Workshop: Intro to Sampling

DSI SUDS Skills Day

Ciara Zogheib

Outline for Today's Workshop

- The Sampling Basics
- Sampling Design
- Estimates and Sample Size
- Errors in Sampling
- Survey Quality
- Sampling Ethics

Today's Goal:

Be able to choose and justify a sampling approach based on your research objectives

1 The Sampling Basics

What is sampling and why do we do it? ❓

“Sampling consists of selecting some part of a population to observe so that one may estimate something about the whole population”

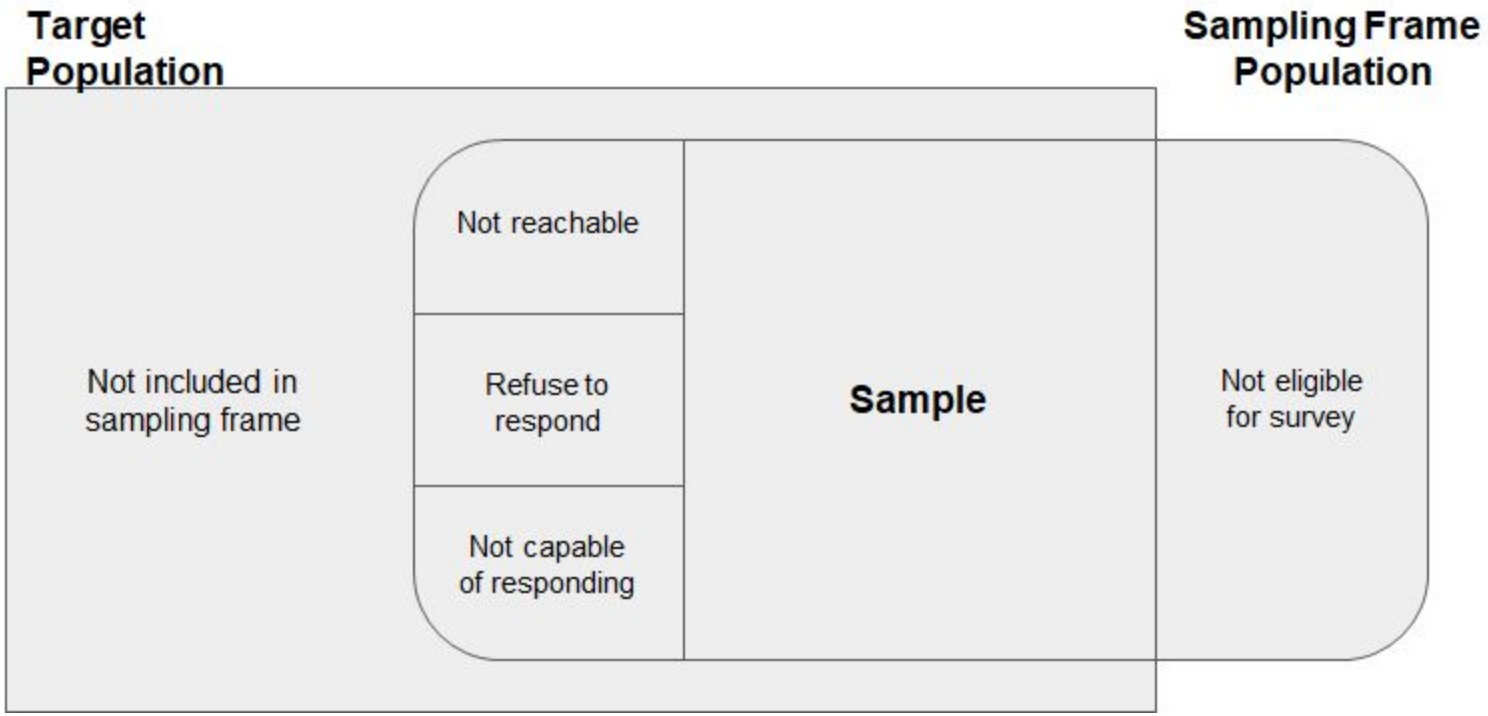
Thompson, ‘Sampling: Third Edition’ (2012)

What is a population?

- Population = Real or hypothetical set of units with characteristics of attributes that can be modeled by random variables
- Can be infinite or finite
- Inclusion and exclusion criteria must be clearly defined

Three levels of population

- Target population = all units covered by our study objective
- Frame population = all units covered by our **sampling frame** (the parts of our population we take our sample from)
- Sampled population = what we actually collected responses from


Target Population
Sampling Frame Population
Not included in sampling frame
Not reachable
Refuse to respond
Not capable of responding
Sample
Not eligible for survey

Why sampling?

- Lets us collect data from only a subset of our target population, called our sample, and make inferences about the whole population
- More cost- and time-efficient than sampling an entire population; usually easier and more realistic

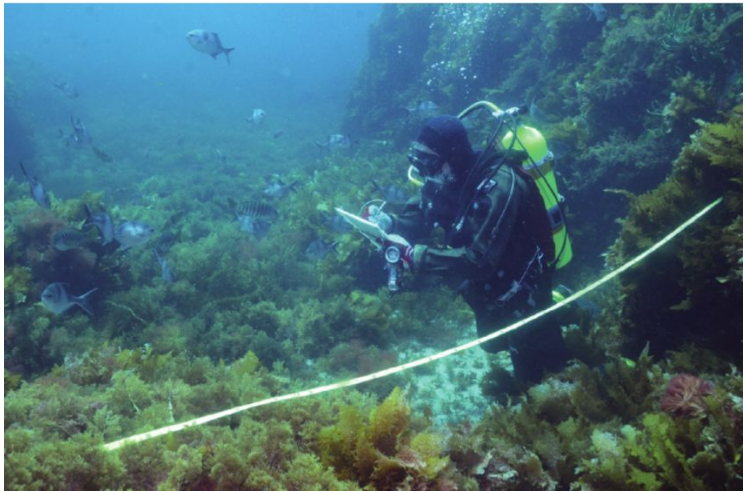
What's a survey?

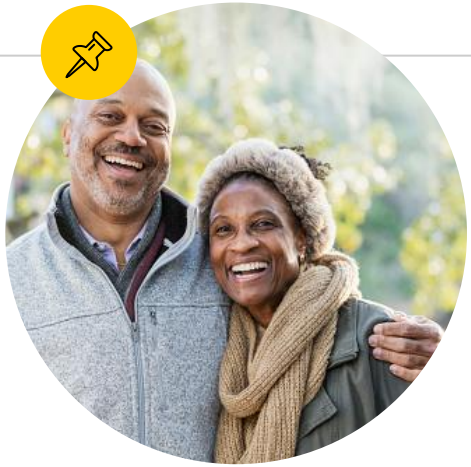
- "A survey is an investigation about the characteristics of a given population by means of collecting data from a sample of that population and estimating their characteristics through the systematic use of statistical methodology." (OECD)

What's a survey?

- "The term survey covers any activity that collects or acquires statistical data. Included are censuses, sample surveys, the collection of data from administrative records and derived statistical activities" (Statistics Canada)

What's a survey?





Activity

A pilot survey for The Canadian Longitudinal Study on Aging (CLSA) was conducted in the province of Ontario. The survey intended to cover the general population of the province with age 45–80 (inclusive). Survey questionnaires were sent to randomly selected individuals through regular mail. Individuals and their mailing addresses were selected and obtained from the Provincial Health Records.

What are the target population, sampling frame, sampling units, frame population, and sampled population?

Activity

A teacher wanted to find the average number of hours each week spent on watching TV by 4 and 5 year old children in Waterloo. She conducted a survey using the list of 123 kindergartens administered by the Waterloo Region District School Board. She first randomly selected ten kindergartens from the list. Within each kindergarten, she was able to obtain a complete list of all 4 and 5 year old children, with contact information for their guardians. She then randomly selected 50 children from the list and mailed the survey to their guardians. The sample data were compiled from completed and returned surveys..

What are the target population, sampling frame, sampling units, frame population, and sampled population?

2

Sampling Design

What types of samples are there and when do we use them?📊

Two types of sampling

- Probability sampling = entire target population has a chance of being randomly selected; results can represent the whole population
- Non-probability sampling = non-random selection (some in target population have no chance of selection); not generalizable but more convenient

Probability sampling

- All individuals/units in our population have a chance to be included in our sample, we can calculate the probability of any one of them being in the sample, and non-response must be random
- Good for quantitative research and projecting to the whole population

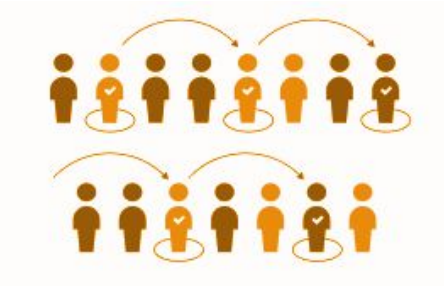
Types of probability sampling

Simple Random Sampling

Numbers are assigned to individuals in the population and randomly selected from sampling. Any group of *N* individuals is as likely to be chosen as any other group of *N* individuals.

Systematic Sampling

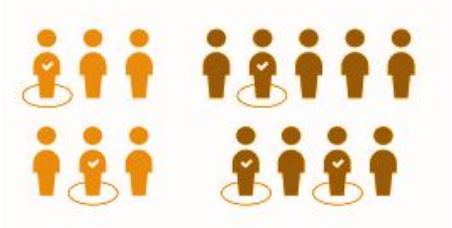
Sampling from a random starting point and then in fixed intervals.



Types of probability sampling

Stratified Sampling

Population is divided into groups called ‘strata’, then a proportionate number is taken from each stratum.

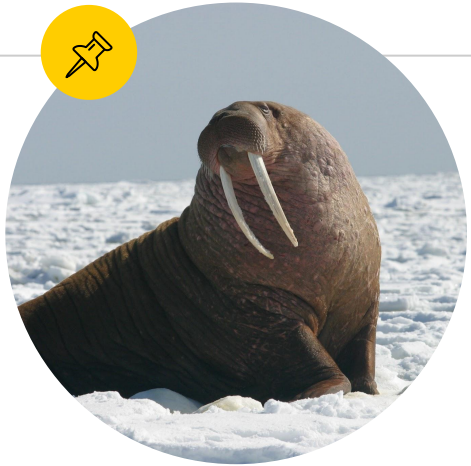


Cluster Sampling

The population is divided into clusters, a set of clusters are randomly selected, then every individual in the chosen clusters is included in the sample.

Multi-stage sampling

- Like cluster sampling or stratified sampling followed by an additional stage of simple random sampling
- Important: Each stage must involve a probability sampling method



Activity

What kind of stratification variable(s) would you use for each of the following situations? (AKA You want to do stratified sampling, what strata would you divide the population into?)

1. A political poll to estimate the percentage of registered voters in Arizona that approve of the governor's performance.
2. A sample of public libraries in Ontario to study the availability of computer resources, and the per capita expenditures.
3. An aerial survey to estimate the number of walrus in the pack ice near Alaska between 173 degrees East and 154 degrees West longitude.

Sampling with and without replacement

Consider a population of size N.

- **Simple random sampling with replacement:**
  1. Select one unit for measurement, with probability $1/N$ of being selected
  2. Sampled unit is returned to the population
  3. Select second unit for measurement, with probability $1/N$
  4. Repeat until desired sample size is obtained
- **Simple random sampling without replacement:**
  1. Select one unit for measurement, with probability $1/N$
  2. Select second unit for measurement, with probability $1/(N-1)$
  3. Repeat until desired sample size is obtained. The final unit in the sample will be selected with probability $1/(N-n+1)$, where n is total sample size.

Non-probability sampling

- Selecting units for our sample based on availability or some purpose → some of our population has no chance of being chosen for our sample and we don't know the exact portion we're excluding
- Means we can't estimate how much our sample is really representative of our population

Types of non-probability sampling

Convenience

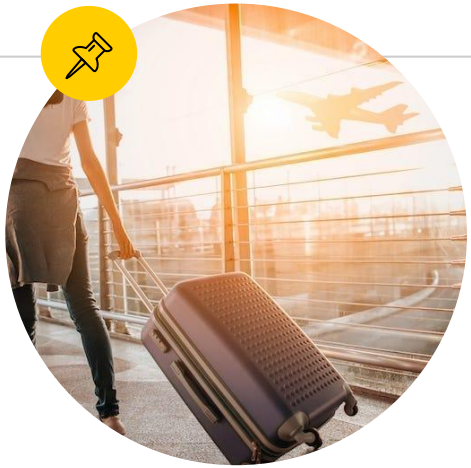
Selecting units from accessible populations, recruiting volunteers.

Purposeful

Choosing units with specific characteristics or expertise.

Snowball

Participants recruit other participants and so on until it ‘snowballs’.



Practice

Name the sampling method used in each of the following situations

A man in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. He does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all travelers who are sitting near gates and not taking naps while they wait.

Practice

Name the sampling method used in each of the following situations

A teacher wants to know if her students are doing homework, so she randomly selects rows two and five and then calls on all students in row two and all students in row five to present the solutions to homework problems to the class.

Practice

Name the sampling method used in each of the following situations

The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out asking for information about age, as well as about other variables of interest.

Practice

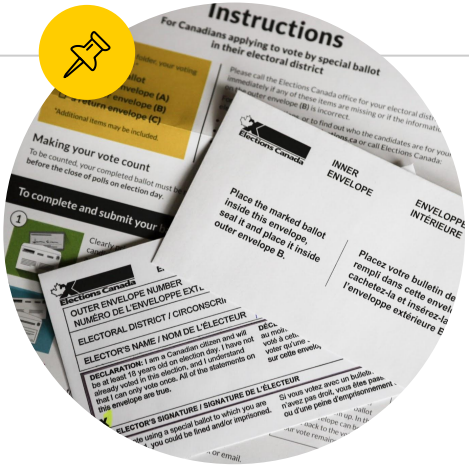
Name the sampling method used in each of the following situations

The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which they mark whether books are checked out by an adult or a child. They record this data for every fourth patron who checks out books.

Practice

Name the sampling method used in each of the following situations

A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked who they intend to vote for and whether the debate changed their opinion of the candidates.



Activity

Let's say there's an election for prime minister coming up and we as political researchers want to find out which candidate Toronto voters like the best. Assuming we want to give our survey to a sample rather than the entire voting population, **what might it look like to collect a simple random sample, a systematic sample, a cluster sample, and a convenience sample? What are some drawbacks or considerations of each approach?**

Respondent-driven sampling

- Researchers select a small number of individuals from the target population (“seeds”)
- Seeds recruit other members of the target population from their own social networks
- New sample members recruit additional sample members, and so on

Why use respondent-driven sampling

- Small target population size, or difficulty distinguishing target population from general population
- Difficulty accessing target population members

Hidden populations!

Studying hidden populations

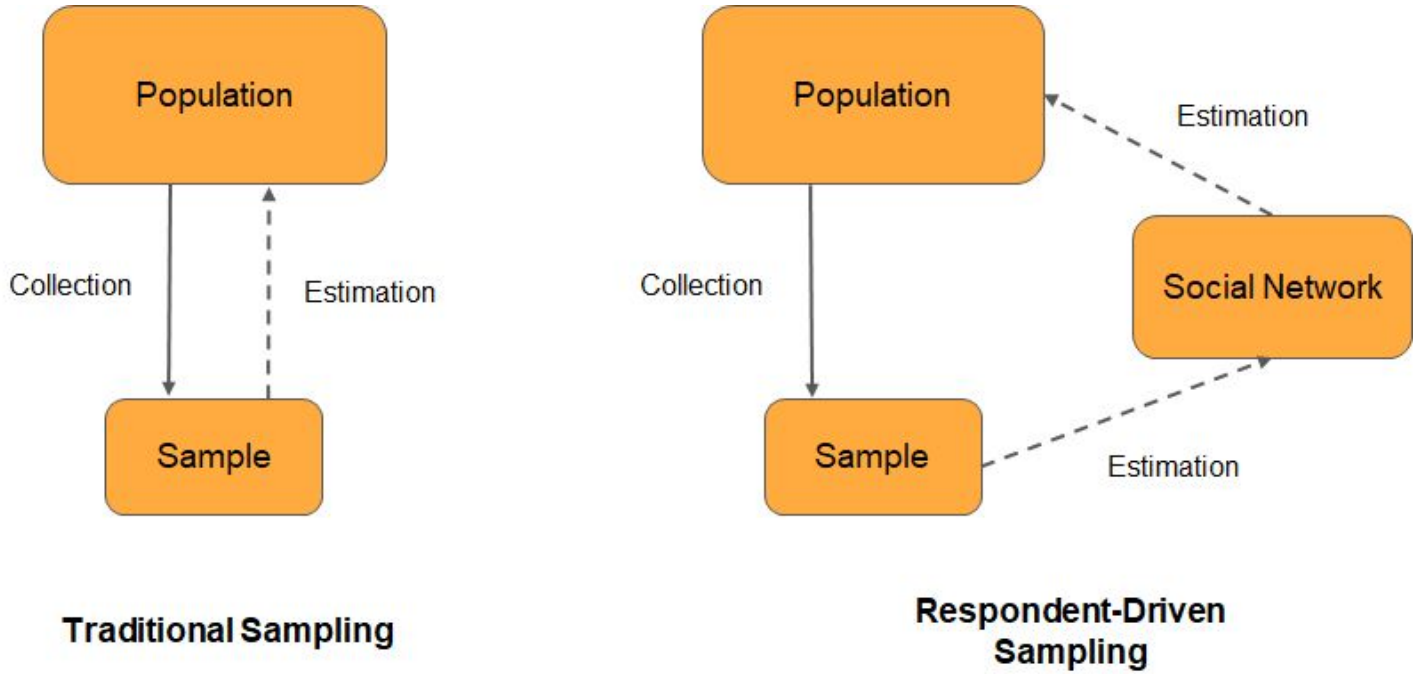
- Targeted sampling
- Time-space sampling

Respondent-driven sampling: Procedure

- Wave 0
  - Select initial **'seeds'** based on existing connections with target population.
  - Each seed is interviewed/surveyed.
  - Each seed receives c unique recruitment coupons to give to others in the target population.
  - Seeds are incentivised for participation and recruitment.
- Wave 1
  - Individuals who received coupons from original seeds are interviewed/surveyed.
  - Each individual receives c new recruitment coupons.
  - Wave 1 individuals are incentivised both for participation and recruitment.
- Recruitment and survey/interview process continues until desired sample size is reached.

Respondent-driven sampling: Procedure

Respondent-driven sampling



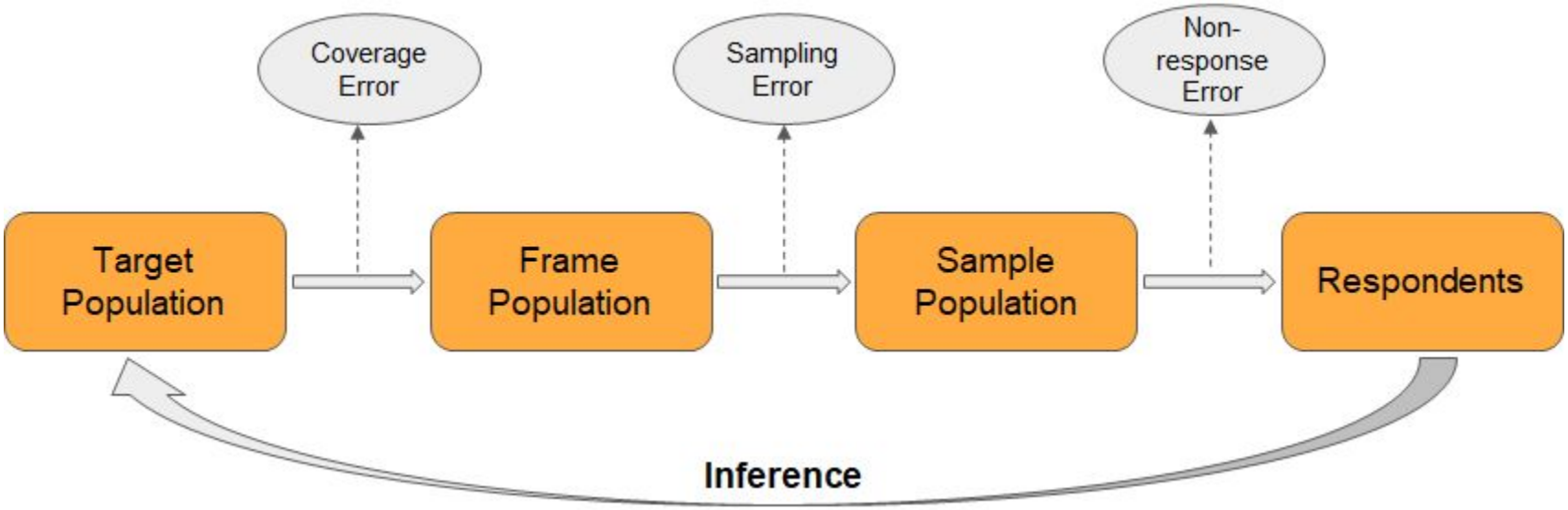
3 Estimates and Sample Size

You thought you could get through today without doing math 😰

What makes a good sample?

- Ideally: sample population = target population
  - Not likely
- Realistic best case: our sample is **representative**, AKA the characteristics of our sampled population match the characteristics of the target population

Representation



Mandatory Pre-Math Disclaimer Time

Sample size - general rules

- If target population <100, try to survey all of them
- ‘A good sample size is 10% of the population as long as it doesn’t exceed 1000’
- BUT - we don’t just want to use rules of thumb

Estimating sample size

1. Determine desired precision for the quantities we're going to estimate
2. How do our sample size *n* and desired precision relate mathematically?
3. Estimate the unknown qualities in our equation and solve for *n*

Precision

- **Confidence interval** = ‘margin of error’; the +/–
- **Confidence level** = how sure can we be that the actual value of interest falls within our confidence interval?
- Common confidence levels: 95%, 99%

Sample size equation

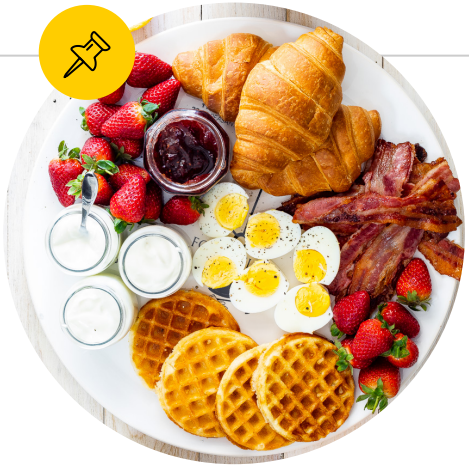
$$\text{Sample size} = \frac{(Z)^2 * (\text{Standard Deviation}) * (1 - \text{Standard Deviation})}{(\text{Margin of Error \%})^2}$$

Standard deviation

- **Standard deviation** = how much variability do we expect within our sample? (How spread out are our data from the sample mean?)
  - Different from standard error (how much will our sample mean vary from the population mean)
- When we're calculating sample size, we have to estimate standard deviation since we don't have our data yet
- Generally safe decision = 0.5

Z-score

- **Z-score** = lets us compare our value to the mean; ‘on a normal distribution, how many standard deviations is our value from the mean?’
- Good news: we can get our Z-scores using a Z table!
- 95% confidence level → Z-score = 1.96
- 99% confidence level → Z-score = 2.326



Practice

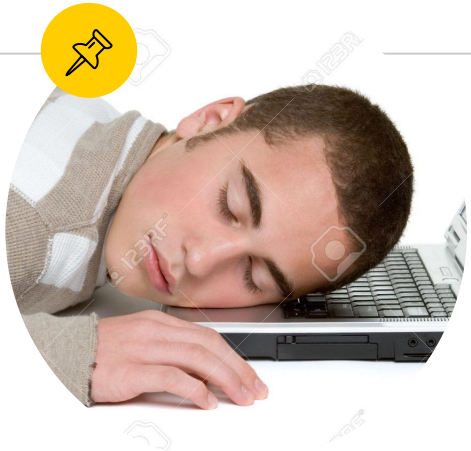
Suppose we are doing a study on students in residence at U of T, and want to find out how many students eat breakfast in the mornings. We don't have much information on the subject to begin with, so we're going to assume that half of the students eat breakfast: this gives us maximum variability (standard deviation = 0.5). We want 95% confidence (Z=1.96), and at least 5 percent—plus or minus—precision. **Calculate sample size.**

$$\text{Sample size} = \frac{(Z)^2 * (\text{Standard Deviation}) * (1 - \text{Standard Deviation})}{(\text{Margin of Error \%})^2}$$

Practice - sample size equation

$$\text{Sample size} = \frac{(1.96)^2 * (0.5) * (1 - 0.5)}{(0.05)^2}$$

Activity



As the next part of our residence study, we want to study whether students at U of T get 8 hours of sleep a night. We assume that half the students get 8 hours a night, for maximum variability, so our standard deviation = 0.5. We want 99% confidence (Z=2.326) and for our margin of error to be +/- 10%. **What sample size do we need?**

$$\text{Sample size} = \frac{(Z)^2 * (\text{Standard Deviation}) * (1 - \text{Standard Deviation})}{(\text{Margin of Error \%})^2}$$

Resources for sample size

- Sample size calculators:
  - https://www.surveysystem.com/sscalc.htm
  - https://www.openepi.com/SampleSize/SSPropor.htm
- Can also explore various survey packages in R and Python

Considerations for sample size

- Often the sample size we calculate will be much larger than is feasible
- Larger sample reduces sampling error BUT may increase non-sampling errors

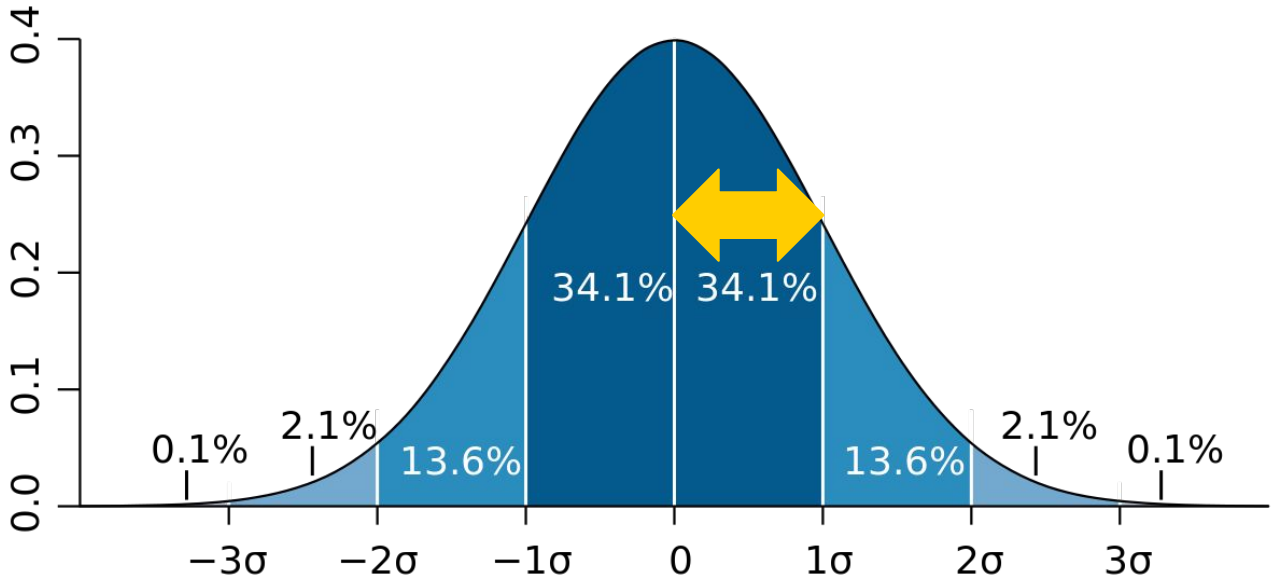
Estimating population mean

- Once we choose our sampling approach and identify the needed sample size, we collect our data
- Then it is time to estimate our population mean from our sample mean

Standard deviation

- Recall: standard deviation (s) is how we measure how spread out our sample data is around our sample mean

Standard deviation (visual)



Standard deviation calculation

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

1. Find our sample mean ($\bar{x}$)
2. From each value (x) subtract the sample mean and square the results
3. Sum the squared differences from the mean
4. Divide by N–1 where N is our sample population
5. Take the square root

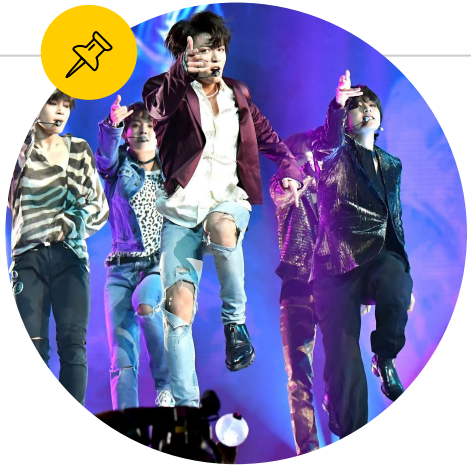
Margin of error calculation

$$\text{Margin of Error} = Z \times \frac{S}{\sqrt{n}}$$

Where *S* is standard deviation, *n* is our sample size, and *Z* is our Z-value (1.96 for 95%, 2.326 for 99%)

Estimating population mean

- Now we just add and subtract our margin of error from the sample mean to get our 95% confidence level ranges for our population mean
- So, if our sample size and approach was appropriate, we can say that if we took our sample 100 times, our population mean will be within the range of our sample mean +/- our margin of error 95 times



Activity - Step 1

We are interested in studying the heights of boyband members. We start by surveying a sample of boyband members and recording their heights. **What is the standard deviation of our sample?**

Name	Height (cm)
Namjoon	183
Michael	179
Niall	173
Joey	168
Justin	176
Nick	171

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Activity - Step 2

Now we need to estimate our population mean (average boyband member height) from our sample mean of 6 boyband members. We want to estimate with 95% confidence level (Z=1.96). **What is our population mean?**

$$\text{Margin of Error} = Z \times \frac{S}{\sqrt{n}}$$

4

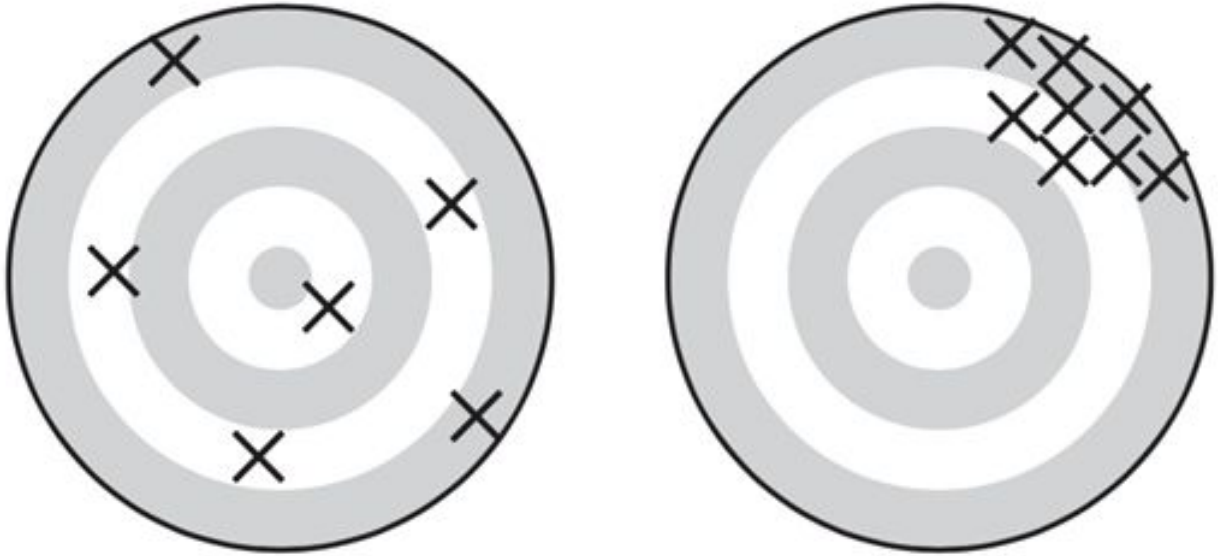
Errors in Sampling

How we mess up and how to avoid it 🙈

Variance vs. Bias

- **Variance** = Random error, describes the variability of our calculated quantities
- **Bias** = Systematic error, difference between the calculated and true quantities

Variance vs. Bias



Sampling Error

- Error that results from taking a given sample instead of measuring the whole population; error from variability
- General rule: smaller sample size produces greater sampling error

Non-sampling Error:

- Does **not** occur as a result of sample variability
- Often systemic (e.g. because of study design)
- Includes selection bias, measurement error, nonresponse

Selection bias

- **Selection bias** = when some population units are unintentionally excluded from the sample population
- Sources:
  - Selection procedure depends on some characteristic associated with what we're measuring
  - Substituting a convenient respondent when one is not available
  - Only sampling volunteers
  - Coverage error

Coverage error

- **Coverage error** = when the sampling frame does not match the target population
  - **Undercoverage** = Failing to include all the target population in the sample frame
  - **Overcoverage** = Including units in the sample frame that are not in the target population
- **Coverage bias** = when coverage error makes sample estimates differ from the population value

Measuring coverage

- A challenge!
- Compare to known values from the population, or
- Compare to external studies or data sources

Measurement error

- **Measurement error** = when survey responses differ from true population value
- **Measurement bias** = when measurement errors occur consistently in one direction
- Sources:
  - Respondent untruth, misunderstanding, or forgetfulness
  - Trying to impress an interviewer
  - Presentation (question order, interviewer persona)

Nonresponse

- **Nonresponse error** = Some members of our sample do not respond or participate; missing data for whatever reason
- **Nonresponse bias** = When respondents are systematically different from nonrespondents; means our results will not represent our whole target population

Types of nonresponse

- **Unit nonresponse** = When an entire sampling unit is missing (e.g. a person does not respond)
- **Item nonresponse** = When a specific measurement or variable for a given sampling unit is missing (e.g. a person responds but does not answer a particular question)

Consequences of nonresponse

- Non-representative samples
- Bias!
- Misallocation of resources

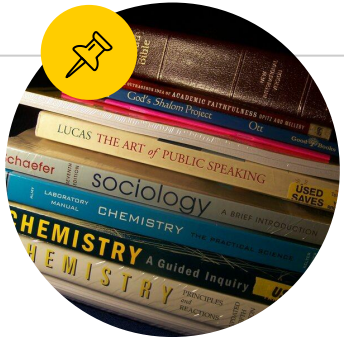
Processing error

- Errors in data entry or editing
- Sources:
  - Incorrect transcription
  - Typos in data entry
  - Open-ended question (e.g. trying to code multiple responses as a single response)

Total survey error

Total Survey Error = Representation Errors + Measurement Errors + Processing Errors

Basically: two types of errors (bias and variance), and sources of these errors include representation (coverage, sampling error, selection bias, nonresponse), measurement, and processing.



Activity

Suppose U of T has 10,000 part-time students (the population). We are interested in the average money a part-time student spends on books. We take two different samples. First, we use convenience sampling and survey ten students from an organic chemistry class. Many of these students are taking first term calculus. The amount of money they spend on books is as follows: $128, $87, $173, $116, $130, $204, $147, $189, $93, $153 The second sample is taken using a list of senior citizens who take P.E. classes and taking every fifth senior citizen on the list, for a total of ten senior citizens. They spend: $50, $40, $36, $15, $50, $100, $40 $53, $22, $22. It is unlikely that any student is in both samples.

Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population? Why?

Activity

A local radio station has a fanbase of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Asking all 20,000 listeners is an almost impossible task. The station uses convenience sampling and surveys the first 200 people they meet at one of the station's music concert events. 24 people said they'd prefer more talk shows, and 176 people said they'd prefer more music. **Do you think that this sample is representative of (or is characteristic of) the entire 20,000 listener population?**

Activity

Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys six flights from Toronto to Montreal to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.

1. **List three things wrong with the way the survey was conducted.**
2. **List three ways that you would improve the survey if it were to be repeated.**

5 Survey Quality

How to measure it and how to improve it ✅

Measures of Quality

- Relevance
- Accuracy
- Timeliness
- Accessibility
- Comparability
- Coherence
- Completeness

Validity

- How well does the data gathering method measure what it's supposed to measure?
- Types:
  - **Internal validity** = validity of results within our study
  - **External validity** = can our results be applied outside of our specific study (is it representative)?
- Sampling practice directly impact external validity

Questionnaire design

- Keep questions simple and clear
- Be specific in your questions
- Forced choice vs. ‘Select all that apply’
- Question order matters
  - Anchoring
  - Priming
- Avoid leading questions

Questionnaire design

- Consider social desirability bias
- No double negatives
- Avoid providing too many options for multiple choice questions
- No double-barrelled questions

Survey mode

- AKA the medium in which our survey is conducted
- For example: mail, email, telephone, website pop-up
- Different modes give different coverage issues, which we can address by
  - Comparing our sampling frame with external sources
  - Choosing a survey mode with high coverage **for our target population**

Sensitive questions

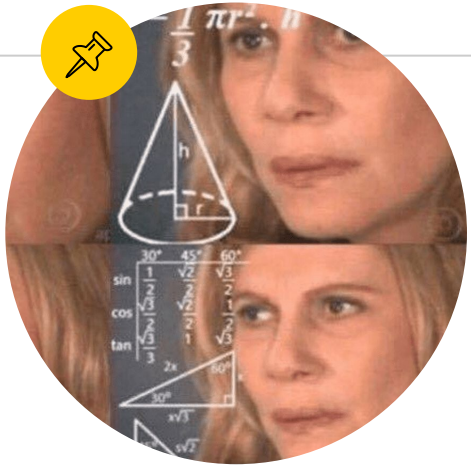
- Questions about:
  - Illegal behaviours
  - Information that poses risk to respondent if disclosed (health conditions, cheating on a partner)
  - Private information (income, home address)
  - Emotionally upsetting topics
  - Socially or politically charged topics

Asking sensitive questions

- Explain and inform
- Consider survey mode
- Who is asking the questions?
- Confidentiality and anonymity

More factors that can affect survey responses

- Distribution time
- Incentives
- Follow-up
- Pilot studies and ‘user testing’



Activity

To help me evaluate my performance as facilitator of this sampling workshop, I've designed a questionnaire. Unfortunately, I'm bad at survey design and this questionnaire is the worst document ever created. **Go through my survey and identify everything wrong with it and why each thing is bad.**

Resampling

- **Resampling** = estimating our parameter or variable of interest multiple times from our sample
- Resampling can help improve the accuracy of our estimate and to quantify the uncertainty of our estimates
- Can be computationally expensive

Types of resampling

- **Randomization** (shuffle observations among groups of samples)
- **Bootstrap** (drawing randomly with replacement from our sample)
- **Jackknife** (‘cutting’ our sample)

Resources for resampling

- Chapter 10 of Wu & Thompson (2020) [Free PDF download]: https://link.springer.com/book/10.1007/978-3-030-44246-0
- General overview of different resampling methods: http://strata.uga.edu/8370/lecturenotes/resampling.html
- Tutorial with example code for bootstrapping in R: https://towardsdatascience.com/a-practical-guide-to-bootstrap-with-r-examples-bd975ec6dcea
- Tutorial with example code for bootstrapping in Python (machine learning focused but still useful!): https://carpentries-incubator.github.io/machine-learning-novice-python/07-bootstrapping/index.html

Imputation

- Imputation = assigning values to missing items in a dataset
  - Reduces nonresponse bias, produces cleaner datasets
- Different methods:
  - Random sample from observations, mean of existing observations, linear regression, prediction based on other respondents with similar non-missing responses

Activity

We recorded song length and rating for a random sample of songs, but our dataset has some missing values. **Use mean imputation to complete the dataset.**

Song Name	Song Length in seconds	Song Rating out of 5
Start of Something New		3
Getcha Head In the Game	155	4
What I've Been Looking For	229	5
Stick to the Status Quo	122	
Bop to the Top		3
Breaking Free	59	2
We're All In This Together	249	

Activity

We recorded song length and rating for a random sample of songs, but our dataset has some missing values. **Use mean imputation to complete the dataset.**

Song Name	Song Length in seconds	Song Rating out of 5
Start of Something New	**163**	3
Getcha Head In the Game	155	4
What I've Been Looking For	229	5
Stick to the Status Quo	122	**3.4**
Bop to the Top	**163**	3
Breaking Free	59	2
We're All In This Together	249	**3.4**

Activity

Mean imputation is the simplest solution to missing data, but generally not the best. **Why can mean imputation be problematic?**

6 Sampling Ethics

Being a good person 101 💕

Ethics approval example

- Let’s go through a real example of U of T ethics review application form (approved and shared publicly)
- This is for doctoral research, so the bar/necessary rigour is quite high

Respondent burden

- **Respondent burden** = any risk, inconvenience, discomfort that participants will face from participating in a study or survey
- **Net burden** = when respondent burden > any personal benefit from participating in the study

Total Participant Burden = Net burden per participant x Number of participants

Respondent burden impacts

- Respondent burden can affect our data quality:
  - Nonresponse
  - Measurement error
  - Sustainability of survey

Study value

- **Study value** = expected social or clinical benefits of a given study
- Can also include benefits to individual respondents
- ‘Ethical seesaw’ → Minimize burden while maximizing chance of statistically valid results

Privacy


CONSOLIDATION
Personal Information Protection and Electronic Documents Act
CODIFICATION
Loi sur la protection
renseignements personn
les documents électroniques
TRI-COUNCIL POLICY STATEMENT
Ethical Conduct for Research Involving Humans
2014
Canadian Institutes of Health Research
Natural Sciences and Engineering Research
Social Sciences and Humanities Rese
Municipal Freedom of Information and Protection of Privacy Act
R.S.O. 1990, CHAPTER M.56
Personal Health Information Protection Act, 2004
ONTARIO REGULATION 329/04
GENERAL
Consolidation Period: From July 1, 2022 to the e-Laws currency date.
Last amendment: 423/22.

Privacy - Key terms

- **Data privacy** = the ability to control when and where your personal data are shared
- **Informational Risk** = potential for harm resulting from the disclosure and sharing of data
- **Personally identifying information (PII)** includes: name, address, telephone number, age, gender, personal opinions or views

Privacy - Key terms

- **Anonymization** = process of removing PII from a dataset
- **Confidentiality** = granted to respondents; only the researcher knows the identities of respondents
- **Anonymity** = granted to respondents; identities of individual respondents not known to researcher

Confidentiality considerations

- Data collection medium
- Data storage
- Who has access to the data?
- Retention and disposal schedule
- Clarify limits on confidentiality
- If no confidentiality – why?

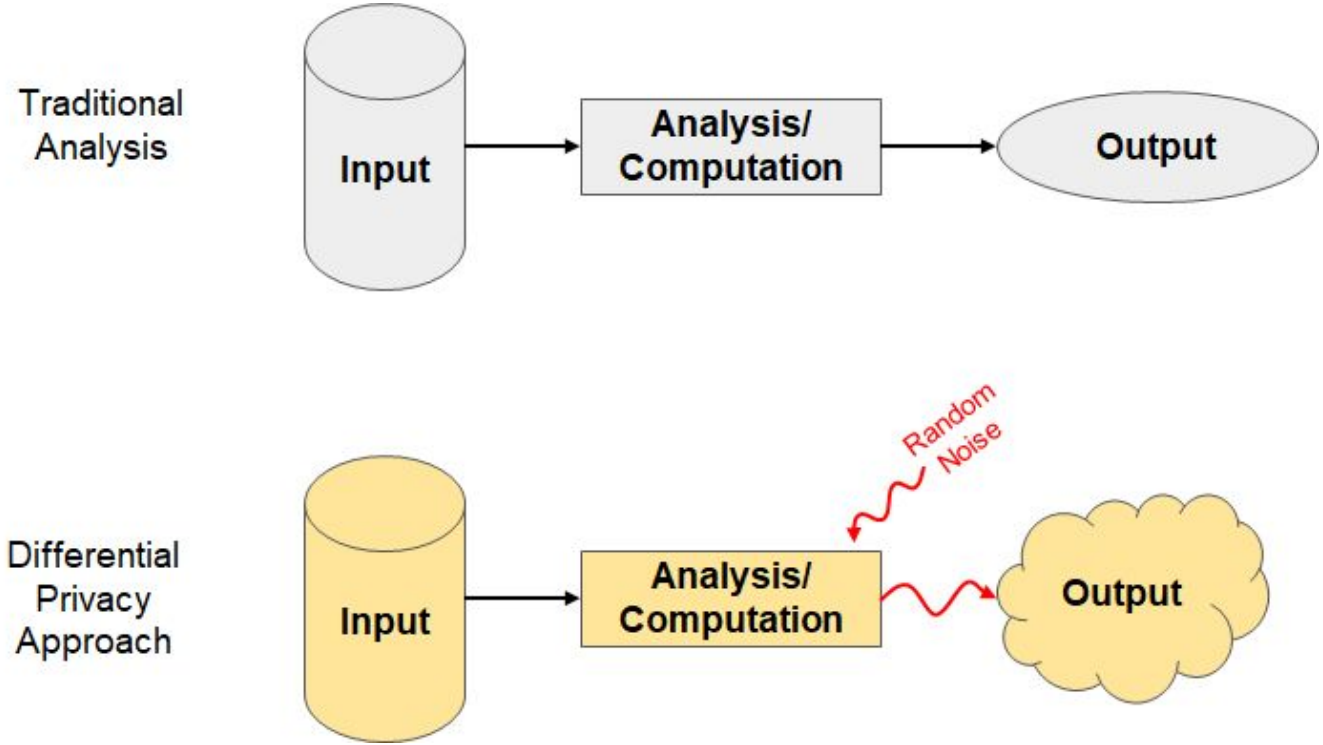
Privacy checklist

- Let's view an example from a public sector survey context!
- https://www.ipc.on.ca/wp-content/uploads/2015/04/best-practices-for-protecting-individual-privacy-in-conducting-survey-research.pdf

Differential privacy

- **Objective: analyze and share information about a dataset without revealing information about any individual within the dataset**
- Differential privacy adds random noise to computations on our dataset
- Randomness obscures the PII of any one individual respondent

Differential privacy



Differential privacy - Video

Reproducibility

- The ability to reproduce the results we computed
  - Reproducible in principle
  - Reproducible results
- One way to support reproducibility: open data!

Data documentation

- Having access to data behind a study is useless if you know nothing about how data were collected or manipulated prior to analysis
- **Data documentation** = recording all of the steps taken to obtain and process your data

What should we document?

What should we document?

- Populations
- Sampling methodology
- Mode
- Timeline
- Response rate
- Data cleaning procedure
- Accuracy

How should we document?

- Share raw data files (Privacy!!)
- Summarize all steps taken during the survey design and sampling
- Keep all original questionnaire or survey materials
- Make available code that was used for processing
- Be sure code is reproducible by using comments, seeds, etc.

Sampling Project

Apply everything we've learned today at hyperspeed

	Canadian Election Study Online Survey, 2019
Sample type *SRS, cluster sample, etc. Include any relevant additional information, like how strata were defined.*	
Sample size	
Target population	
Sampling frame	
Survey mode(s)	
Timeline	
Data processing *Cleaning, imputation, etc.*	
Sources of error *Limitations, known biases*	
Ethics or Privacy concerns	

2019 Canadian Election Study: Online Survey

Stephenson, L; Harell, A; Rubenson, D; Loewen, P, https://doi.org/10.7910/DVN/DUS88V/HRZ21G

Today's Goal:

Be able to choose and justify a sampling approach based on your research objectives