Lecture 1: Introduction to Statistics

Sravani Vadlamani

QMB 3200

08/25/2020

Story Time

- Joy Milne and her husband Les
- From Perth, UK
- Made international news when she claimed she had amazing talent

Smelling Parkinson's Disease

- Joy Milne was sitting with her husband, Les. She noticed that he smelled funny. So Les showered but it didn't help. After a couple of weeks, they went to a doctor. Les had been diagnosed with Parkinson's disease. Joy told the doctor that she knew something was wrong because of the smell. The doctor was flabbergasted because no one had ever claimed that a symptom of Parkinson's is a smell. It was like Joy was claiming to have some sort of superpower.

Smelling Parkinson's Disease

- How do you feel about Joy's claim? Do you believe her – why or why not?

Smelling Parkinson's Disease

- How do you feel about Joy's claim? Do you believe her – why or why not?

Needs evidence

Smelling Parkinson's Disease

- What could Parkinson's researchers do to verify Joy's claim that the disease has a smell?

Smelling Parkinson's Disease

- What could Parkinson's researchers do to verify Joy's claim that the disease has a smell?

Test

Smelling Parkinson's Disease

- If you were a Parkinson's disease researcher and you want to know if Joy's claim was true, how could you test her?

The Experiment

- Researchers had 12 people wear a shirt for a day. Some of them had Parkinson's while the others did not. The shirts were put in front of Joy. She was asked to sniff each shirt and then announce if that shirt came from a Parkinson's patient or from a person without Parkinson's.

The Experiment

- How many of Joy's guesses (out of 12) do you expect to be correct if she was just guessing at random?

The Experiment

- How many of Joy's guesses (out of 12) do you expect to be correct if she was just guessing at random?

50 – 60%

The Experiment

- How many correct guesses would convince you that Joy really does have the ability to smell Parkinson's?

The Experiment

- How many correct guesses would convince you that Joy really does have the ability to smell Parkinson's?

At least 10

The Experiment

- Suppose Joy did well during the experiment. What are two possible explanations for her performance?

The Experiment

- Suppose Joy did well during the experiment. What are two possible explanations for her performance?

Has the skill
Good Guesser

Why Statistics?

- Help us understand our world
  - By summarizing and reducing the amount and detail of data
- Conclusions are valid if and only if
  - Appropriate statistical techniques are implemented
  - Assumptions are reasonable
  - Random samples

Why Statistics?

- Multiple ways to analyze the same data
- Potential to provide contradicting information
- Statistics can be used deceitfully
  - There are three kinds of lies: lies, damned lies, and statistics” – Disraeli
  - “Round numbers are always false” – Samuel Johnson

Example

- Research undertaken by the University of Texas Health Science Center appears to show a link between consumption of *diet* soda and weight gain.

Example

- The study of more than 600 normal-weight people found, eight years later, that they were 65 percent more likely to be overweight if they drank one diet soda a day than if they drank none.

- And if they drank two or more diet sodas a day, they were even more likely to become obese. To the astonishment of many, it seems that those who drank diet soda had a greater chance of becoming overweight than did those who drank regular, full-calorie soda.

Example (Cont…)

- “By no means does this state that diet soda causes obesity - but there is a strange pattern at play here. Diet Soda has *zero* calories so what gives?”

- “One possibility: A person who drinks a diet soda may feel it's OK to make up for those calories with another high-calorie food.”

Statistics affects your lives

- Car insurance rates
  - Gender
  - Good student discounts
  - Age brackets
  - Marital status
  - Type of car
  - Color of car (?)
- College Admissions
  - SAT scores, High School GPA
- NCAA eligibility
- Minimum GPA to be a statistics major

Statistics for inference

- What classes to take, what professors to choose
  - Ratemyprofessor.com
- Weather
  - 30% chance of rain
- What university to attend?
  - Rankings, ratings, etc.

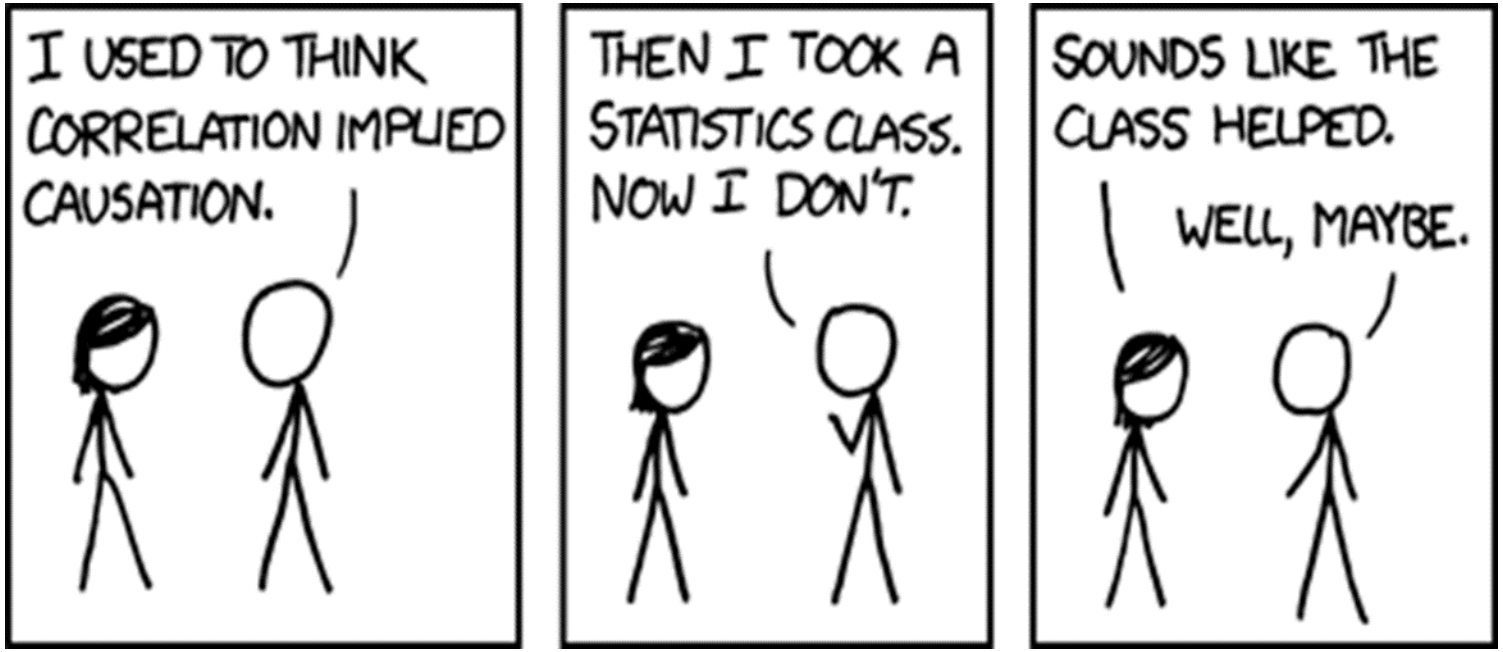
Statistics for inference

- Why go to college?
  - College graduates made an average of $51,554 in 2004, compared with $28,645 for adults with a high school diploma.
  - High school dropouts earned an average of $19,169, and those with advanced college degrees made an average of $78,093.

To go to college or not?

- But anybody who has gotten a passing grade in statistics knows what's wrong with this line of argument. A correlation between B.A.s and income is not proof of cause and effect. It may reflect nothing more than the fact that the economy rewards smart people and smart people are likely to go to college.

Correlation vs. Causation



In Summary

- Goal of statistical procedure
  - Simplify things
  - Information is lost but we assume it is unimportant
- Statistics involve DATA
  - We should be able to describe our data
  - We need to be able to make inferences from our data

Branches of Statistics

- Descriptive Statistics
  - Summarizing data from a research study
  - Example: Participants ranged from 25 to 35 years of age with a mean of 28 years and standard deviation of 1.5
- Inferential Statistics
  - Drawing conclusions from a study based on the data collected
  - Example: Students at FLPoly have greater working memory capacity than students at USF

Data

Data Generation

https://tinyurl.com/qmb3200data

Measurement

- Assigning quantitative values to some attribute of an object (or person) relative to some standard.

- *“ If a thing exists, it exists in some amount; and if it exists in some amount, it can be measured.”*
  - E. L. Thorndike (1914)

Measurement

- Height (cm, inches etc.)
- Weight (lbs., kgs)
- Intelligence, anxiety etc
  - Likert scale

Variables

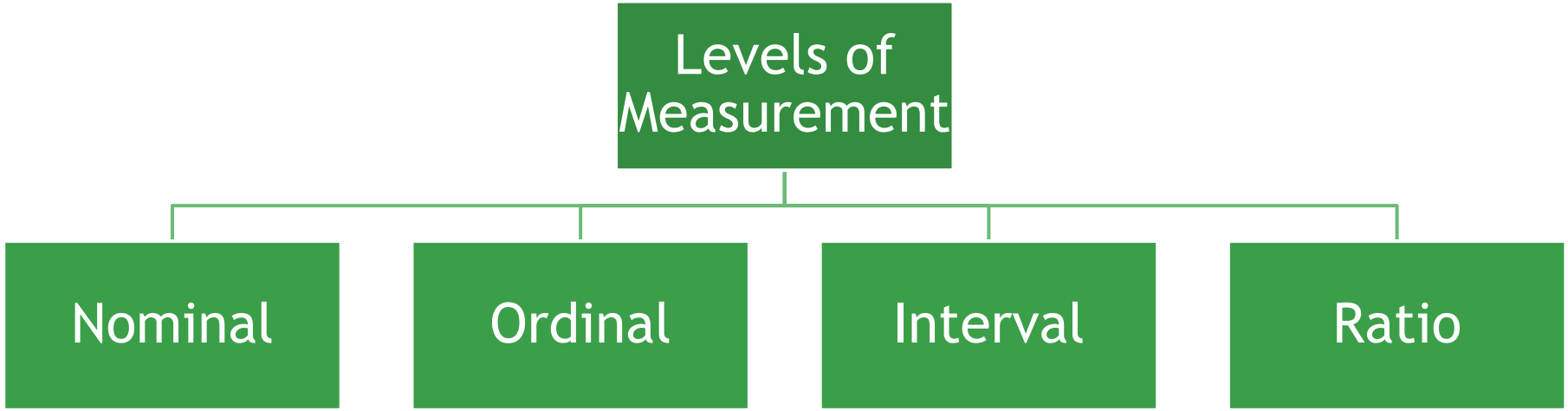
- A **variable** is an entity that can take on different *values*
  - Physical variables
    - Height, weight, sex
  - Psychological variables
    - Anxiety, depression, attachment, cognitive ability

- A **constant** takes on only one value

$$\hat{Y}_i = a + bX_i$$

a & b are constants

Measurement



Levels of Measurement

- Nominal
  - Categorical
- Ordinal
  - Ordering is important
  - Differences are not meaningful
- Interval
  - Ordering + Equal Interval
  - Differences are meaningful
- Ratio
  - Ordering, equal interval, + absolute zero
  - Division is meaningful

Nominal Scale

- Name Only
- No quantitative information
- 2 or more categories
  - Examples
    - Political party, Religion, Gender

Ordinal Scales

- Rank Ordering
  - Hierarchy of values
  - Unequal measurement intervals
  - Examples
    - Rankings of the best colleges
    - Hardness scale

Interval Scales

- Equal intervals between sequential values
  - No true 0
  - Degree Fahrenheit

- Linear transformation does not ruin the scale properties

$$\hat{Y}_i = a + bX_i$$

$$F = 32 + 9/5*C$$

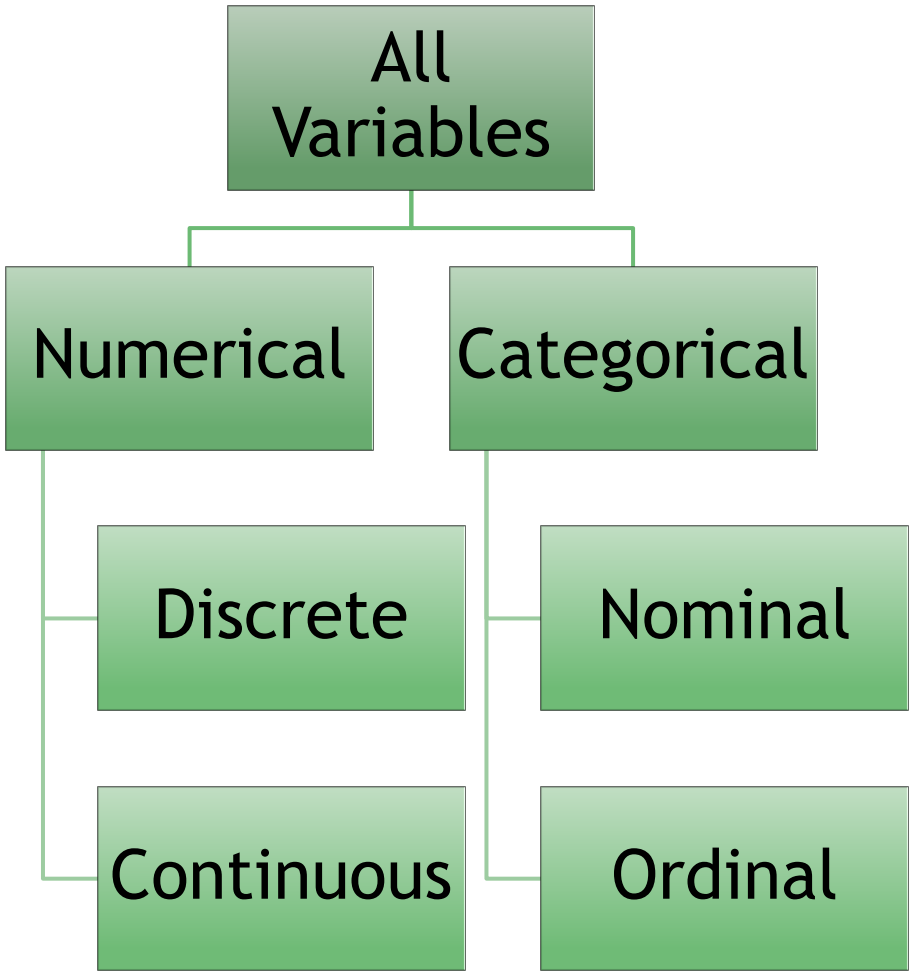
Ratio Scale

- Properties of the Interval Scale + absolute 0
  - Examples
    - Height, Weight
- Ratios make sense
  - 80 degrees F is not twice as hot as 40 degrees F
    - Interval Scale
    - 0 degrees F is not the absence of heat
  - 80 inches is twice as tall as 40 inches
    - Ratio scale: True zero point
    - 0 inches is the absence of height

Summary of Measurement Scales

- Measurement scales differ by how many of these attributes they have:
  - Magnitude
  - Equal intervals between adjacent units
  - Absolute zero point
- Nominal: none
- Ordinal: magnitude
- Interval: magnitude + equal intervals
- Ratio: magnitude + equal intervals + true zero

Types of Variables



Types of Variables

- Numeric Variables (vs. Nominal/Categorical Variable)
  - Quantitative Variables
  - Values are numbers with meaning (ordering is informative)
- Discrete Variable (vs. Continuous Variable)
  - Limited number of specific values
  - E.g., Number of times you went to the dentist
  - E.g., Wage categories such as $10,000 - $19,999, $20,000 - $29,999
- Continuous Variable (vs. Discrete Variable)
  - Theoretically infinite number of possible values
  - E.g., Wages in $$, Age, Height, Weight

Types of Variables (Cont.)

- Independent Variable (IV)
  - Assigned by the experimenter
  - Experimenter ***assigns*** values for each person
  - Usually categorical (e.g., control vs. experimental group)

- Dependent Variable (DV)
  - Measured by the experimenter
  - Experimenter ***measures*** attribute for each person
  - Output/Outcome Variable
  - Can be continuous or categorical (depression score on the CES-D or Depression diagnosis)

Back to Data Generated

- Look at the data generated from the survey and answer the following questions
  - How many cases were included in the data?
  - How many numerical variables were included in the data? Indicate what they are and if they are continuous or discrete?
  - How many categorical variables are included in the data and what are they? List the corresponding levels or categories?)

Sampling

- When we want to learn something about human behavior we don’t try to collect data from every single human being alive

- Instead we collect data from a small subsets of “humanity” (sample) and make generalizations to all of “humanity” (population)

Population

- Entire collection of events of interest
- Entire group of individuals that we want to learn something about
  - The group to which we *generalize to*
  - Never is completely observed

- All research questions are about populations

Population value called a **parameter**

- Use Greek letters (Mean = μ, Standard Deviation = σ)
- In statistical applications, parameters are estimated from a sample

Population

- Key idea of a population- includes **all** members of the category
- Can refer to any entity: rats, factories, schools, etc.
  - Doesn't have to be people
- We rarely (i.e., never) conduct research using whole populations
  - Not feasible
    - Too expensive to measure everybody
    - Can't get everybody to respond
  - One possible exception
    - U.S. census
    - Even census cannot measure everyone
      - Illegal immigrants
      - Homeless

Examples

- How does vocabulary knowledge change as an adult ages?
  - Population: Human population

- People who switched to Geico (or any other insurance company) saved an average of $350
  - Population: People who switched to Geico
  - Population is not people who inquired about switching to Geico

Examples

- Can juvenile delinquency be predicted from characteristics of a child in preschool?
  - Talking about any preschool-age child in the human population
  - May only be interested in children in the U.S.
    - Better to say, “In the U.S. can juvenile delinquency be predicted from characteristics of a child in preschool?”
    - Population: U.S. children

Sample

- Any subset of the population
- Sample value called a **statistic**
  - Use Latin letters (Mean = *M*, Standard Deviation = *S*)

Samples in Research

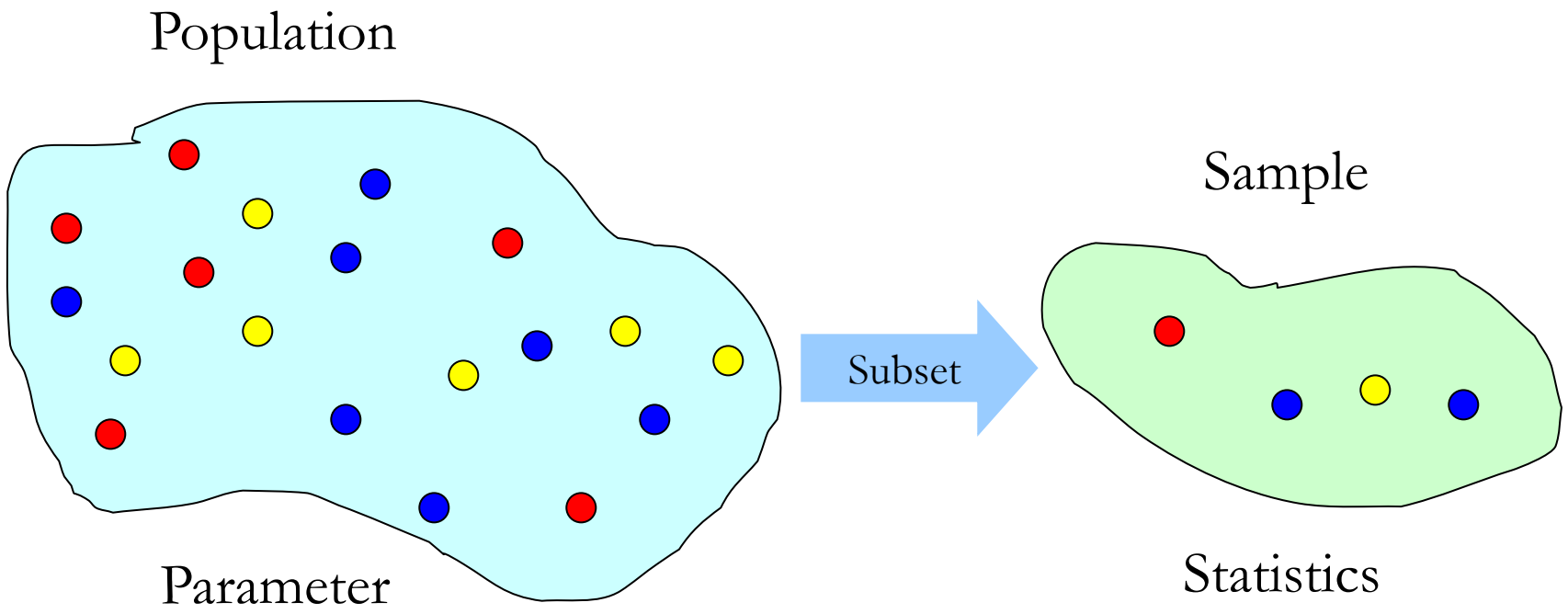
- Research studies use samples
- The group we generalize *from*
- Primary goal of the field of statistics is to learn about populations from samples
  - Take information from the sample we observe
  - Make generalizations to a population even though we don’t observe the entire population

Samples in Research

- What can we say about the population is based on the limited information about the population we get from the sample

- Example:
  - Children whose mothers graduated from high school have higher IQs than children with mothers who did not graduate from high school.

Sampling Strategies

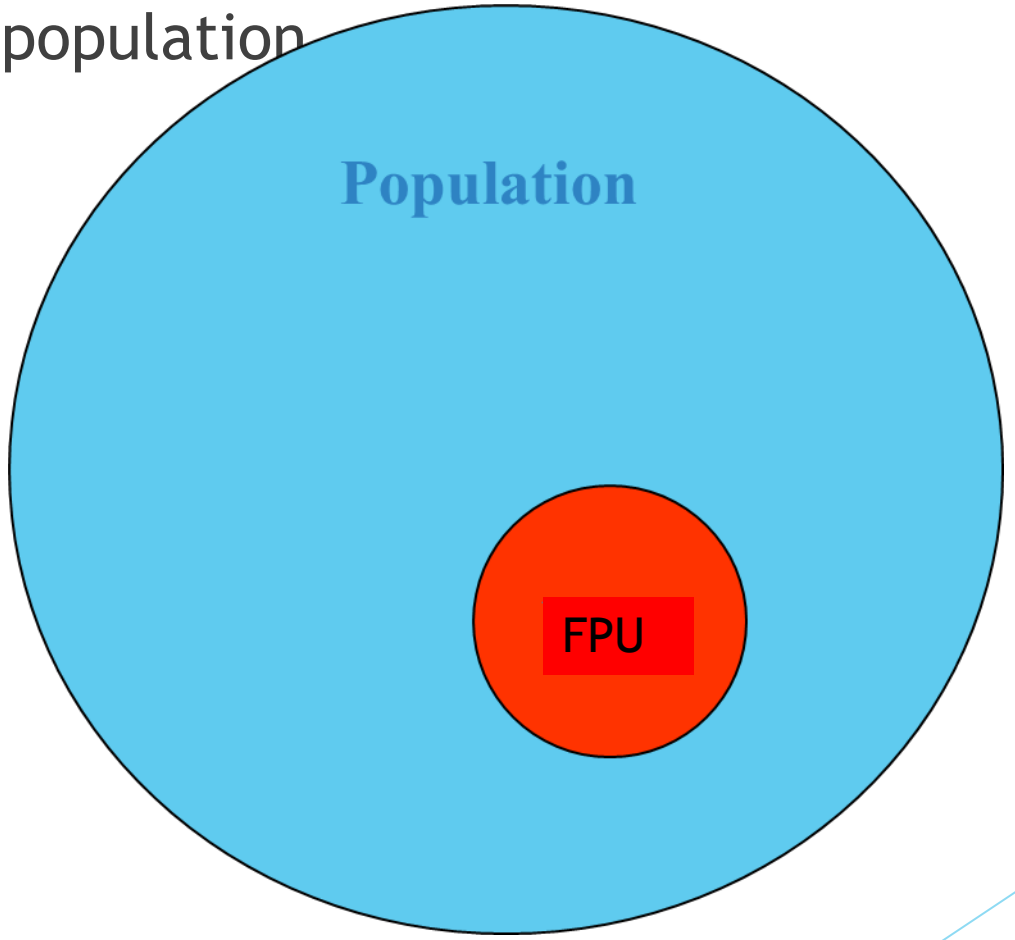


Sampling Strategies

- Convenient sampling
- Random sampling
- Stratified sampling
- Cluster sampling

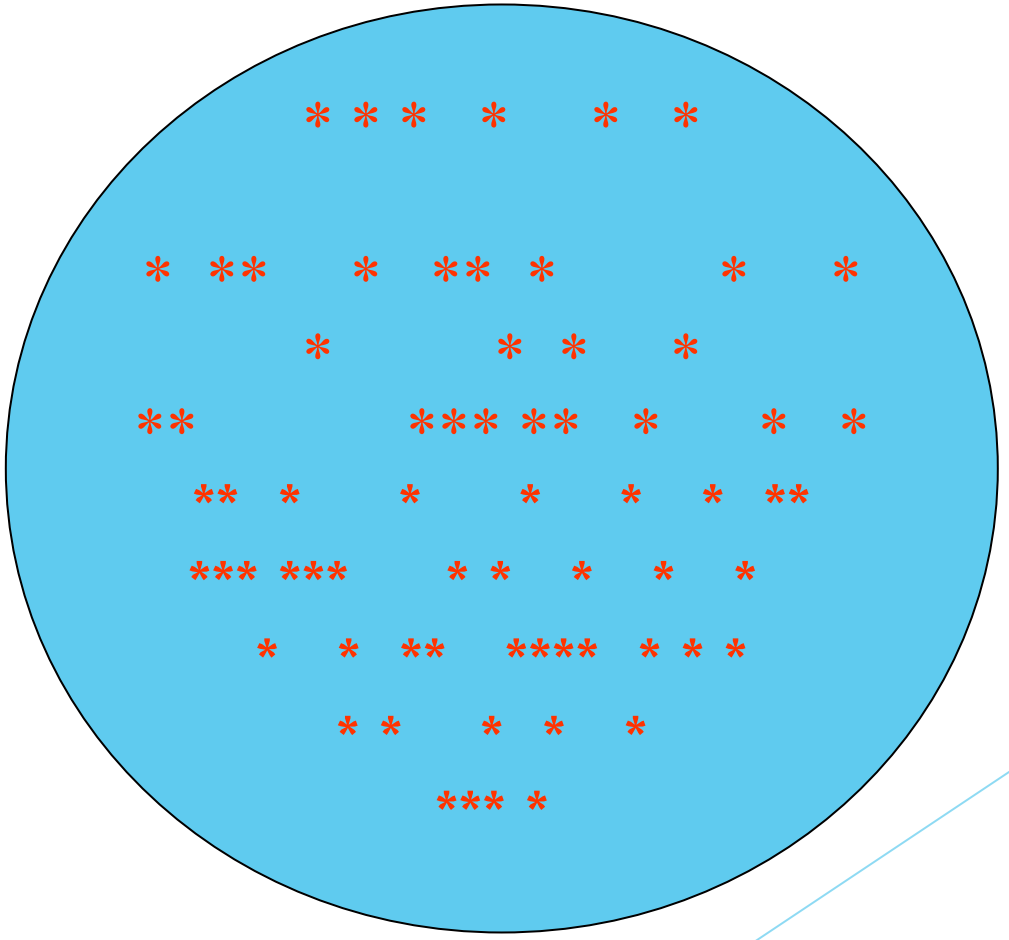
Convenient Sample

- Sample of participants drawn from a small (and often select) portion of the population
  - Non-random
  - Sampling bias
  - Non-response



Random Sample

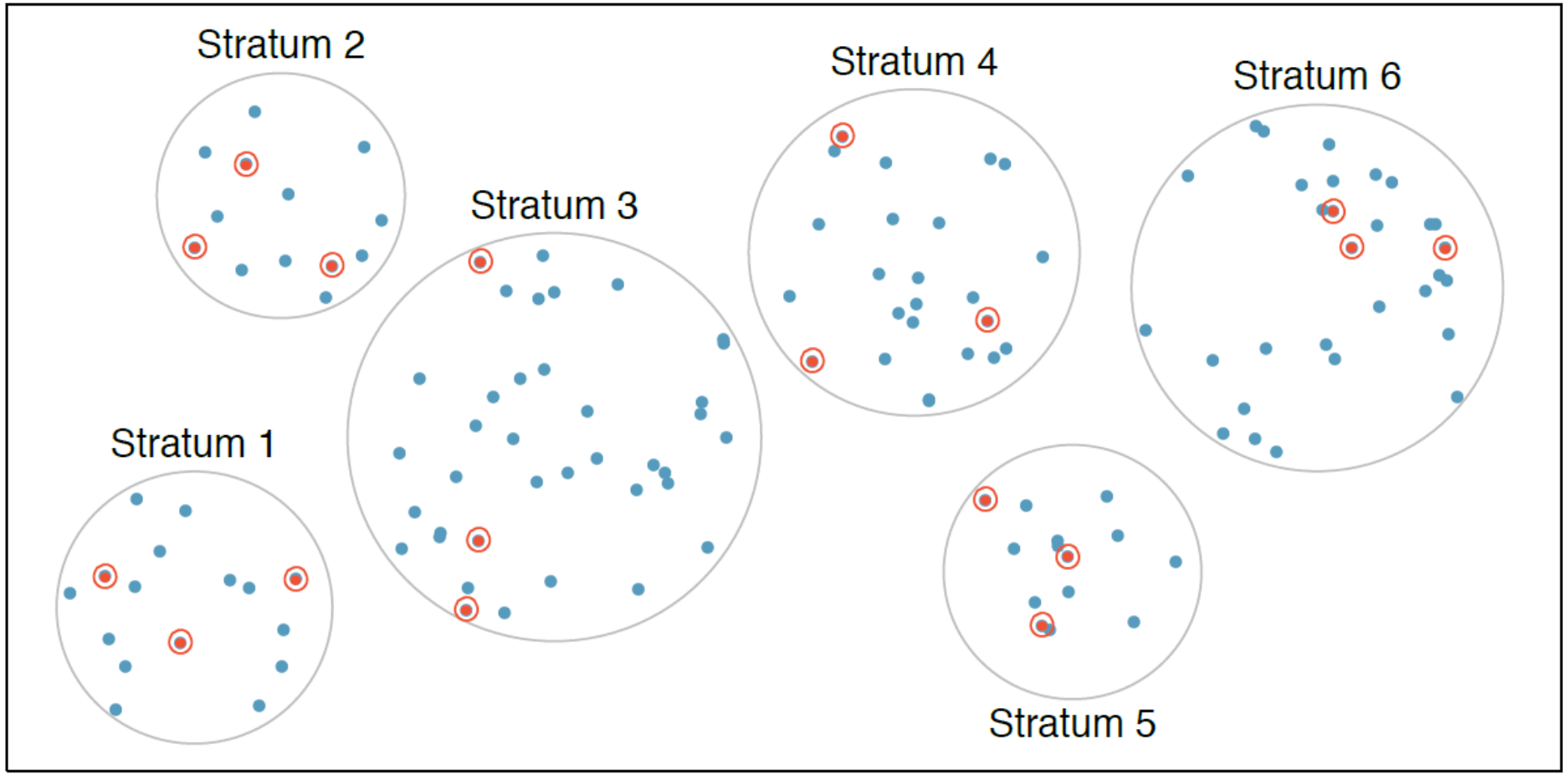
- Every person in the population has the same opportunity (probability) of being selected into the study



Stratified Sampling

- Divide population into groups called strata
- Choose strata such that similar cases are grouped together
- Use simple random sampling to choose a certain number of samples from each stratum
- Useful when cases in each stratum are similar

Stratified Sampling



Cluster Sampling

- Divide population into groups called clusters
- Sample fixed number of clusters
- Include all observations from the cluster

Random Assignment

- Each participant is independently and randomly assigned to one (and only one) group
- Each participant has the same probability of being assigned to each group
- Most often the groups are made to have the same number of participants
  - A violation of true random assignment

Validity

- Random sampling is important to establish **external** validity
  - Results will generalize to the population of interest

- Random assignment is important to establish **internal** validity
  - Results will replicate

Research Studies

- Observational Studies
  - collect data in a way that does not directly interfere with how the data arise
  - Collect information through surveys, review company or medical records
  - Provide evidence of naturally occurring association between variables but cannot show a causal connection
- Experimental Study
  - Randomly assign subjects to treatments to establish causal connections between variables

Example

- In an experiment or observation, the independent variable typically has an effect on the dependent variable.
- **Recall example of diet soda and weight gain, diet soda consumption is your independent variable and weight gain is your dependent variable.**
  - Is this study observational or experimental?
  - What conclusions might be drawn?
  - What might be another variable(s) that impacts the weight gain?

Confounding Variables

- A confounding variable is an “extra” variable that you didn’t account for.
- Issues
  - Results that does not make sense or are useless.
  - They can ruin an experiment and give you useless results.
  - Suggest correlation when in fact there isn’t.
  - Introduce bias.
- Important to know what one is, and how to avoid getting them into your experiment in the first place.

Confounding Variables - Example

- Suppose an observational study tracked sunscreen use and skin cancer and it was found the more sunscreen someone used, the more likely a person was to have skin cancer. Does this mean sunscreen causes skin cancer?

Experiments

- Randomized experiments are important when trying to prove a causal connection between two variables.
  - Many factors can lead to the wrong conclusion.

A good experimental design will include:

- Controlling the factors that affect the outcome.
- Randomly assign subjects to treatment groups (treatment and control)
  - Evens out the conditions that can't be controlled.
  - Prevents accidental bias.
- Replication: collect a sufficiently large sample. The more cases the greater accuracy the effect of the independent variable on the dependent variable.
- Other controls: Blocking (stratification of groups), Blind and Double-blind

Experiments

- Difference between blocking and explanatory variables:
  - Factors are conditions we can impose on the experimental units.
  - Blocking variables are characteristics that the experimental units come with, that we would like to control for.
  - Blocking is like stratifying, except used in experimental settings when randomly assigning, as opposed to when sampling.

Practice

- A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are equally represented in each group. Which of the below is correct?}
- (a) There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)
- (b) There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)
- (c) There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)
- (d) There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

Practice

- A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are equally represented in each group. Which of the below is correct?}
- (a) There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)
- *(b) There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)*
- (c) There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)
- (d) There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

More Terminology

- Placebo: fake treatment, often used as the control group for medical studies
- Placebo effect: experimental units showing improvement simply because they believe they are receiving a special treatment
- Blinding: when experimental units do not know whether they are in the control or treatment group
- Double-blind: when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group

Practice

What is the main difference between observational studies and experiments?

(a) Experiments take place in a lab while observational studies do not need to.

(b) In an observational study we only look at what happened in the past.

(c) Most experiments use random assignment while observational studies do not.

(d) Observational studies are completely useless since no causal inference can be made based on their findings.

Practice

What is the main difference between observational studies and experiments?

(a) Experiments take place in a lab while observational studies do not need to.

(b) In an observational study we only look at what happened in the past.

(c) *Most experiments use random assignment while observational studies do not.*

(d) Observational studies are completely useless since no causal inference can be made based on their findings.

Practice

A statistics student is curious about the relationship between the amount of time students spend on social networking sites and their performance at school. He decides to conduct a survey and examines the following research strategies. For each, name the sampling method proposed and any bias you might expect.

1. He randomly samples 40 students from the study's population, gives them the survey, asks them to fill it out and bring it back the next day.
2. He gives out survey only to his friends, making sure each one of them fills out the survey
3. He posts a link to an online survey on FB and asks his friends to fill the survey
4. He randomly samples 5 classes and asks a random sample of students from those classes to fill out the survey

Practice – answers marked in red

A statistics student is curious about the relationship between the amount of time students spend on social networking sites and their performance at school. He decides to conduct a survey and examines the following research strategies. For each, name the sampling method proposed and any bias you might expect.

1. He randomly samples 40 students from the study's population, gives them the survey, asks them to fill it out and bring it back the next day. Simple random- non response
2. He gives out survey only to his friends, making sure each one of them fills out the survey. Convenience, not representative of population and non-response
3. He posts a link to an online survey on FB and asks his friends to fill the survey. Convenience, not representative of population and non-response
4. He randomly samples 5 classes and asks a random sample of students from those classes to fill out the survey. Multistage, no bias if classes are similar except for potential non-response

Ethics in Data Collection

- Ethics are the values by which human behavior is morally evaluated

- Research must be done in an ethically responsible manner

Ethics in Data Collection

- Tuskegee Syphilis Study
  - Examined the progression of syphilis in a group of 400 poor African-Americans
  - Started in 1932, continued until 1972
  - Effective treatment (penicillin, discovered in 1943) was withheld in the name of science
  - Subjects were not informed of the nature of the research nor that there was treatment available elsewhere
  - Is this ethically justifiable?

Ethics in Data Collection

- Milgram Obedience Experiments
  - Subjects were told to give a dangerous and even fatal electric shock to another person when requested or ordered to do so by a person in authority
  - Published in 1963
  - Used deception
    - Person receiving shock was confederate
    - Shock was faked
  - Is this ethically justifiable

Research Guidelines

- Principles
  - Respect for persons and their autonomy
    - Informed consent
      - Prospective participants must know what they are getting into
      - Voluntarily agree
      - If unable to give consent (e.g., kids), must have appropriate representative
    - In rare cases where deception is necessary
      - Still must give informed consent but with incomplete disclosure
      - Complete disclosure after the completion of the research (debriefing)

Research Guidelines

- Principles
  - Beneficence and Non-maleficence
    - Research must have some conceivable benefit
      - Usually to scientific knowledge
      - Sometimes to participant
    - Research must avoid doing harm
      - Usually to the participant
    - Does not mean that risk of harm is never OK
      - Benefit must clearly outweigh potential harm
      - ex.: drug that has chance of curing terminal disease, but chance of killing patient

Research Guidelines

- Principles
  - Justice
    - Fairness
    - Benefits of scientific inquiry must be available to all
    - If drug is showing clear benefits, stop experiment and give to all participants
      - Even those in placebo group
    - In psychological research
      - No potential participant may be excluded from participation without cause

Research Guidelines

- Principles
  - Trust
    - Maintain confidentiality
    - Maintain anonymity
  - Fidelity and scientific integrity
    - Honesty in scientific research
    - Common violations
      - Misreporting results
      - Plagiarism

How to determine if a research is ethical

- Each institution has an Institutional Review Board (IRB)
  - Consists of scientists in various fields plus at least one community member
  - IRBs review every proposal for scientific research
    - Determine if ethical standards are met

- In the end, it is the responsibility of the researcher to maintain ethical behavior

Suggested Readings

- Chapter 1 from OIS