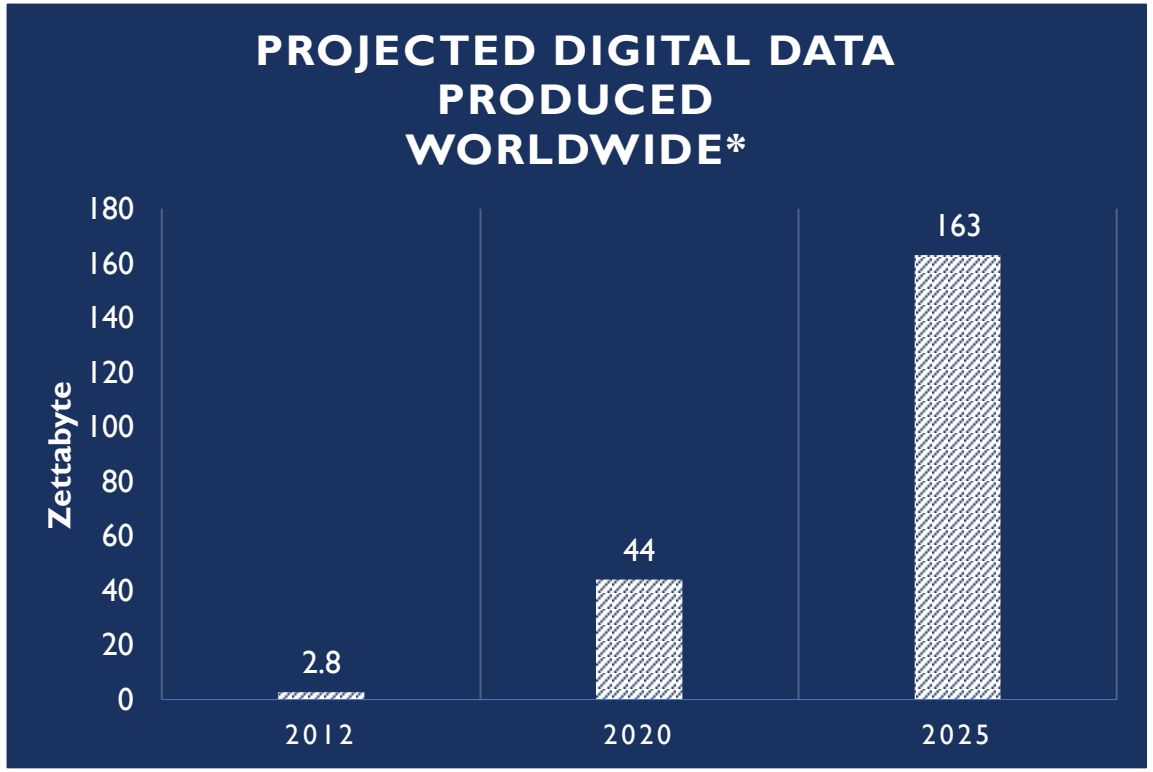
GATHERING DATA

ST101 – DR. ARIC LABARR

GATHERING DATA

- Data is everywhere.
- With all this data being gathered and stored, we need to understand good practices of gathering data.
- Data gathered without thinking ahead of time leaves itself open for problems later.



*IDC *Digital Universe*

GATHERING DATA

- Main concepts in this section of the course:
  - Samples and populations
  - Randomness
  - Good vs. bad sampling methods
  - Ethical concerns around data

WHY DO WE CARE?

- Why are we collecting data?

WHY DO WE CARE?

- Why are we collecting data?
  - Make better decisions around a group of people, places, things, etc.

- Why would data help with that?

WHY DO WE CARE?

- Why are we collecting data?
  - Make better decisions around a group of people, places, things, etc.
- Why would data help with that?
  - If data represents the things we are interested in, it can provide insights.

WHY DO WE CARE?

- Why are we collecting data?
  - Make better decisions around a group of people, places, things, etc.

- Why would data help with that?
  - **If data represents the things we are interested in**, it can provide insights.

This is NOT trivial!
It is foundational!

WHY DO WE CARE?

- Why are we collecting data?
  - Make better decisions around a group of people, places, things, etc.

- Why would data help with that?
  - If data represents the things we are interested in, it can provide insights.
  - If data **doesn't** represent the things we are interested in, it can provide misleading results and lead to incorrect decisions.

EXAMPLE – HEIGHT

- Imagine you wanted to know the average height of the adult population in the United States because you are designing a new clothing line for adults.
- You take a **sample** of people (subset of people) since you think it will be impossible to ask everyone in the United States what their height is.
- Your sample consists entirely of professional basketball players.
- Do you see any problem here?

EXAMPLE – HEIGHT

- Do you see any problem here?
  - Professional basketball players are probably taller than most adults in the United States.
  - If their heights are taller, then our guess will be too tall!
  - Clothes will not be designed for common adults to wear which will lead to poor sales numbers and wasted resources on producing many clothes that not many people buy.

EXAMPLE – HEIGHT

- The data we made decisions from did not represent the people we wanted to serve!
- The data wasn't bad, just collected in a way that didn't provide the insights we wanted.
- How do we ensure we don't make this mistake?
  - Samples and populations
  - Randomness
  - Good vs. bad sampling methods

SUMMARY

- Data gathered without thinking ahead of time leaves itself open for problems later.
- If data represents the things we are interested in, it can provide insights.
- If data doesn't represent the things we are interested in, it can provide misleading results and lead to incorrect decisions.

SAMPLES AND POPULATIONS

GATHERING DATA

GATHERING DATA

- Before we start gathering data, it is good for us to know who or what we are interested in gathering information about.
- Should also consider what we want to know about this group we are interested in.

POPULATION

- **Population** – set of all objects/individuals of interest.
- Usually too large to obtain information from entire population.
- Example:
  - Want to know average height of **adults in the United States**.
  - Impossible to actually get information from all adults in United States.

POPULATION

- **Population** – set of all objects/individuals of interest.
- Usually too large to obtain information from entire population.
- Example:
  - Want to know average height of **adults in the United States**.
  - Impossible to actually get information from all adults in United States.

- Obtaining information from the whole population is called a **census**.

POPULATION DETAILS

- **Population** – set of all objects/individuals of interest.
- Example:
  - Want to know average height of adults in the United States.

- Must pay attention to **details** of the population.
  - What do you consider an adult?
  - If this is for marketing a new clothing line, do you want ALL adults? Adults of certain age range? Business or casual? Certain region of the country?
  - Lots of problems with sampling comes from not fully defining the population.

POPULATION PARAMETER

- Population – set of all objects/individuals of interest.
- Example:
  - Want to know **average height** of adults in the United States.

- **Parameter** – measures computed from a population.

SAMPLE

- **Sample** – subset of the population that information is actually obtained.
  - Should represent the population well.

- **Sampling frame** – actual list from which the sample is taken.
  - May not equal the population.

SAMPLE STATISTIC

- Sample – subset of the population that information is actually obtained.
  - Should represent the population well.

- **Statistic** – measures computed from a sample.
- Sample statistics is the **point estimate** of the population parameter.
  - Point estimate is a single number estimate of an unknown parameter.

PUT IT ALL TOGETHER

Population

Sample

- Population – set of all objects/individuals of interest.

Parameter

Statistic

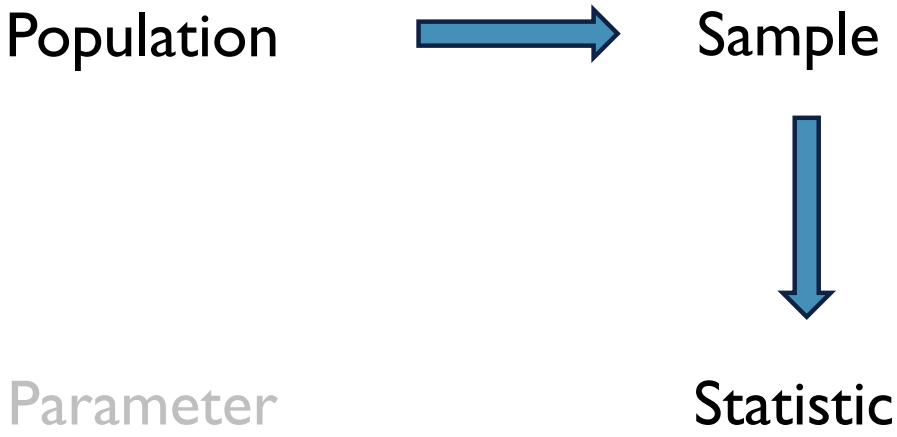
PUT IT ALL TOGETHER

Population → Sample

Parameter Statistic

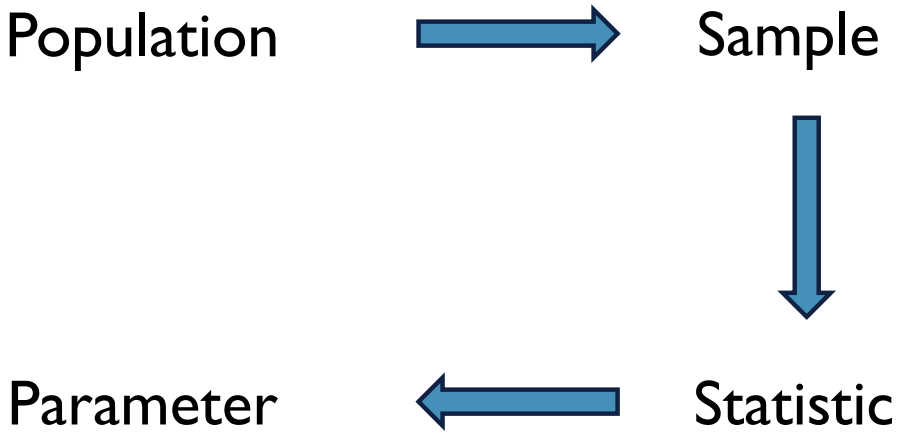
- Population – set of all objects/individuals of interest.
- Sample – subset of the population that information is actually obtained.

PUT IT ALL TOGETHER



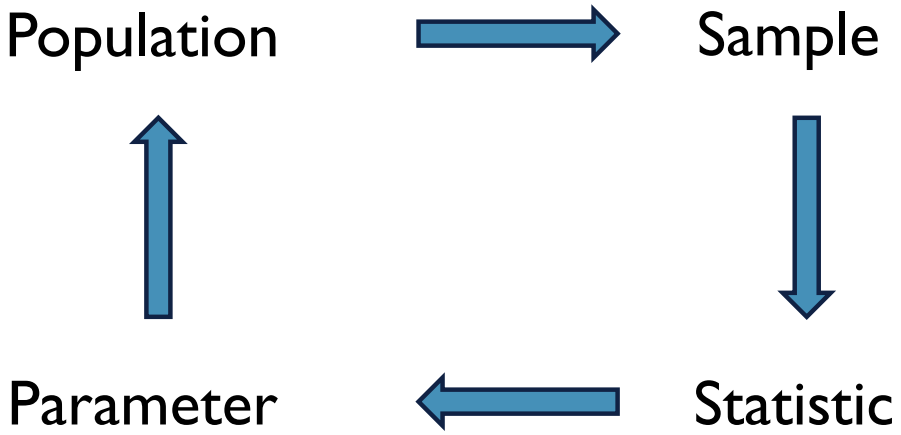
- Population – set of all objects/individuals of interest.
- Sample – subset of the population that information is actually obtained.
- Statistic – measures computed from a sample.

PUT IT ALL TOGETHER



- Population – set of all objects/individuals of interest.
- Sample – subset of the population that information is actually obtained.
- Statistic – measures computed from a sample.
- Parameter – measures computed from a population.

PUT IT ALL TOGETHER



- Population – set of all objects/individuals of interest.
- Sample – subset of the population that information is actually obtained.
- Statistic – measures computed from a sample.
- Parameter – measures computed from a population.

EXAMPLE

- A retail chain is trying to determine if a new product they introduced is selling well across their stores. The retail chain has 2135 stores nationwide. The analyst in charge of this project is tasked to estimate the average daily sales of this new product across all stores. Older computing technology forces the company to randomly pick 179 stores spread evenly throughout the nation to calculate gather data from. The average daily sales from these 179 stores is $129.19.

- Identify population, sample, parameter, statistic.
- Any sampling frame issues?

EXAMPLE

- A retail chain is trying to determine if a new product they introduced is selling well across their stores. The retail chain has **2135 stores nationwide**. The analyst in charge of this project is tasked to estimate the average daily sales of this new product across all stores. Older computing technology forces the company to randomly pick 179 stores spread evenly throughout the nation to calculate gather data from. The average daily sales from these 179 stores is $129.19.

- Identify **population**, sample, parameter, statistic.
- Any sampling frame issues?

EXAMPLE

- A retail chain is trying to determine if a new product they introduced is selling well across their stores. The retail chain has 2135 stores nationwide. The analyst in charge of this project is tasked to estimate the average daily sales of this new product across all stores. Older computing technology forces the company to randomly pick **179 stores spread evenly throughout the nation** to calculate gather data from. The average daily sales from these 179 stores is $129.19.

- Identify population, **sample**, parameter, statistic.
- Any sampling frame issues?

EXAMPLE

- A retail chain is trying to determine if a new product they introduced is selling well across their stores. The retail chain has 2135 stores nationwide. The analyst in charge of this project is tasked to estimate **the average daily sales of this new product across all stores**. Older computing technology forces the company to randomly pick 179 stores spread evenly throughout the nation to calculate gather data from. The average daily sales from these 179 stores is $129.19.

- Identify population, sample, **parameter**, statistic.
- Any sampling frame issues?

EXAMPLE

- A retail chain is trying to determine if a new product they introduced is selling well across their stores. The retail chain has 2135 stores nationwide. The analyst in charge of this project is tasked to estimate the average daily sales of this new product across all stores. Older computing technology forces the company to randomly pick 179 stores spread evenly throughout the nation to calculate gather data from. The average daily sales from these 179 stores is **$129.19**.

- Identify population, sample, parameter, **statistic**.
- Any sampling frame issues?

EXAMPLE

- A retail chain is trying to determine if a new product they introduced is selling well across their stores. The retail chain has 2135 stores nationwide. The analyst in charge of this project is tasked to estimate the average daily sales of this new product across all stores. Older computing technology forces the company to randomly pick 179 stores spread evenly throughout the nation to calculate gather data from. The average daily sales from these 179 stores is $129.19.

- Identify population, sample, parameter, statistic.
- Any sampling frame issues? **NO**

SUMMARY

- Population – set of all objects/individuals of interest.
- Sample – subset of the population that information is actually obtained.
  - Sampling frame – actual list from which the sample is taken.
- Statistic – measures computed from a sample.
- Parameter – measures computed from a population.

RANDOMNESS

GATHERING DATA

EXAMPLE

- A retail chain is trying to determine if a new product they introduced is selling well across their stores. The retail chain has 2135 stores nationwide. The analyst in charge of this project is tasked to estimate the average daily sales of this new product across all stores. Older computing technology forces the company to **randomly** pick 179 stores spread evenly throughout the nation to calculate gather data from. The average daily sales from these 179 stores is $129.19.

- Identify population, sample, parameter, statistic.
- Any sampling frame issues? NO

RANDOMNESS

- What do you think of with randomness?

RANDOMNESS

- What do you think of with randomness?
  1. Not knowing what is going to happen…
  2. Fairness (equal chance for outcomes)…

RANDOMNESS

- **Random** – an outcome is random if we know the particular outcomes that something could have but are unsure of which of those outcomes is about to happen.

RANDOMNESS

- **Random** – an outcome is random if we know the particular outcomes that something could have but are unsure of which of those outcomes is about to happen.

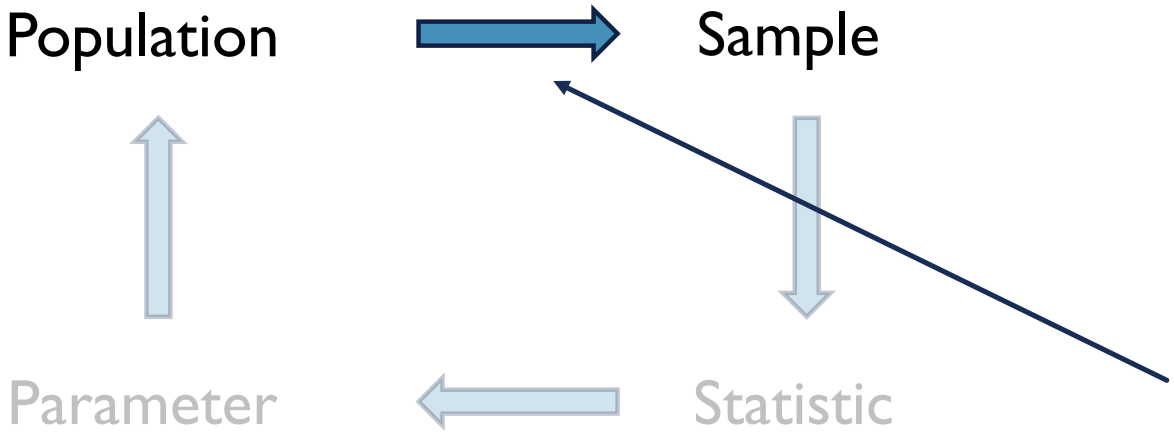
- Not knowing what is going to happen…
  - Kind of true. We know what **could** happen, but not which of the outcomes will happen.
  - Flip a fair coin → could be heads or tails, but not sure which.

RANDOMNESS

- **Random** – an outcome is random if we know the particular outcomes that something could have but are unsure of which of those outcomes is about to happen.

- Not knowing what is going to happen…
  - Kind of true. We know what **could** happen, but not which of the outcomes will happen.
  - Flip a fair coin → could be heads or tails, but not sure which.
- Fairness (equal chance of outcomes)…
  - Could be true, but not required.
  - An unfair coin is still random, but the outcomes are not even.

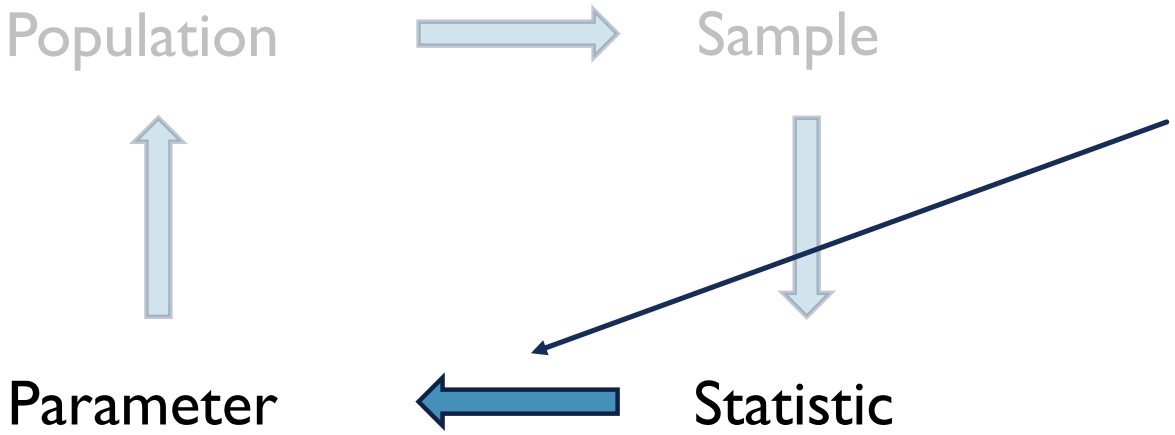
RANDOMNESS AND SAMPLING



Having randomness helps make the sample representative of the population.

Protects us from having certain pieces of information overly influence our sample.

RANDOMNESS AND SAMPLING



Having a good representative sample means the inference we make from the statistic to the parameter is reasonable!

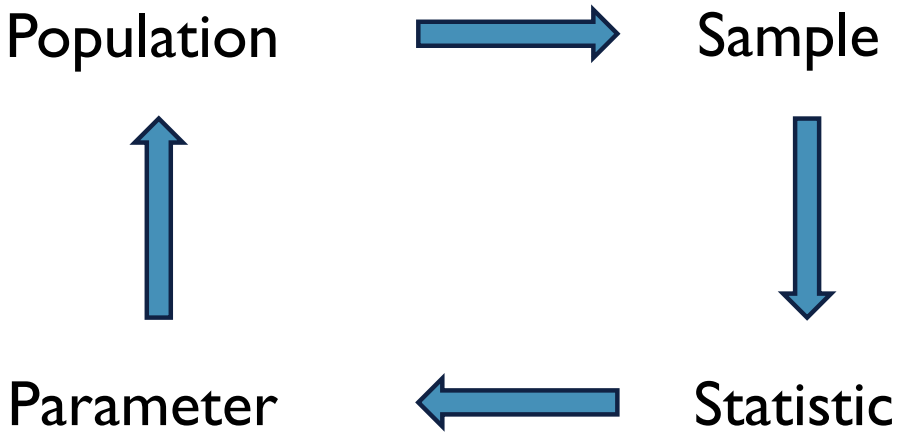
SUMMARY

- Random – an outcome is random if we know the particular outcomes that something could have but are unsure of which of those outcomes is about to happen.
- Having randomness helps make the sample representative of the population.
- Having a good representative sample means the inference we make from the statistic to the parameter is reasonable.

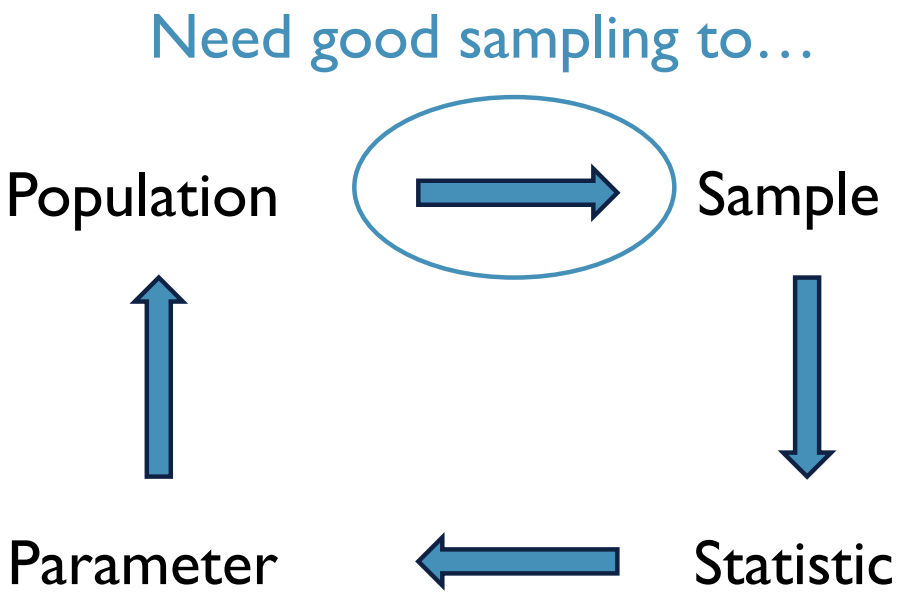
BAD SAMPLING METHODS

GATHERING DATA

PARAMETERS VS. STATISTICS

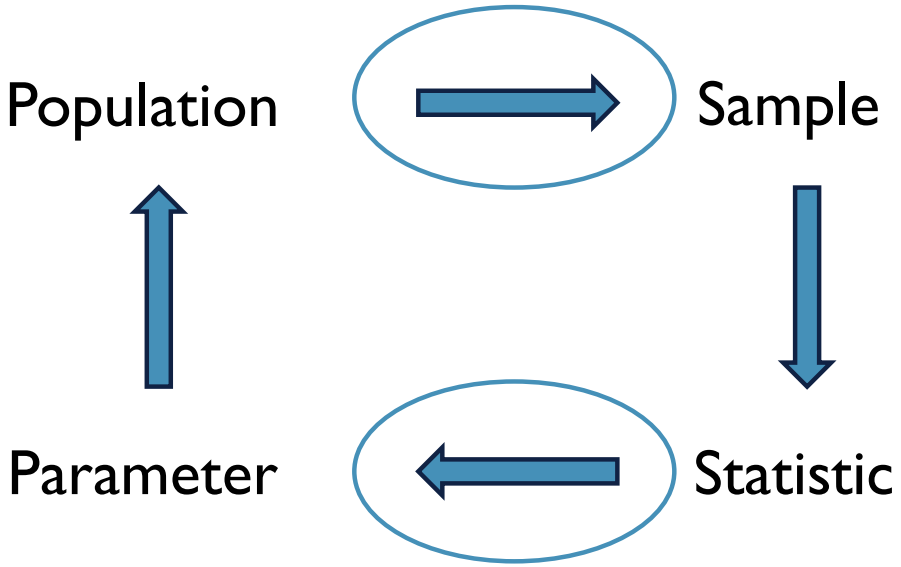


PARAMETERS VS. STATISTICS



PARAMETERS VS. STATISTICS

Need good sampling to…



…have good estimates.

SAMPLING

- There are many different ways to sample data from population.
- Mistakes in sampling can lead to **bias** in the sample.

- **Bias** – certain outcomes are favored over other outcomes in samples.

TYPES OF BIAS

- **Bias** – certain outcomes are favored over other outcomes in samples.
- 2 Common Types of Bias:
  1. Selection Bias
  2. Sampling Bias

TYPES OF BIAS

- **Bias** – certain outcomes are favored over other outcomes in samples.
- 2 Common Types of Bias:
  1. Selection Bias
  2. Sampling Bias

TYPES OF BIAS

- **Bias** – certain outcomes are favored over other outcomes in samples.
- 2 Common Types of Bias:
  1. Selection Bias
     a) Undercoverage
     b) Nonresponse
  2. Sampling Bias

UNDERCOVERAGE

- **Undercoverage** – sampling frame and population are not equal.

- Problem:
  - Sample doesn't represent the population of interest.
  - Incorrect and biased inference is made.

- Example – Phone book.

NONRESPONSE

- **Nonresponse** – subject in sample cannot / will not respond or be measured.

- Problem:
  - Those who respond don't represent the population as a whole.
  - Incorrect and biased inference is made.

- Example – Telemarketers.

TYPES OF BIAS

- **Bias** – certain outcomes are favored over other outcomes in samples.
- 2 Common Types of Bias:
  1. Selection Bias
  2. Sampling Bias

TYPES OF BIAS

- **Bias** – certain outcomes are favored over other outcomes in samples.
- 2 Common Types of Bias:
  1. Selection Bias
  2. Sampling Bias
     a) Convenience sampling
     b) Voluntary sampling

CONVENIENCE SAMPLING

- **Convenience sampling** – technique that selects subjects from population based on accessibility and ease.

- Problem:
  - Just because subjects are easy to talk with, doesn't mean they represent the population of interest as a whole.
  - Incorrect and biased inference is made.

- Example – Shopping store surveyors.

VOLUNTARY SAMPLING

- **Voluntary sampling** – technique where subjects volunteer themselves to sample.

- Problem:
  - People who volunteer don't necessarily represent the population of interest as a whole.
  - Incorrect and biased inference is made.

- Example – Marriage questionnaire.

SUMMARY

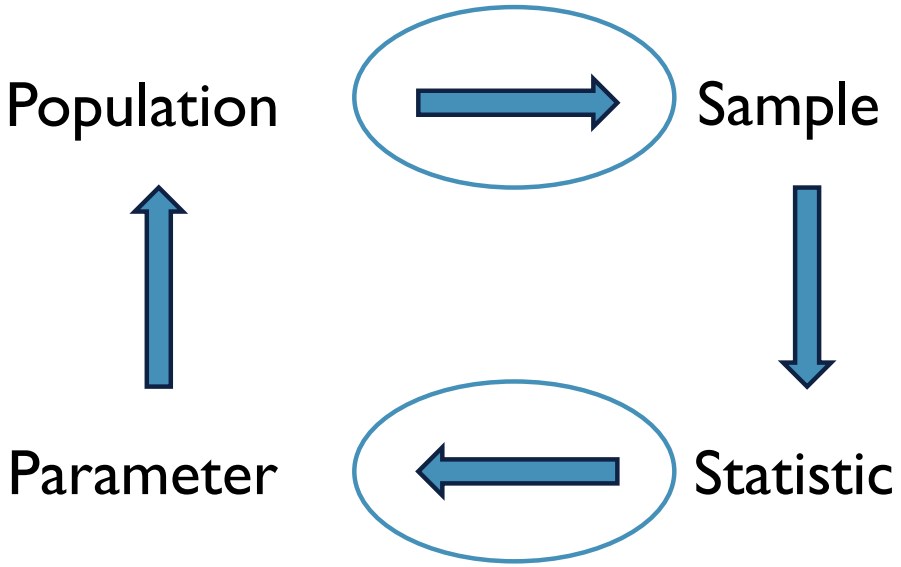
- Need good sampling to have good estimates.
- Bias – certain outcomes are favored over other outcomes in samples.
- 2 Common Types of Bias:
  1. Selection Bias
     a) Undercoverage
     b) Nonresponse
  2. Sampling Bias
     a) Convenience sampling
     b) Voluntary sampling

GOOD SAMPLING METHODS

GATHERING DATA

PARAMETERS VS. STATISTICS

Need good sampling to…



…have good estimates.

STATISTICAL TECHNIQUES

- **Statistical sampling techniques** use selection methods based on chance selection instead of convenience or judgement.
- 4 Common Techniques:
  1. Simple Random Sampling (SRS)
  2. Stratified Random Sampling
  3. Cluster Sampling
  4. Systematic Sampling

SIMPLE RANDOM SAMPLING (SRS)

- A method of sampling items from a population such that **every possible sample** of a specified size has an **equal chance** of being selected.

- Advantages:
  - No statistical bias, no previous information about sample needed ahead of time.
- Disadvantages:
  - Expensive, time consuming, hard to implement, need list of population.

STRATIFIED RANDOM SAMPLING (STS)

- A method of sampling items where the population is divided *beforehand* into subgroups, called **strata**, so that each member in the population belongs to only one strata. Sample items from **every** strata (with SRS for example).

- Advantages:
  - Smaller sample sizes can achieve same accuracy as SRS, more information about parts of population.
- Disadvantages:
  - Need information about population ahead of time to split on!

CLUSTER SAMPLING

- A method of sampling items where the population is divided *beforehand* into subgroups, called **clusters**, so that each member in the population belongs to only one cluster. Sample items from **a sample** of *m* clusters selected randomly.

- Advantages:
  - Overcome issues with travel, time, and expense; Easier to implement than SRS or STS.
- Disadvantages:
  - Need information about population ahead of time to split on – but not total list!
  - May have slight bias if random clusters aren't representative.

SYSTEMATIC SAMPLING

- A method of sampling items that involves selecting every k^{th} item in the population after randomly selecting a starting point between 1 and k.
- The value k is determined as the ratio of the population size over the desired sample size.

- Advantages:
  - Very easy to get sample.
- Disadvantages:
  - May be biased, especially if order of list of population matters.

EXAMPLE

- A large worldwide financial company wants to develop a new retirement plan for the company. They want to survey different managers of branches around the world to find out the most important strategies the new retirement plan should contain. They have 5000 branches worldwide and want to personally interview these branch managers. They have information about the branch size (small, medium, large), and the state/province location of the branch. They want to talk to 50 branch managers.

- Develop four separate strategies to sample these branch managers based on the four different statistical sampling techniques discussed previously.

EXAMPLE – SIMPLE RANDOM SAMPLE

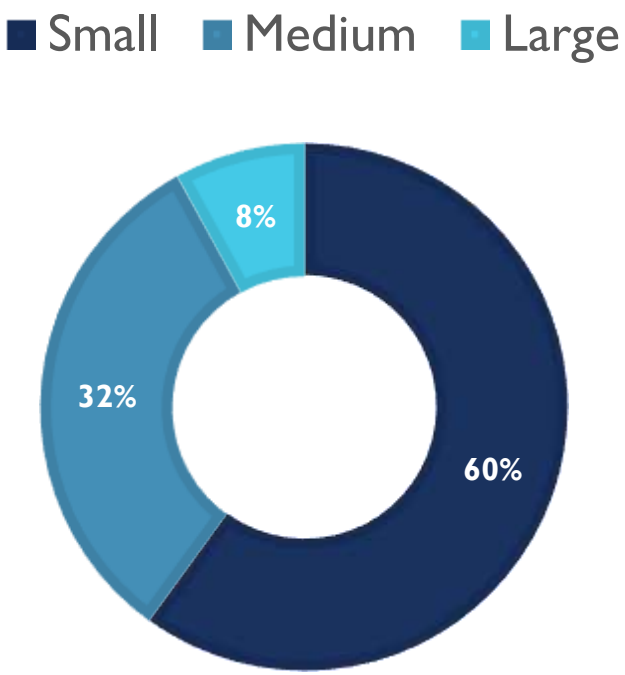
- Randomly sample 50 branches to interview their managers.
- Need a list of branches to randomly sample from.

Branch List:
1, 2, 3, …, 4998, 4999, 5000

50 Branches:
434, 938, 2582, …, 3218, 3439, 4134

EXAMPLE – STRATIFIED RANDOM SAMPLE

- Split branches into small, medium, and large since we want some of each represented.
- Randomly sample (SRS) proportionally from each group to make sure it looks like population.
  - $0.08 \times 50 = 4$
  - $0.32 \times 50 = 16$
  - $0.60 \times 50 = 30$



EXAMPLE – CLUSTER SAMPLING

- Split branches up by state.
- Randomly sample (SRS) 5 states.
- Randomly select (SRS) 10 branches in each state.

EXAMPLE – CLUSTER SAMPLING

- Split branches up by state.
- Randomly sample (SRS) 5 states.
- Randomly select (SRS) 10 branches in each state.

- Potential bias – what if these 5 states don't represent the population of all states well?

EXAMPLE – SYSTEMATIC SAMPLING

- Split list of branches into groups of 5000 / 50 = 100.
- Randomly select (SRS) starting point in first group of 100.
- Take same point in each group.

Branch List:
1, 2, 3, …, 4998, 4999, 5000

First Group List:
1, 2, 3, …, 98, 99, 100

50 Branches:
9, 109, 209, …, 4709, 4809, 4909

EXAMPLE – PUT ALL TOGETHER

- Develop four separate strategies to sample these branch managers based on the four different statistical sampling techniques discussed previously.
  1. SRS – Randomly sample 50 branches to interview their managers.
  2. STS – Stratify by size and select SRS from each.
  3. Cluster – Randomly select sample of states/provinces, then select branches at random from those states/provinces.
  4. Systematic – Select every 100^{th} branch in list of branches.

SUMMARY

- Need good sampling to have good estimates.
- 4 Common Techniques:
  1. Simple Random Sampling (SRS)
  2. Stratified Random Sampling (STS)
  3. Cluster Sampling
  4. Systematic Sampling

EXPERIMENTS

GATHERING DATA

TYPES OF STUDIES

- Data collection studies usually classified as observational or experimental.
- **Observational** – researcher does not interfere or intervene in the process of collecting data.
  - Requires selecting a sample.
- **Experimental** – researcher manipulates the conditions in which the study is carried out.
  - Requires selecting a sample and conducting and designing an experiment.

OBSERVATIONAL EXAMPLE

- Imagine you wanted to know the average height of the adult population in the United States because you are designing a new clothing line for adults.

- Observational study – just observing what has happened (height) in our population of interest.

EXPERIMENT TERMINOLOGY

- In an experiment, the researcher randomly assigns treatments to experimental units.
  - **Factor** – variable used to predict that takes on a finite number of values (categorical variable)
  - **Level** – setting a factor can take on.
  - **Treatment** – specific experimental condition, either the level of a factor (if only 1 factor) or the combinations of the levels from several factors.

EXPERIMENT EXAMPLE

- A mechanical engineer wanted to determine which variables influence gas mileage of a certain year and model of a car.
- Gas mileage is the variable we are interested in.
- Factors studied:
  - Tire pressure (low, standard).
  - Octane rating of fuel (regular, midgrade, premium).
- Held constant the following variables:
  - Weather conditions.
  - Route.
  - Tire type.

EXPERIMENTS

- The key thing that makes this study an **experimental study** is the active role the research plays in manipulating the environment.
- Makes it difficult in some situations to have a true experiment.
  - Effects of smoking on children?
  - Effects of family unit income as child for college performance?

DESIGN OF EXPERIMENTS

- Three key components to a well-designed experiment

1. Randomization – treatments are randomly assigned to experimental units
2. **Replication** – multiple subjects are assigned the same treatment
   - Subjects with the same treatment are called **replicates.**
   - More replication allow us to have more confidence in our study conclusions

DESIGN OF EXPERIMENTS

- Three key components to a well-designed experiment

1. Randomization – treatments are randomly assigned to experimental units
2. Replication – multiple subjects are assigned the same treatment
3. **Control -** some study conditions are held constant in order to reduce variability.
   - Controlling certain variables (sometimes called nuisances) that can impact what we are interested in.
   - This makes it easier to see differences due to our treatments

SUMMARY

- Observational study – researcher does not interfere or intervene in the process of collecting data.
- Experimental study – researcher manipulates the conditions in which the study is carried out.
- Three key components to a well-designed experiment
  1. Randomization
  2. Replication
  3. Control

DATA ETHICS

GATHERING DATA

GATHERING DATA

- The gathering of data leads to questions around the ethical collection and use of that data.
- As Christians we are held to an even higher standard around ethical considerations.

AREAS OF ETHICAL CONCERNS

- In observational studies / experiments we must keep the interest of the subject we are collecting data from at the forefront.

- 1964 Helsinki Declaration of the World Medical Association:
  - "The interests of the subject must always prevail over the interests of science and society."

SAFEGUARDS

- Collection of data:
  - Institutional review boards
  - Informed consent
  - Confidentiality

INSTITUTIONAL REVIEW BOARDS

- People have to exist that have the best interest of the subjects of the data collection in mind.
- Medical studies require **institutional review boards** to evaluate every study before it is conducted so that subjects are not put into any harm.
- These are **not** required for a lot of business studies, but the people collecting the data SHOULD take the subject into account before any data collection is performed.

INFORMED CONSENT

- **Informed** – subject should be told what data is needed from them and what potential outcomes come from the data being given to the people collecting it.
  - Must ensure that ALL information is shared.
  - May be hard for those gathering the data since they believe in their work and its usefulness.
  - Must always consider the risks.

INFORMED CONSENT

- **Consent** – after being informed, subjects must agree to the collection of data (usually in writing).
  - Who can give consent?
  - What about children? Mentally ill subjects?
  - Some are afraid that consent is harder to come by if you reveal ALL possible bad outcomes, no matter how unlikely. Is this bad?

CONFIDENTIALITY

- Once data is collected, privacy is VERY IMPORTANT!
- **Confidentiality** – the subjects in the data have their identifying information masked.
- You can report overall statistics about data that is gathered, but not who belonged to a certain outcome (unless you are reporting results to others who own the data).

- Many stories of confidential data being leaked due to computer hacking.

ANONYMITY VS. CONFIDENTIALITY

- **Anonymity** – identifying information about the subjects is NEVER known in the data collection.

- Anonymity is more private than confidentiality!

WEBSITE TESTING EXAMPLE

- You want to know which website design will work better to get people to click on your products. You randomly show one of the two websites to people who visit your website to measure which design performs better.
- Any concerns around…
  - Institutional review?
  - Informed consent?
  - Confidentiality?

WEARABLE MEDICAL DEVICE EXAMPLE

- You wear a watch that tracks your heartrate and sends that information off to the company. That company uses the information to determine trends and characteristics of people at risk for heart disease.
- Any concerns around…
  - Institutional review?
  - Informed consent?
  - Confidentiality?

SUMMARY

- The gathering of data leads to questions around the ethical collection and use of that data.
- As Christians we are held to an even higher standard around ethical considerations.
- In observational studies / experiments we must keep the interest of the subject we are collecting data from at the forefront.
- Collection of data:
  - Institutional review boards
  - Informed consent
  - Confidentiality

COLLECTING DATA INTUITION

GATHERING DATA

GATHERING DATA

- Main concepts in this section of the course:
  - Samples and populations
  - Randomness
  - Good vs. bad sampling methods
  - Ethical concerns around data

INTUITION – POPULATION OF INTEREST

- Who are you REALLY interested in gathering data around?
- The biggest problem with setting a population is not providing enough detail.
- Be very detailed and it will same you time later on.

INTUITION – REPRESENTATIVE SAMPLE

- Does your sample represent your population?
- Good sampling methods that involve randomness help you get a sample that represents the population.
- Still good practice to explore your data to make sure it looks like the population in a commonsense way.
- Example:
  - Possible to randomly get REALLY lucky and select only NBA players for your height study.
  - However, upon investigation, you realize that your sample probably isn't right, so you take another sample.

INTUITION – GOOD SAMPLING

- Does your sampling favor certain outcomes over others?
- Its always good to think about your sampling method to make sure you haven't built in any bias.
- Make sure your sampling method have randomness to help protect you against bias.

INTUITION – ETHICAL CONSIDERATIONS

- Can anyone be harmed or burdened by the collection and use of your data?
- Think about the possible harm the collection of your data could have.
- You must be open and honest with people you are collecting data on.
- Remember, God holds us to a higher standard than the world, let's represent Him well!

INTUITION – OVERALL

- It is EXTREMELY hard to protect yourself and consider all these things by yourself.
- ASK FOR HELP!
- I always like to ask others who I know (especially if they have different perspectives and experiences than I do) to make sure I am not missing anything.

SUMMARY

- Intuition and careful thought can protect you a lot of times when it comes to data gathering.
- Use other people to help make sure you are considering all the things you need to.