in line at the grocery store.

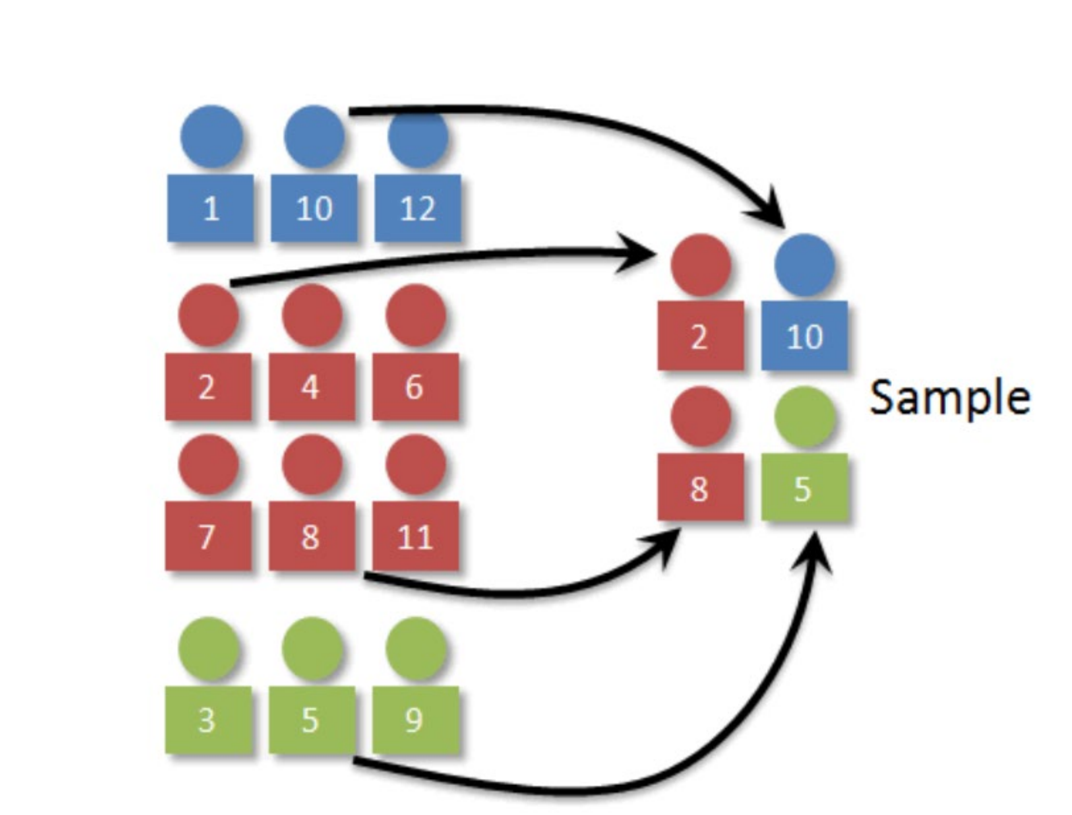


Stratified Random Sampling (or Stratified Sampling)

- The population is first sliced into homogeneous groups, called strata, before the sample is selected. Then simple random sampling is used within each stratum and the results are combined.

Example:

• A company that has 60 females and 30 males and 10 non-binary folk, will randomly select 6 females, 3 males and 1 non-binary folk to take a survey.



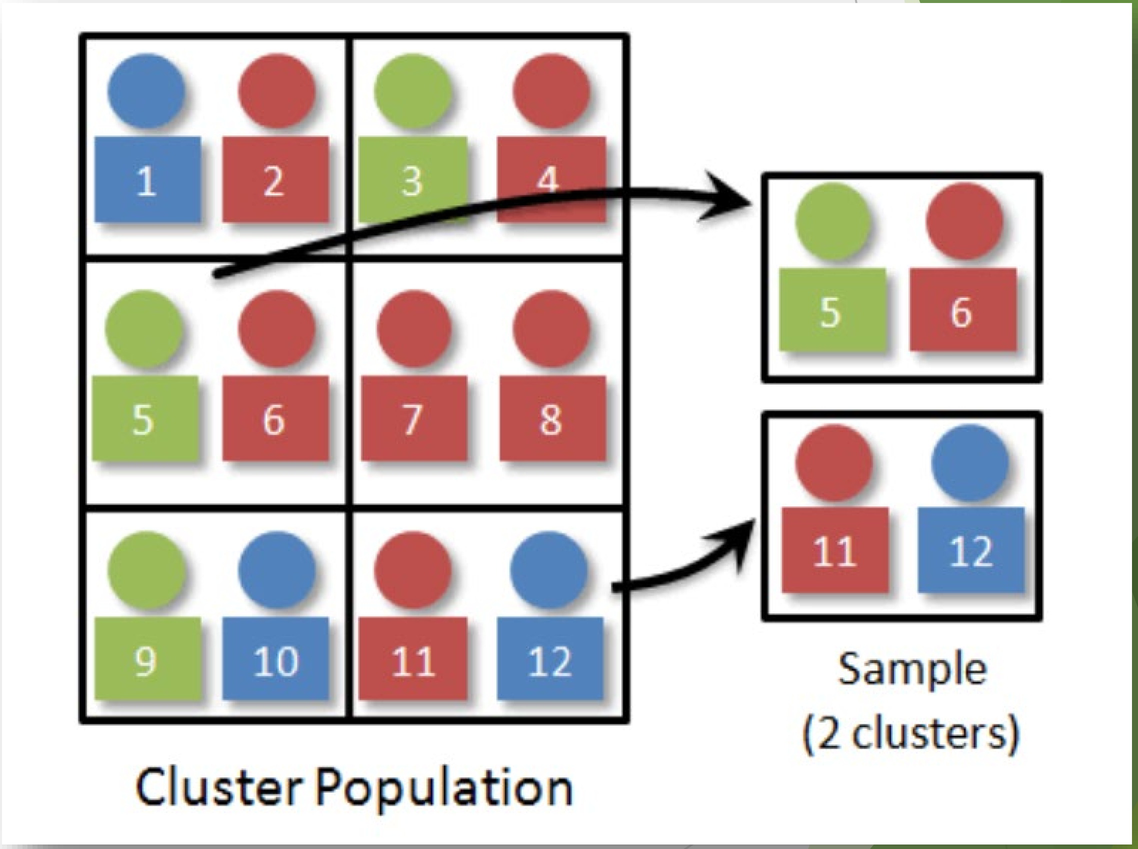
Cluster Sampling

- Splitting the population into similar parts or clusters can make sampling more practical. Then we select clusters at random and perform a census within each of them.

Examples:

• A school will randomly select 10 classes and survey all students in the selected classes.

• A cancer clinical trial will randomly select 5 hospitals and include all oncology patients within each hospital.



In-class Activity

Weekly Attendance Case Study

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 available and due at the end of class.

Weekly Attendance Case Study

Background: A high school is interested in conducting a survey to understand the students' opinions on the cafeteria food. The school principal has asked the research team to design a sampling method that will allow them to gather reliable and useful data from the student body. There are approximately 4000 students in the school, and the school has capacity to survey 400 students.

- **Sampling Method:** The research team is considering surveying the first 100 students in grade 9, the first 100 students in grade 10, the first 100 students in grade 11 and the first 100 students who are in grade 12, who enter the cafeteria during the lunch hour.

<u>Reflection:</u>

Is this a probability-based or non-probability-based sampling approach?

What is the name of this of sampling technique?

What are the pros of using this sampling technique?

What groups of students might be underrepresented by this approach?