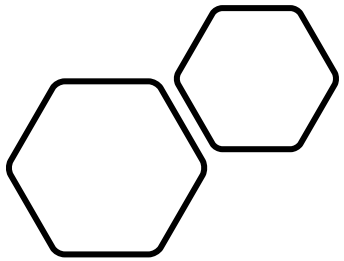
Data - Basics



Sravani Vadlamani

Statistics I

Data Generation

https://tinyurl.com/sta2023datagen

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

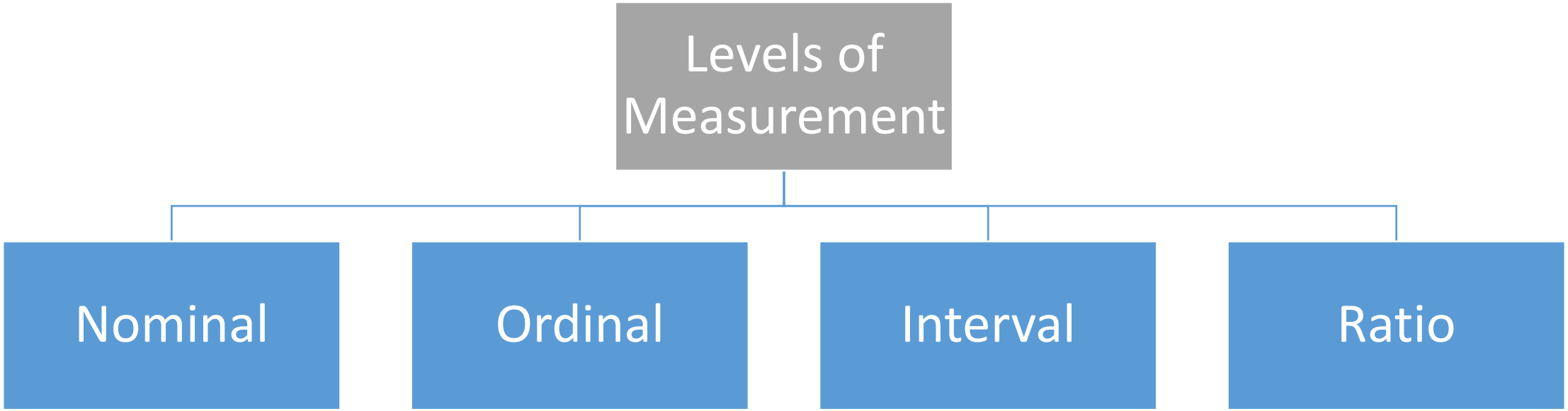
- Height (cm, inches etc.)
- Weight (lbs., kgs)
- Intelligence, anxiety etc
  - Likert scale (0-10, 1-5 etc)

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
  - $\hat{Y}i = a + bXi$, 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 Scale

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

Interval Scale

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

- Linear transformation does not ruin the scale properties

$$\hat{Y}i = a + bXi$$

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

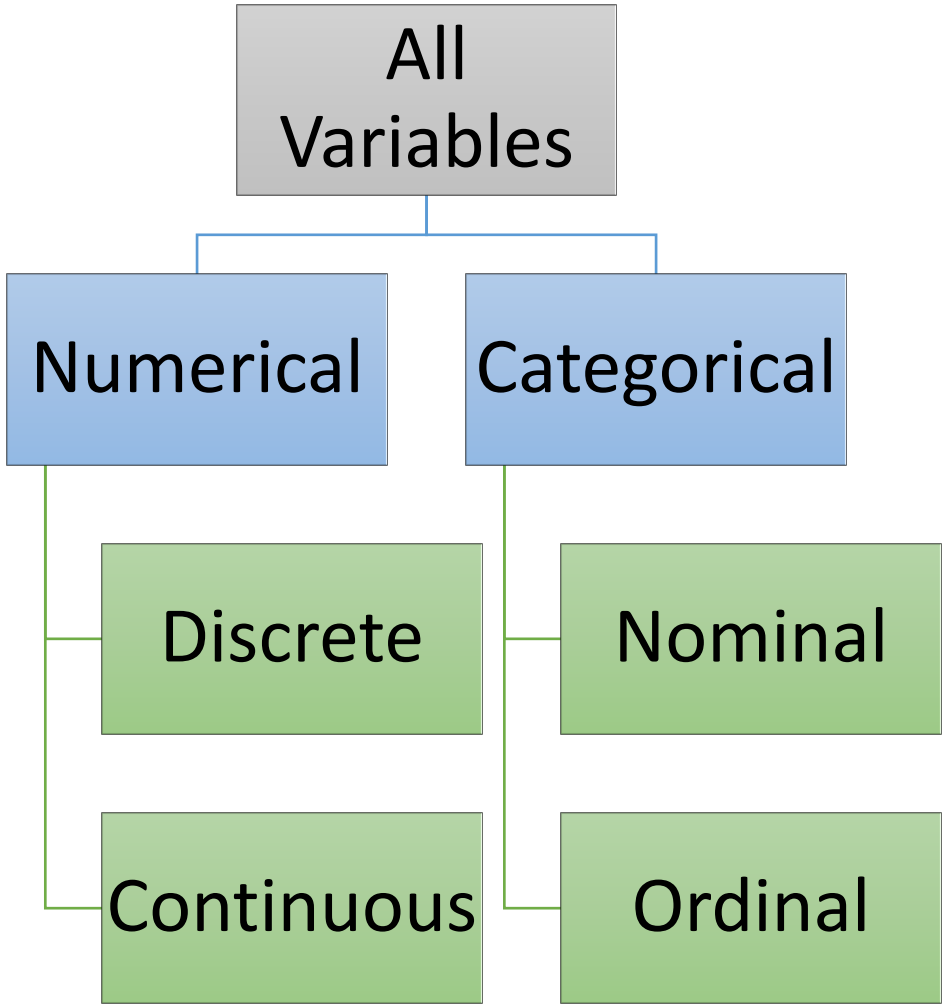
Ratio Scale

- 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

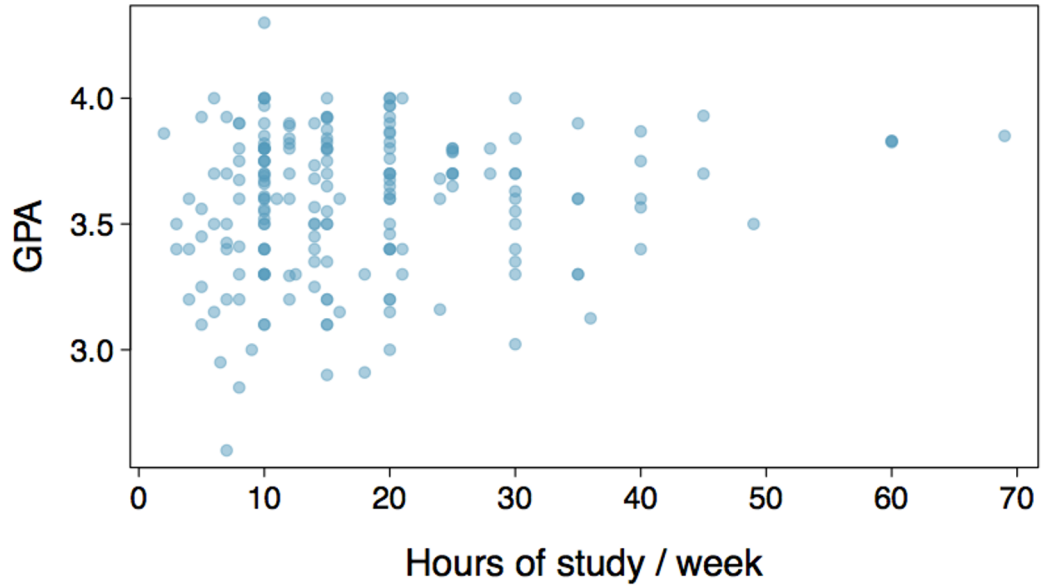
- Numeric Variables (vs. Nominal/Categorical Variable)
  - Quantitative Variables
  - Values are numbers with meaning (ordering is informative)
- Discrete Variable (vs. Continuous Variable)
  - Limited number of specific values
  - E.g., Number of times you went to the dentist
  - E.g., Wage categories 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

Independent vs. Dependent Variables

- When two variables show some connection with one another, they are called associated variables.
  - Associated variables can also be called dependent variables and vice-versa.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be independent.

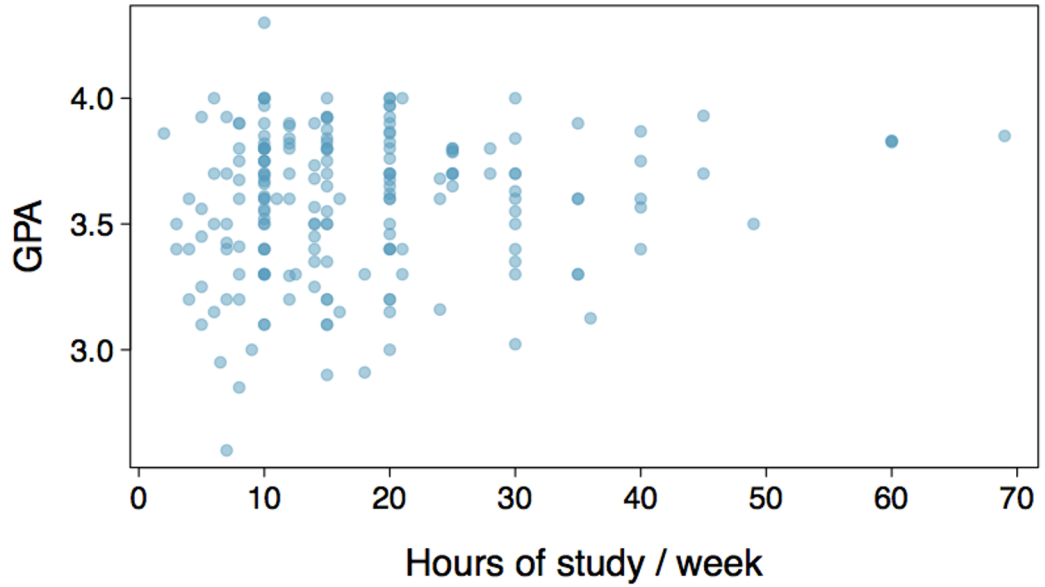
Relationship between variables

- Does there appear to be a relationship between the hours of study per week and the GPA of a student?



Relationship between variables

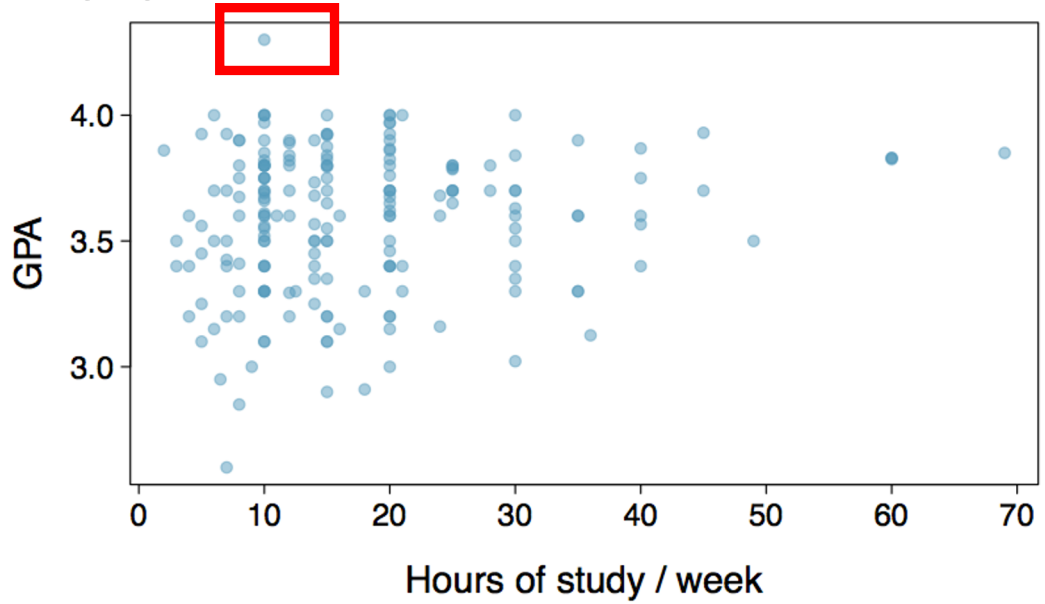
- Does there appear to be a relationship between the hours of study per week and the GPA of a student?



Can you spot anything unusual about any of the data points?

Relationship between variables

- Does there appear to be a relationship between the hours of study per week and the GPA of a student?

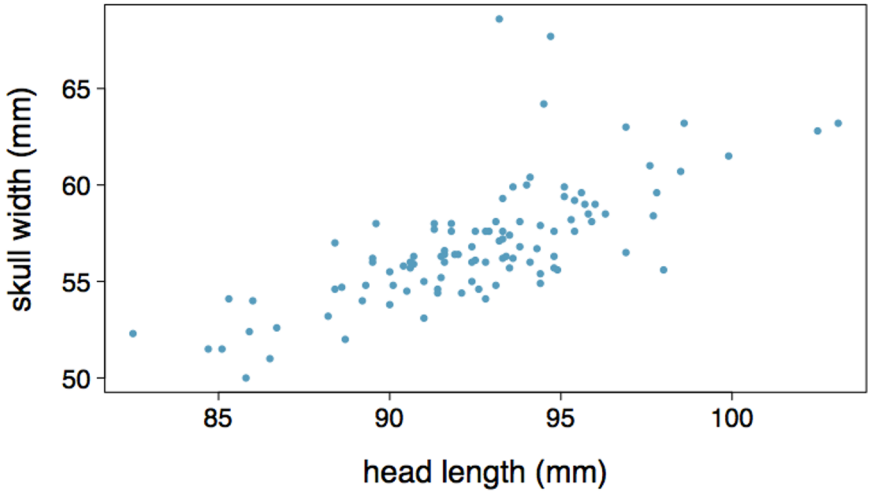


Can you spot anything unusual about any of the data points?

There is one student with *GPA* > 4.0, this is likely a data error.

Relationship between variables

Based on the scatterplot, which of the following statements is correct about the head and skull lengths of possums?



(a) There is no relationship between head length and skull width, i.e. the variables are independent.

(b) Head length and skull width are positively associated.

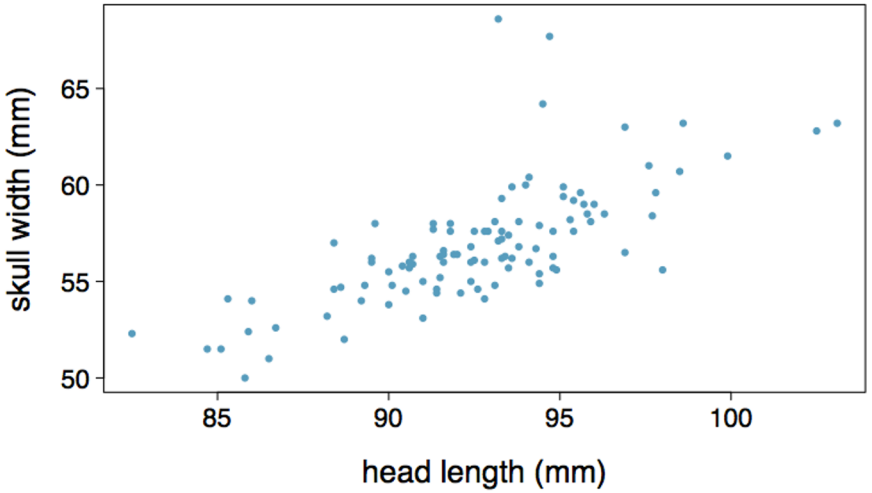
(c) Skull width and head length are negatively associated.

(d) A longer head causes the skull to be wider.

(e) A wider skull causes the head to be longer.

Relationship between variables

Based on the scatterplot, which of the following statements is correct about the head and skull lengths of possums?



(a) There is no relationship between head length and skull width, i.e. the variables are independent.

(b) Head length and skull width are positively associated.

(c) Skull width and head length are negatively associated.

(d) A longer head causes the skull to be wider.

(e) A wider skull causes the head to be longer.

Back to Survey Data

Classroom Survey

- A survey was conducted on students in an introductory statistics course. Below are a few of the questions on the survey, and the corresponding variables the data from the responses were stored in:
  - gender: What is your gender?
  - intro_extra: Are you an introvert or an extrovert?
  - sleep: How many hours do you sleep at night, on average?
  - bedtime: What time do you usually go to bed?
  - countries: How many countries have you visited?
  - dread: On a scale of 1-5, how much do you dread being here?

Data Matrix

variable ↓

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

← *observation*

Variable Types

variable ↓

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

← *observation*

- gender:

Variable Types

variable ↓

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

← *observation*

- gender: *categorical*

Variable Types

variable ↓

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

← *observation*

- gender: *categorical*
- sleep:

Variable Types

variable ↓

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

← *observation*

- gender: *categorical*
- sleep: *numerical, continuous*

Variable Types

variable ↓

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

← *observation*

- gender: *categorical*
- sleep: *numerical, continuous*
- bedtime:

Variable Types

variable ↓

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

← *observation*

- gender: *categorical*
- sleep: *numerical, continuous*
- bedtime: *categorical, ordinal*

Variable Types

variable ↓

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

← *observation*

- gender: *categorical*
- sleep: *numerical, continuous*
- bedtime: *categorical, ordinal*
- countries:

Variable Types

variable ↓

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

← *observation*

- gender: *categorical*
- sleep: *numerical, continuous*
- bedtime: *categorical, ordinal*
- countries: numerical, discrete

Variable Types

variable ↓

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

← *observation*

- gender: *categorical*
- sleep: *numerical, continuous*
- bedtime: *categorical, ordinal*
- countries: *numerical, discrete*
- dread:

Variable Types

variable ↓

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

← *observation*

- gender: *categorical*
- sleep: *numerical, continuous*
- bedtime: *categorical, ordinal*
- countries: *numerical, discrete*
- dread: *categorical, ordinal - could also be used as numerical*

Practice

What type of variable is a telephone area code?

(a) numerical, continuous
(b) numerical, discrete
(c) categorical
(d) categorical, ordinal

Practice

What type of variable is a telephone area code?

(a) numerical, continuous
(b) numerical, discrete
(c) *categorical*
(d) categorical, ordinal

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

Census

- It can be difficult to complete a census: there always seem to be some individuals who are hard to locate or hard to measure. And these difficult-to-find people may have certain characteristics that distinguish them from the rest of the population.
- Populations rarely stand still. Even if you could take a census, the population changes constantly, so it's never possible to get a perfect measure.
- Taking a census may be more complex than sampling.

Population - 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

Population - 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

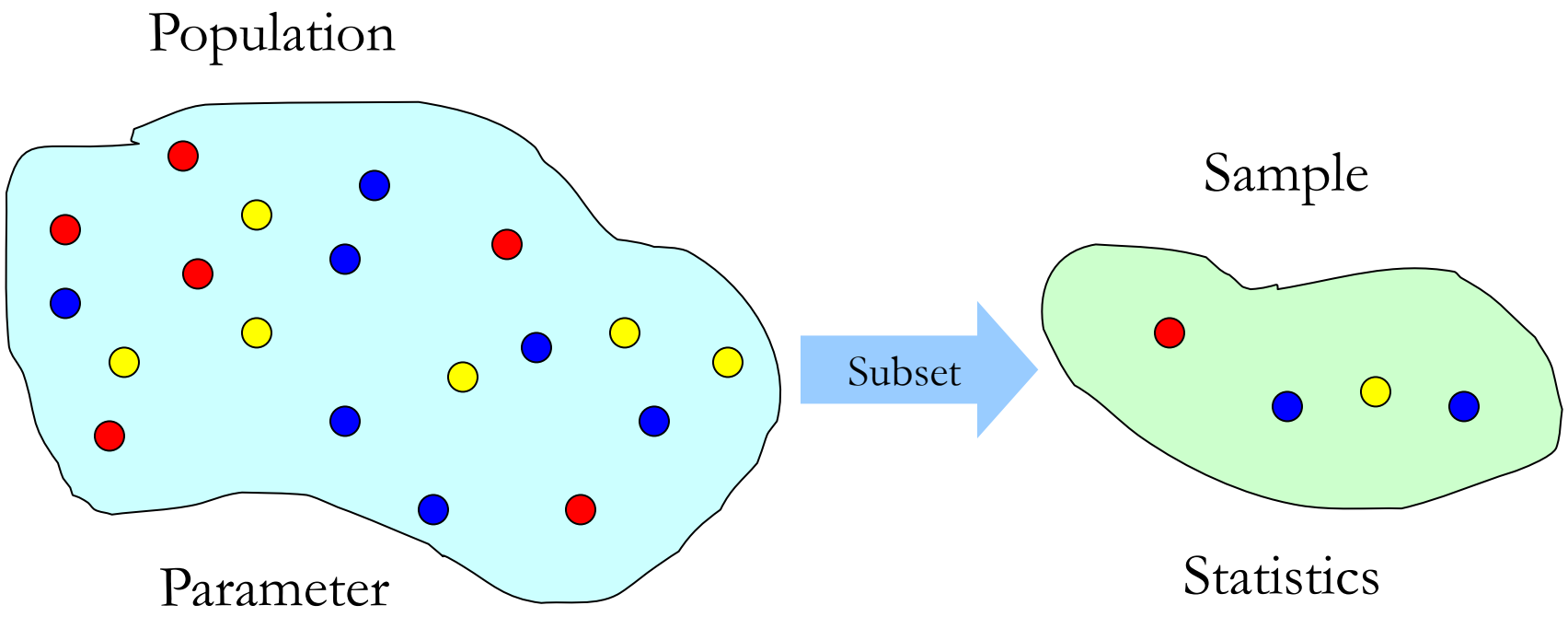
Sample

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

Sample

- Research studies use samples to generalize
- 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

Population vs Sample



Sample

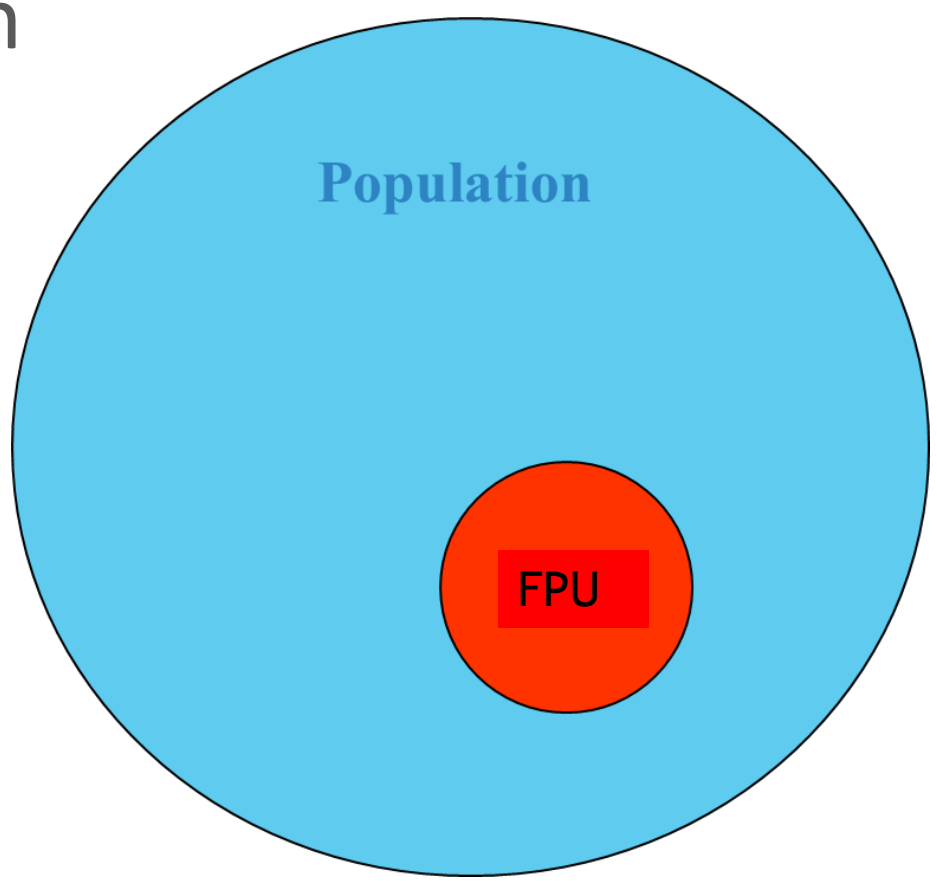
- Sampling is natural.
- Think about sampling something you are cooking - you taste (examine) a small part of what you're cooking to get an idea about the dish as a whole.
- When you taste a spoonful of soup and decide the spoonful you tasted isn't salty enough, that's *exploratory analysis*.
- If you generalize and conclude that your entire soup needs salt, that's an *inference*.
- For your inference to be valid, the spoonful you tasted (the sample) needs to be *representative* of the entire pot (the population).
  - If your spoonful comes only from the surface and the salt is collected at the bottom of the pot, what you tasted is probably not representative of the whole pot.
  - If you first stir the soup thoroughly before you taste, your spoonful will more likely be representative of the whole pot.

Sampling Strategies

- Convenient sample
- Random sample
- Stratified sample
- Cluster sample

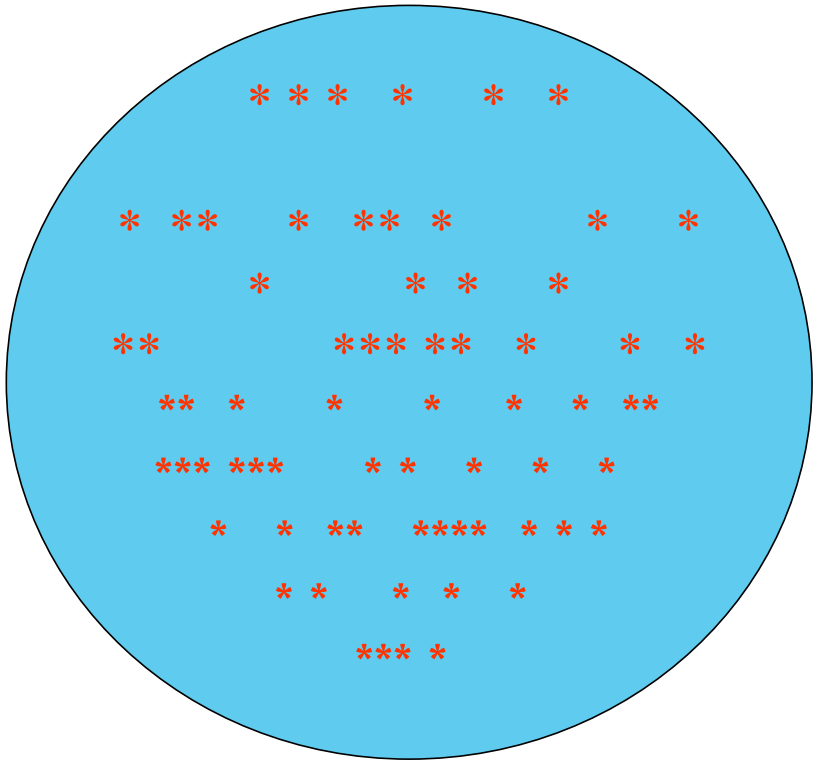
Convenient Sample

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



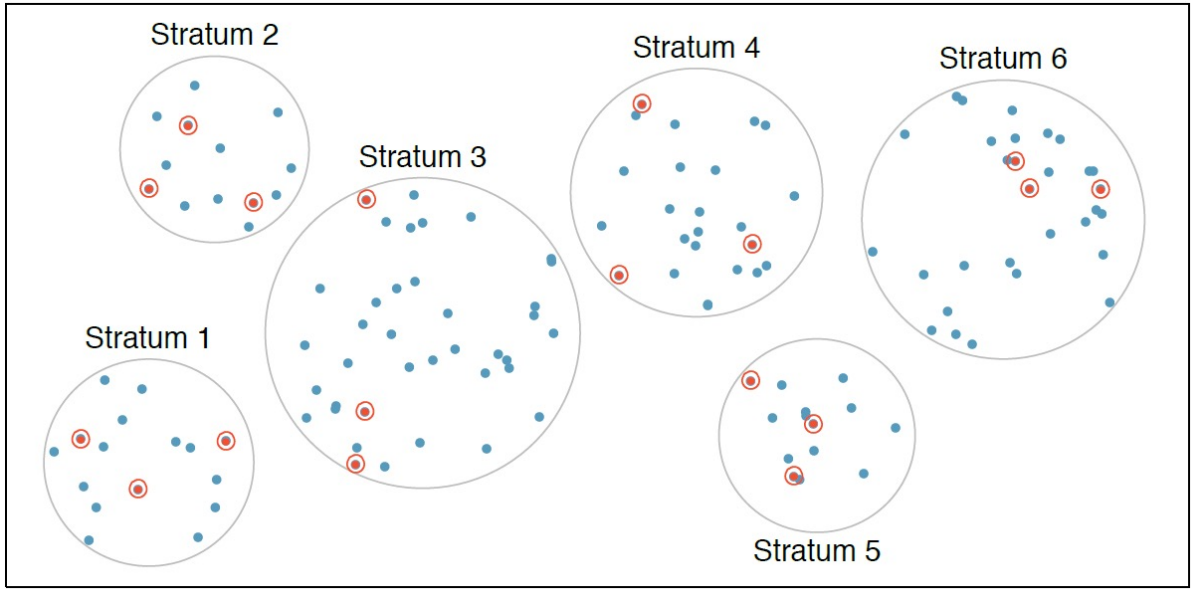
Random Sample

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



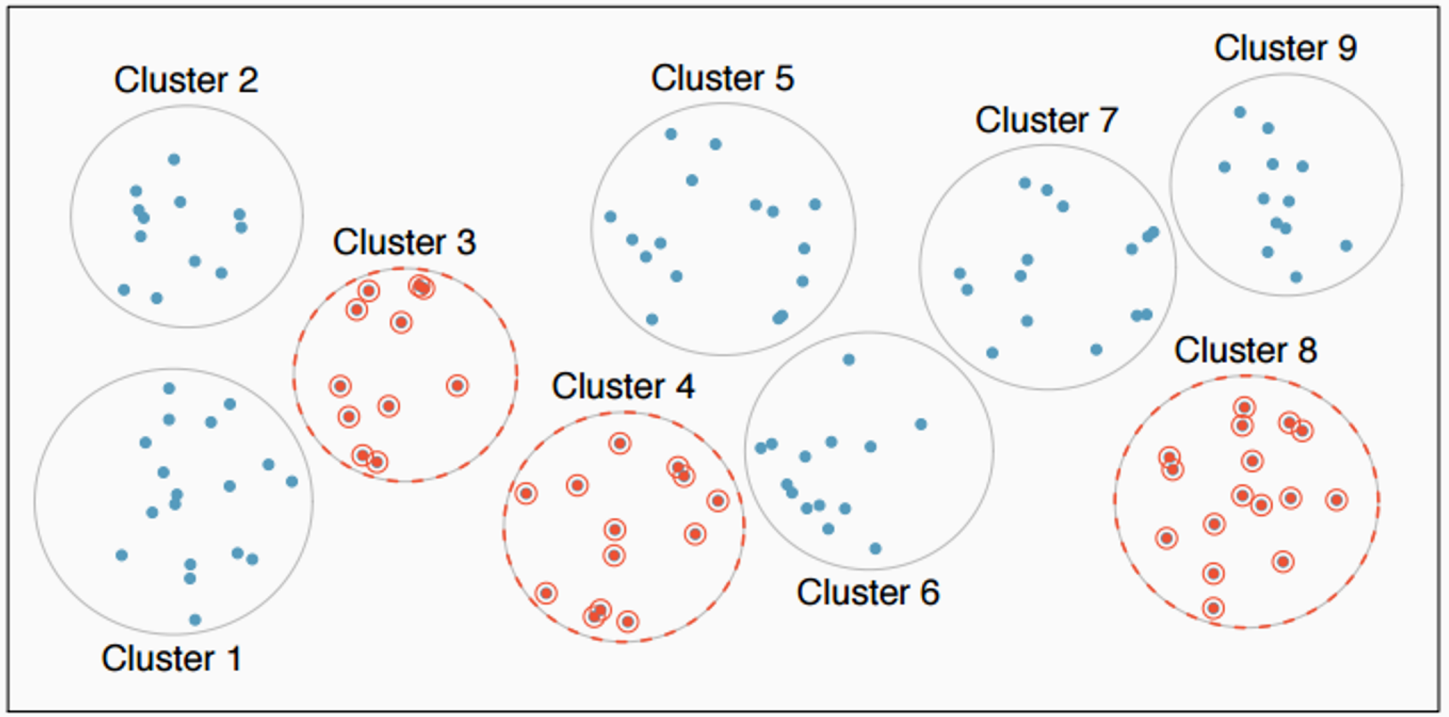
Stratified Sample

- 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



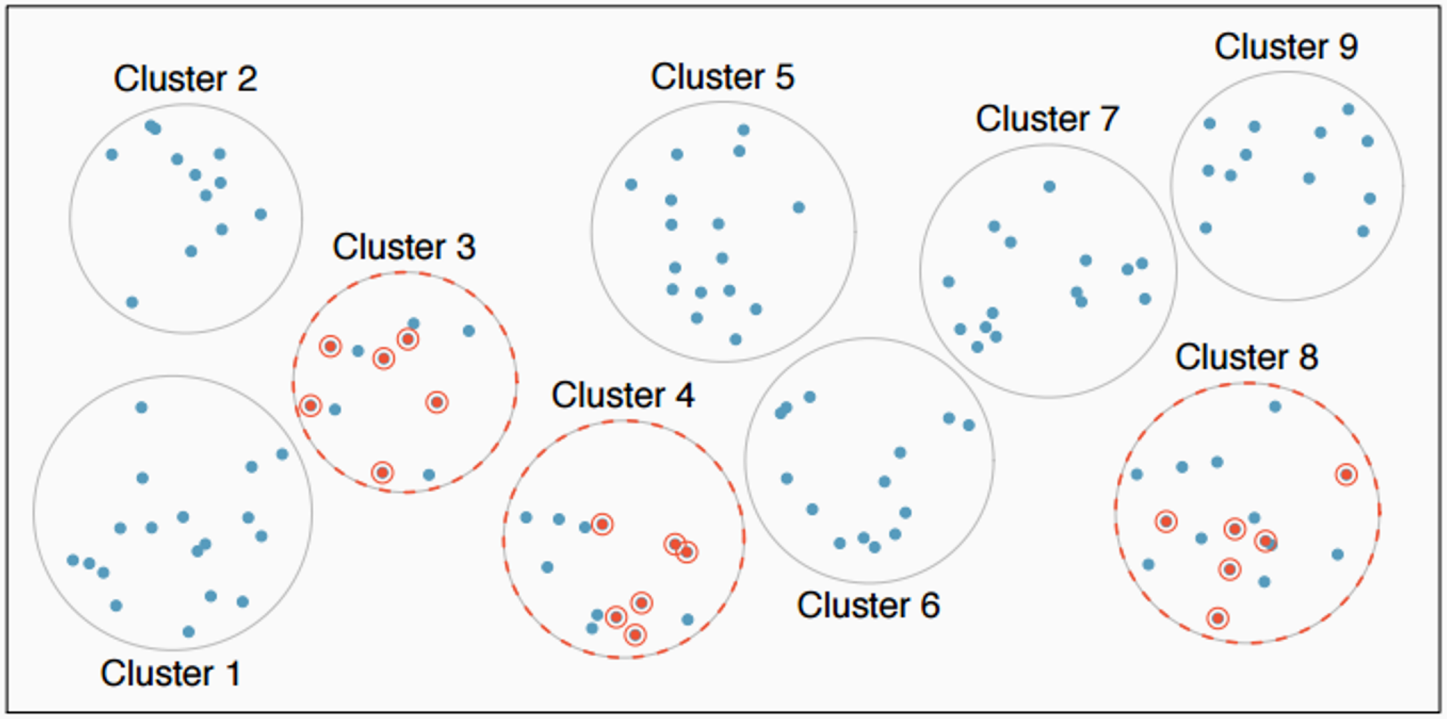
Clustered Sample

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



Multistage Sample

- Clusters are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then take a simple random sample of observations from the sampled clusters



Practice

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments. Which approach would likely be the least effective?

(a) Simple random sampling

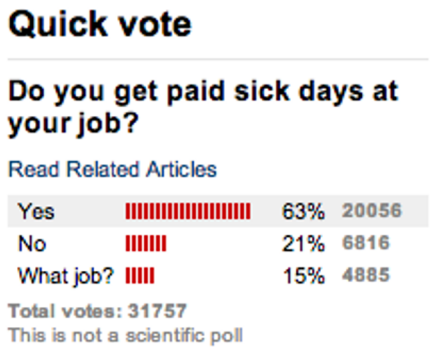
(b) Cluster sampling

(c) Stratified sampling

(d) Blocked sampling

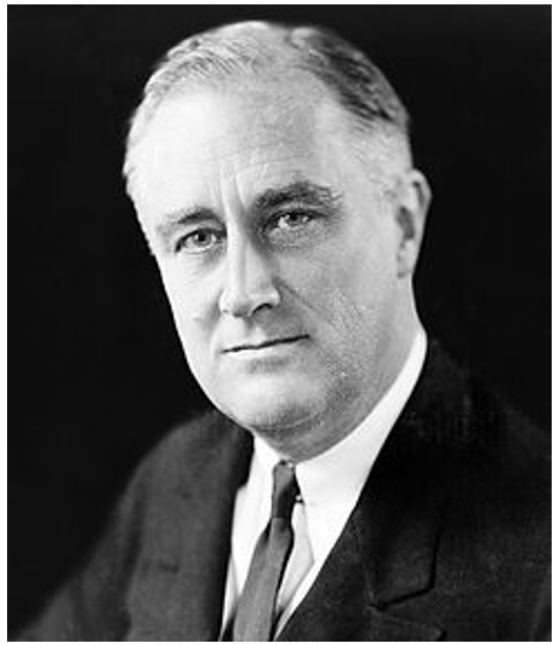
Sampling Bias

- Non-response: If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- Voluntary response: Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue. Such a sample will also not be representative of the population.



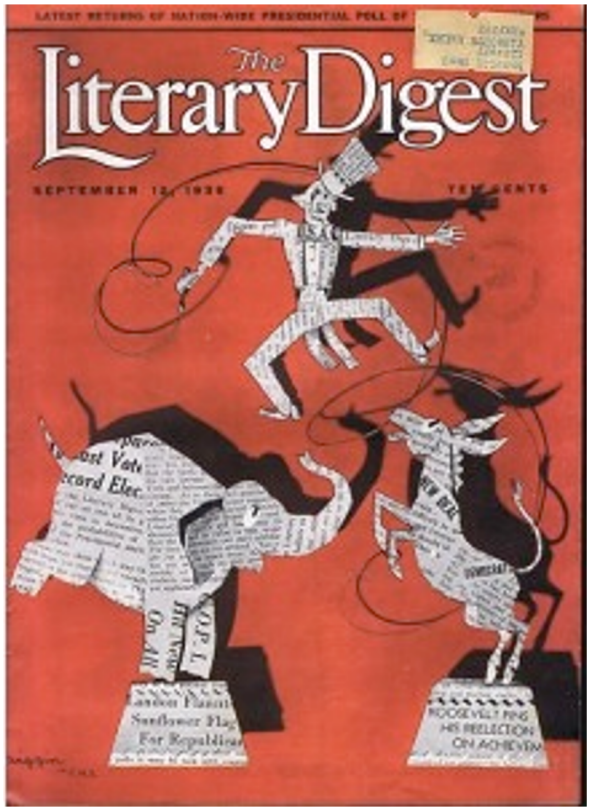
Sampling Bias Example

- A historical example of a biased sample yielding misleading results
- In 1936, Landon sought the Republican presidential nomination opposing the re-election of FDR.



The Literary Digest Poll

- The Literary Digest polled about 10 million Americans, and got responses from about 2.4 million.
- The poll showed that Landon would likely be the overwhelming winner and FDR would get only 43% of the votes.
- Election result: FDR won, with 62% of the votes.
- The magazine was completely discredited because of the poll, and was soon discontinued.



The Literary Digest Poll – What Went Wrong

- The magazine had surveyed
  - its own readers,
  - registered automobile owners, and
  - registered telephone users.

- These groups had incomes well above the national average of the day (remember, this is Great Depression era) which resulted in lists of voters far more likely to support Republicans than a truly typical voter of the time, i.e. the sample was not representative of the American population at the time.

Large Samples are Preferred but…

- The Literary Digest election poll was based on a sample size of 2.4 million, which is huge, but since the sample was biased, the sample did not yield an accurate prediction.

- Back to the soup analogy: If the soup is not well stirred, it doesn't matter how large a spoon you have, it will still not taste right. If the soup is well stirred, a small spoon will suffice to test the soup.

Practice

A school district is considering whether it will no longer allow high school students to park at school after two recent accidents where students were severely injured. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of 6,000 surveys that go out, 1,200 are returned. Of these 1,200 surveys that were completed, 960 agreed with the policy change and 240 disagreed. Which of the following statements are true?

I. Some of the mailings may have never reached the parents.

II. The school district has strong support from parents to move forward with the policy approval.

III. It is possible that majority of the parents of high school students disagree with the policy change.

IV. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only I (b) I and II (c) I and III (d) III and IV (e) Only IV

Practice

A school district is considering whether it will no longer allow high school students to park at school after two recent accidents where students were severely injured. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of 6,000 surveys that go out, 1,200 are returned. Of these 1,200 surveys that were completed, 960 agreed with the policy change and 240 disagreed. Which of the following statements are true?

I. Some of the mailings may have never reached the parents.

II. The school district has strong support from parents to move forward with the policy approval.

III. It is possible that majority of the parents of high school students disagree with the policy change.

IV. The survey results are unlikely to be biased because all parents were mailed a survey.

(a) Only I (b) I and II (c) I and III (d) III and IV (e) Only IV

Observational Studies

- Researchers collect data in a way that does not directly interfere with how the data arise.
- Collect information through surveys, review company or medical records
- Results of an observational study can generally be used to establish an association between the explanatory and response variables.

Prospective vs Retrospective Studies

- A prospective study identifies individuals and collects information as events unfold.
  - Example: The Nurses Health Study has been recruiting registered nurses and then collecting data from them using questionnaires since 1976.

- Retrospective studies collect data after events have taken place.
  - Example: Researchers reviewing past events in medical records.

Experimental Studies

- Randomly assign subjects to treatments to establish causal connections between variables
- 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 Variable

- 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 Variable - 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?

Principles of Experimental Design

- Control: Compare treatment of interest to a control group.
- Randomize: Randomly assign subjects to treatments, and randomly sample from the population whenever possible.
- Replicate: Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.
- Block: If there are variables that are known or suspected to affect the response variable, first group subjects into blocks based on these variables, and then randomize cases within each block to treatment groups.

Blocking

- We would like to design an experiment to investigate if energy gels makes you run faster:
  - Treatment: energy gel
  - Control: no energy gel
  - It is suspected that energy gels might affect pro and amateur athletes differently, therefore we block for pro status:
    - Divide the sample to pro and amateur
    - Randomly assign pro athletes to treatment and control groups
    - Randomly assign amateur athletes to treatment and control groups
    - Pro/amateur status is equally represented in the resulting treatment and control groups

Why is this important? Can you think of other variables to block for?

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?

- There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)

- There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)

- There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)

- 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?

- There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)

- There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)

- There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)

- There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

Blocking vs Explanatory Variables

- 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.

More Terminology on Experimental Design

- 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

Random Assignment

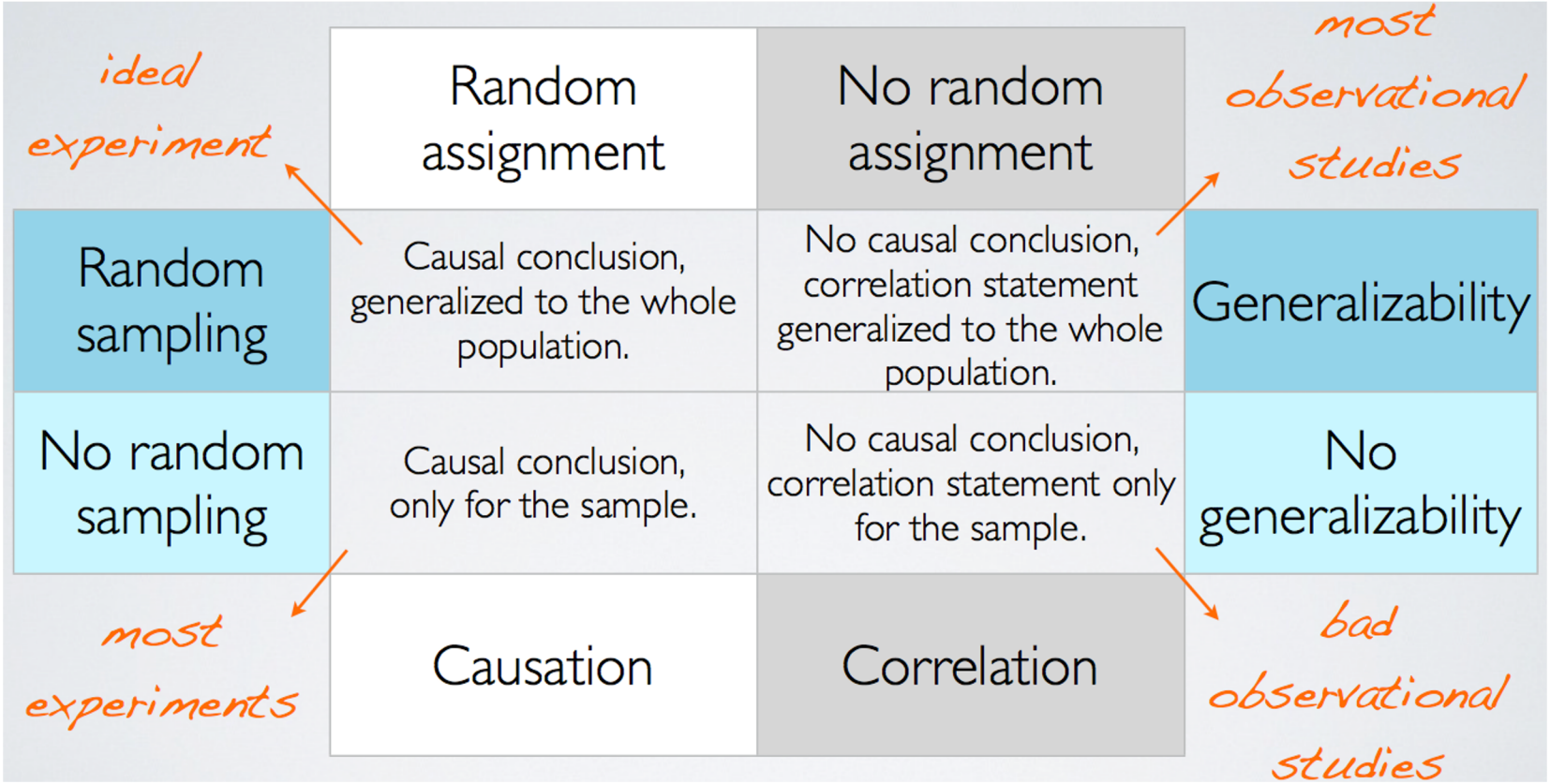
- Each participant is independently and randomly assigned to one (and only one) group
- Each participant has the same probability of being assigned to each group
- Most often the groups are made to have the same number of participants
- A violation of true random assignment

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

Random Assignment vs Random Sampling



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

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

Textbook Reading

Chapter 1- Review Exercises from end of Chapter

- 1.3-1.5
- 1.13-1.16
- 1.22-1.27
- 1.35-1.44

Textbook Problems

- Chapter 1- Review Exercises from end of Chapter
  - 1.3-1.5
  - 1.13-1.16
  - 1.22-1.27
  - 1.35-1.44

Assignment 1

15 problems

You may work with your friend but need to submit your own work