Data colection + Exploratory data analysis

Sergio I. Garcia-Rios

Government 3990: Statistics in the Social Science

Data Collection + Observational studies and experiments

Use a sample to make inferences about the population

1. Use a sample to make inferences about the population

- Ultimate goal: make inferences about populations

1. Use a sample to make inferences about the population

- Ultimate goal: make inferences about populations
- Caveat: populations are difficult or impossible to access

1. Use a sample to make inferences about the population

- Ultimate goal: make inferences about populations
- Caveat: populations are difficult or impossible to access
- Solution: use a sample from that population, and use statistics from that sample to make inferences about the unknown population parameters

1. Use a sample to make inferences about the population

- Ultimate goal: make inferences about populations
- Caveat: populations are difficult or impossible to access
- Solution: use a sample from that population, and use statistics from that sample to make inferences about the unknown population parameters
- The better (more representative) sample we have, the more reliable our estimates and more accurate our inferences will be

1. Use a sample to make inferences about the population

- Ultimate goal: make inferences about populations
- Caveat: populations are difficult or impossible to access
- Solution: use a sample from that population, and use statistics from that sample to make inferences about the unknown population parameters
- The better (more representative) sample we have, the more reliable our estimates and more accurate our inferences will be

1. Use a sample to make inferences about the population

- Ultimate goal: make inferences about populations
- Caveat: populations are difficult or impossible to access
- Solution: use a sample from that population, and use statistics from that sample to make inferences about the unknown population parameters
- The better (more representative) sample we have, the more reliable our estimates and more accurate our inferences will be

Your Turn

Suppose we want to know how many offspring female squirrels have, on average. It's not feasible to obtain offspring data from on all female squirrels, so we use data from the Cornell Squirrel Center. We use the sample mean from these data as an estimate for the unknown population mean. Can you see any limitations to using data from the Cornell Squirrel Center to make inferences about all squirrels?

Sampling is natural

- 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)

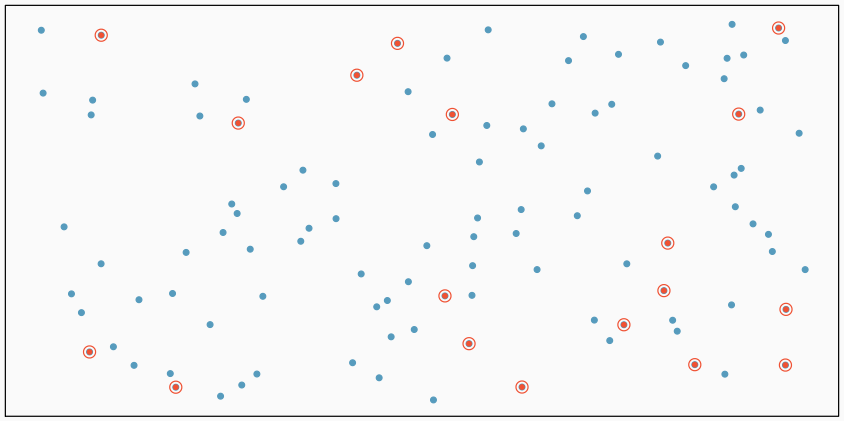
Sampling is natural

Sampling is natural

Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

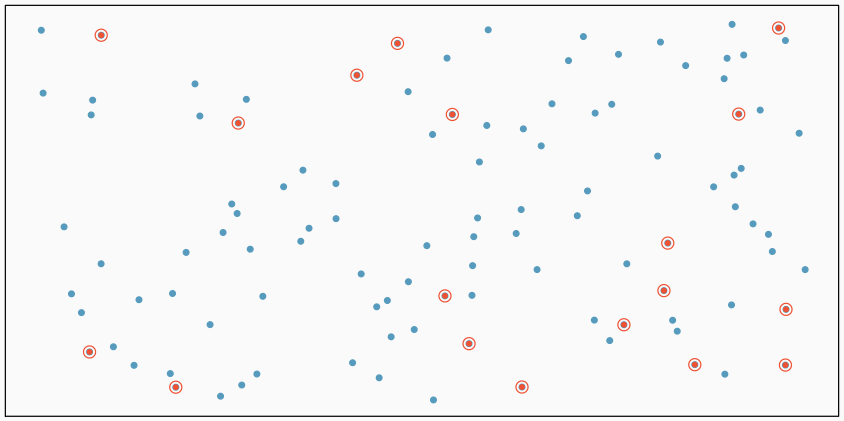
Simple random:

Drawing names from a hat



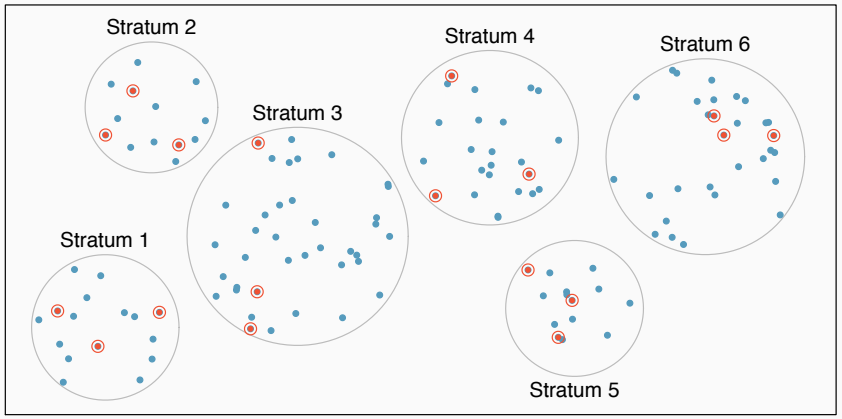
Simple random:

Drawing names from a hat



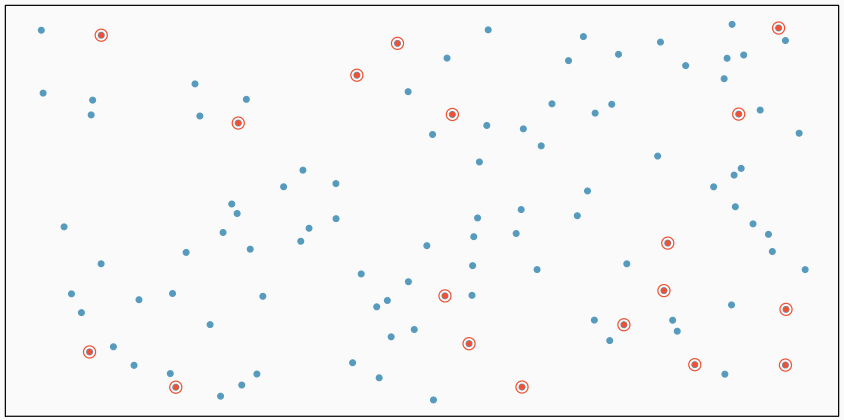
Stratified: homogenous strata

Stratify to control for SES



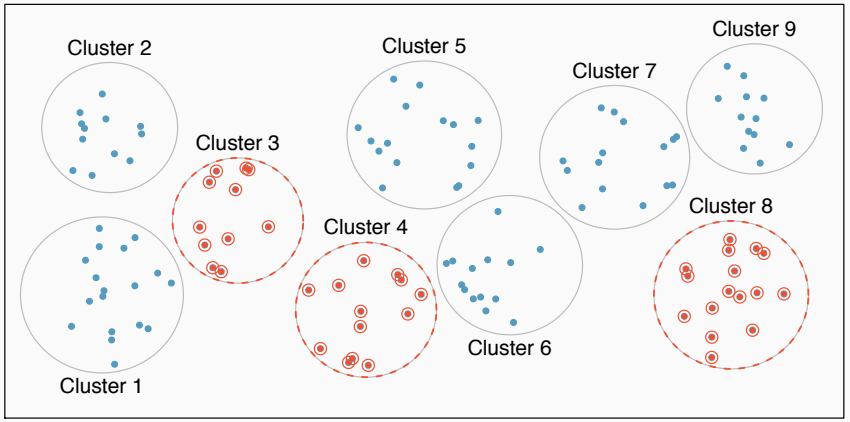
Simple random:

Drawing names from a hat



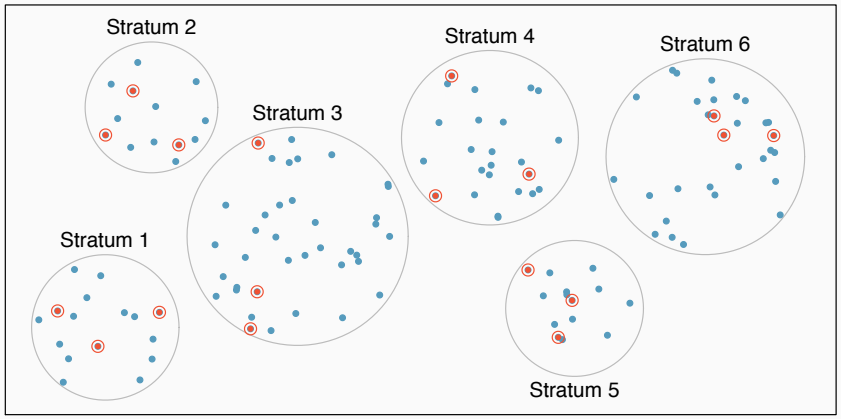
Cluster: heterogenous clusters

Sample all chosen clusters



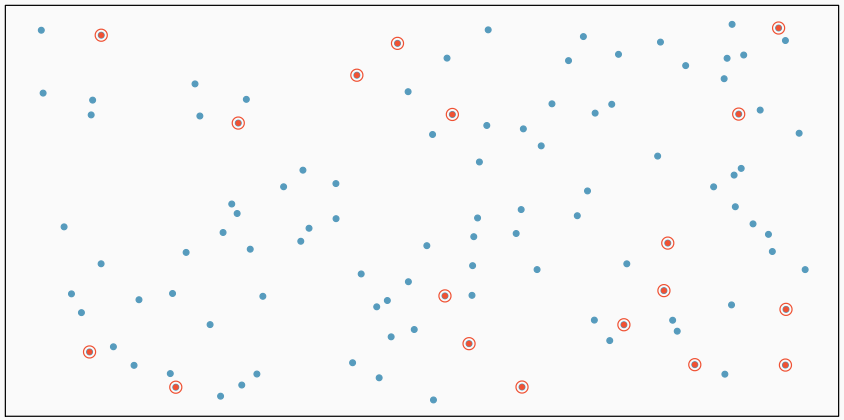
Stratified: homogenous strata

Stratify to control for SES



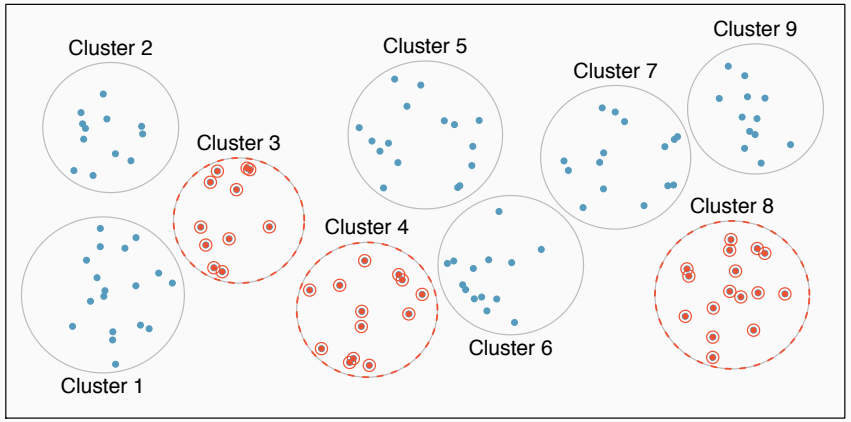
Simple random:

Drawing names from a hat



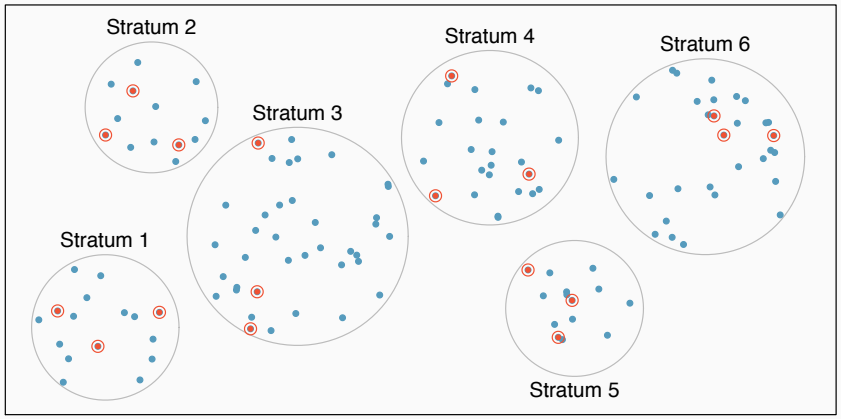
Cluster: heterogenous clusters

Sample all chosen clusters



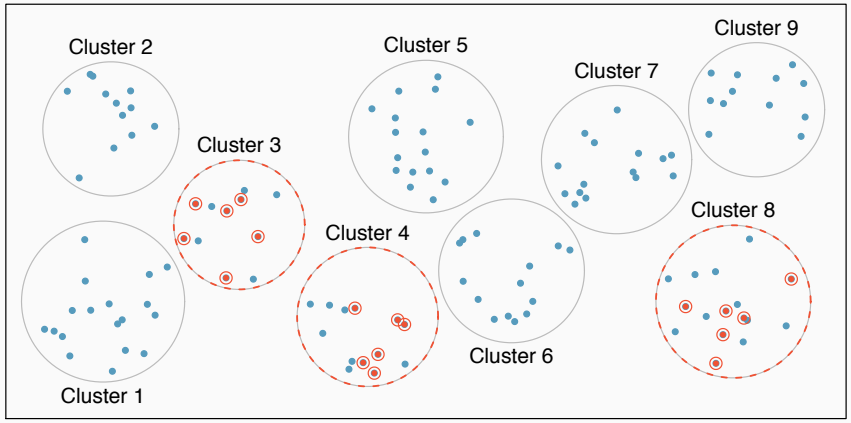
Stratified: homogenous strata

Stratify to control for SES



Multistage:

Random sample in chosen clusters



Your Turn

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, and others a diverse mixture of housing structures. Which approach would likely be the *least* effective?

(a) Simple random sampling

(b) Stratified sampling, where each stratum is a neighborhood

(c) Cluster sampling, where each cluster is a neighborhood

Your Turn

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, and others a diverse mixture of housing structures. Which approach would likely be the *least* effective?

(a) Simple random sampling

(b) Stratified sampling, where each stratum is a neighborhood

(c) **Cluster sampling, where each cluster is a neighborhood**

Sampling schemes can suffer from a variety of biases

3. Sampling schemes can suffer from a variety of biases

- 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

3. Sampling schemes can suffer from a variety of biases

- 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 since such a sample will also not be representative of the population

3. Sampling schemes can suffer from a variety of biases

- 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 since such a sample will also not be representative of the population
- Convenience sample: Individuals who are easily accessible are more likely to be included in the sample

Your Turn

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. Overall, 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

Your Turn

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. Overall, 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

Experiments use random assignment to treatment groups, observational studies do not

What type of study is this? What is the scope of inference (causality / generalizability)?[1]

Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry

By VINDU GOEL JUNE 29, 2014

The New York Times

In an academic paper published in conjunction with two university researchers, the company reported that, for one week in January 2012, it had altered the number of positive and negative posts in the news feeds of 689,003 randomly selected users to see what effect the changes had on the tone of the posts the recipients then wrote.

The researchers found that moods were contagious. The people who saw more positive posts responded by writing more positive posts. Similarly, seeing more negative content prompted the viewers to be more negative in their own posts.

[1] http://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html

4. Experiments use random assignment to treatment groups, observational studies do not

Your Turn

A study that surveyed a random sample of otherwise healthy adults found that people are more likely to get muscle cramps when they're stressed. The study also noted that people drink more coffee and sleep less when they're stressed. What type of study is this?

What is the conclusion of the study?

Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

4. Experiments use random assignment to treatment groups, observational studies do not

Your Turn

A study that surveyed a random sample of otherwise healthy adults found that people are more likely to get muscle cramps when they're stressed. The study also noted that people drink more coffee and sleep less when they're stressed. What type of study is this?

Observational

What is the conclusion of the study?

Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

4. Experiments use random assignment to treatment groups, observational studies do not

Your Turn

A study that surveyed a random sample of otherwise healthy adults found that people are more likely to get muscle cramps when they're stressed. The study also noted that people drink more coffee and sleep less when they're stressed. What type of study is this?

Observational

What is the conclusion of the study?

There is an association between increased stress & muscle cramps.

Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

4. Experiments use random assignment to treatment groups, observational studies do not

Your Turn

A study that surveyed a random sample of otherwise healthy adults found that people are more likely to get muscle cramps when they're stressed. The study also noted that people drink more coffee and sleep less when they're stressed. What type of study is this?

Observational

What is the conclusion of the study?

There is an association between increased stress & muscle cramps.

Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

Muscle cramps might also be due to increased caffeine consumption or sleeping less – these are potential confounding variables.

Four principles of experimental design: randomize, control, block, replicate

5. Four principles of experimental design: randomize, control, block, replicate

- We would like to design an experiment to investigate if increased stress causes muscle cramps:

5. Four principles of experimental design: randomize, control, block, replicate

- We would like to design an experiment to investigate if increased stress causes muscle cramps:
  - Treatment: increased stress
  - Control: no or baseline stress

5. Four principles of experimental design: randomize, control, block, replicate

- We would like to design an experiment to investigate if increased stress causes muscle cramps:
  - Treatment: increased stress
  - Control: no or baseline stress
- It is suspected that the effect of stress might be different on younger and older people: block for age.

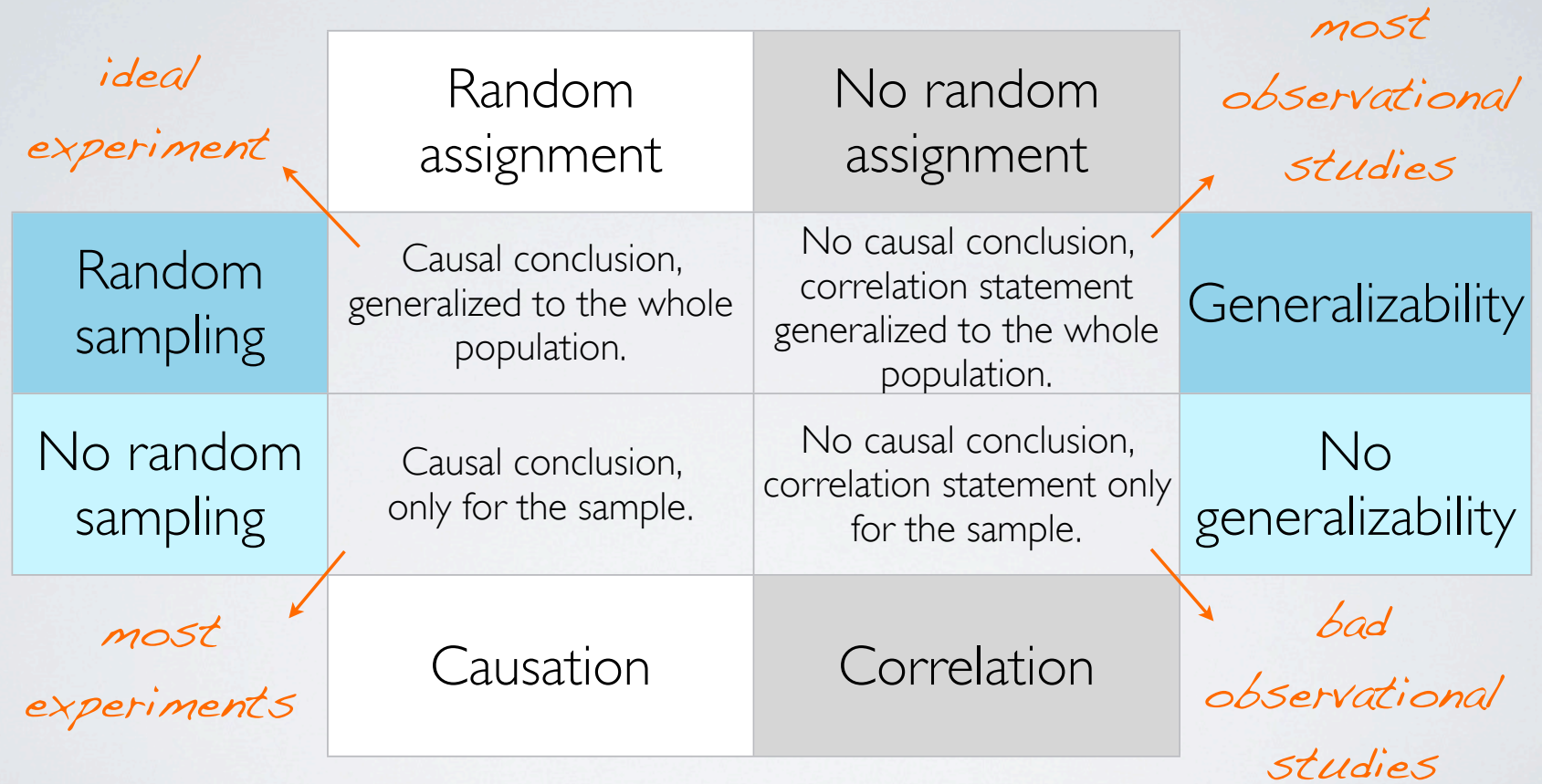
5. Four principles of experimental design: randomize, control, block, replicate

- We would like to design an experiment to investigate if increased stress causes muscle cramps:
  - Treatment: increased stress
  - Control: no or baseline stress
- It is suspected that the effect of stress might be different on younger and older people: block for age.

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

Random sampling helps generalizability, random assignment helps causality

6. Random sampling helps generalizability, random assignment helps causality



Summary

Summary of main ideas

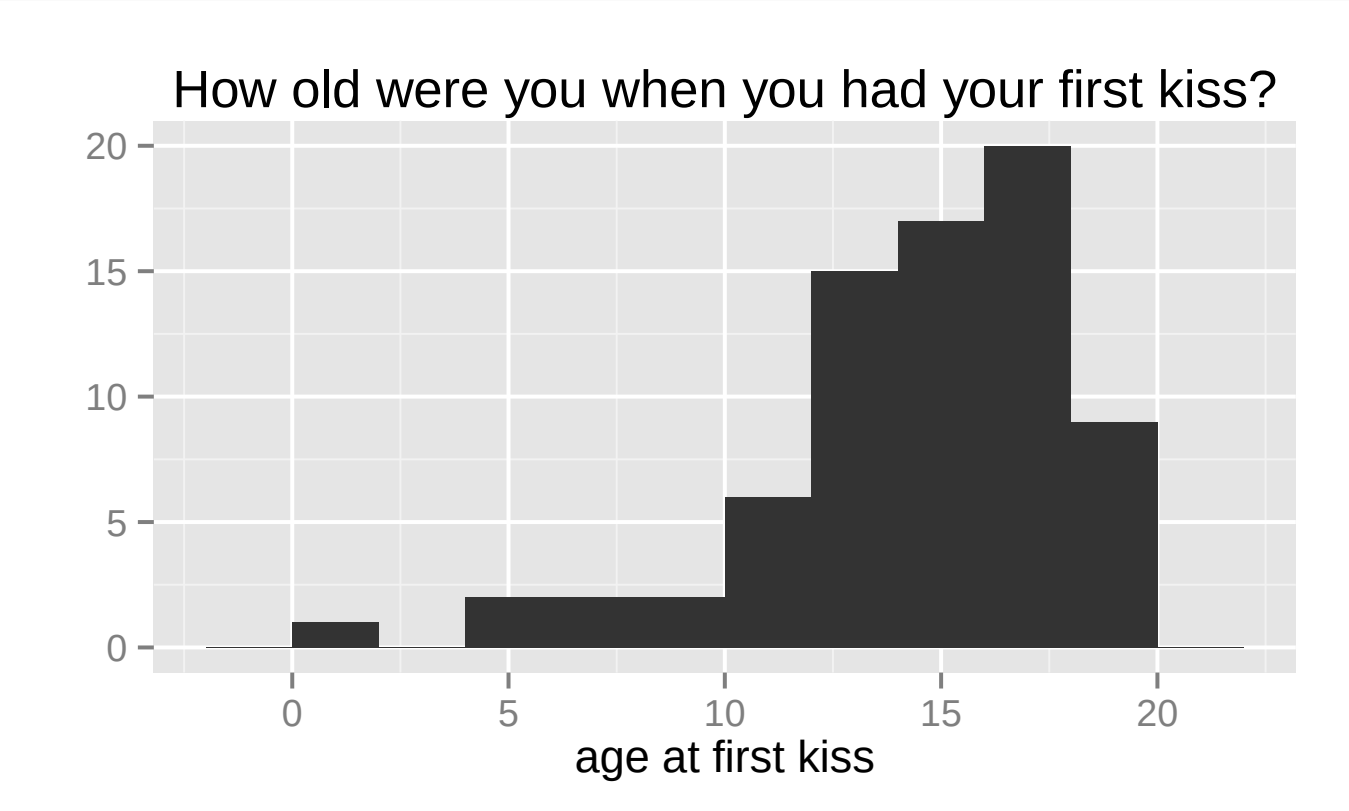
1. Use a sample to make inferences about the population
2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
3. Sampling schemes can suffer from a variety of biases
4. Experiments use random assignment to treatment groups, observational studies do not
5. Four principles of experimental design: randomize, control, block, replicate
6. Random sampling helps generalizability, random assignment helps causality

Exploratory data analysis

Always start your exploration with a visualization

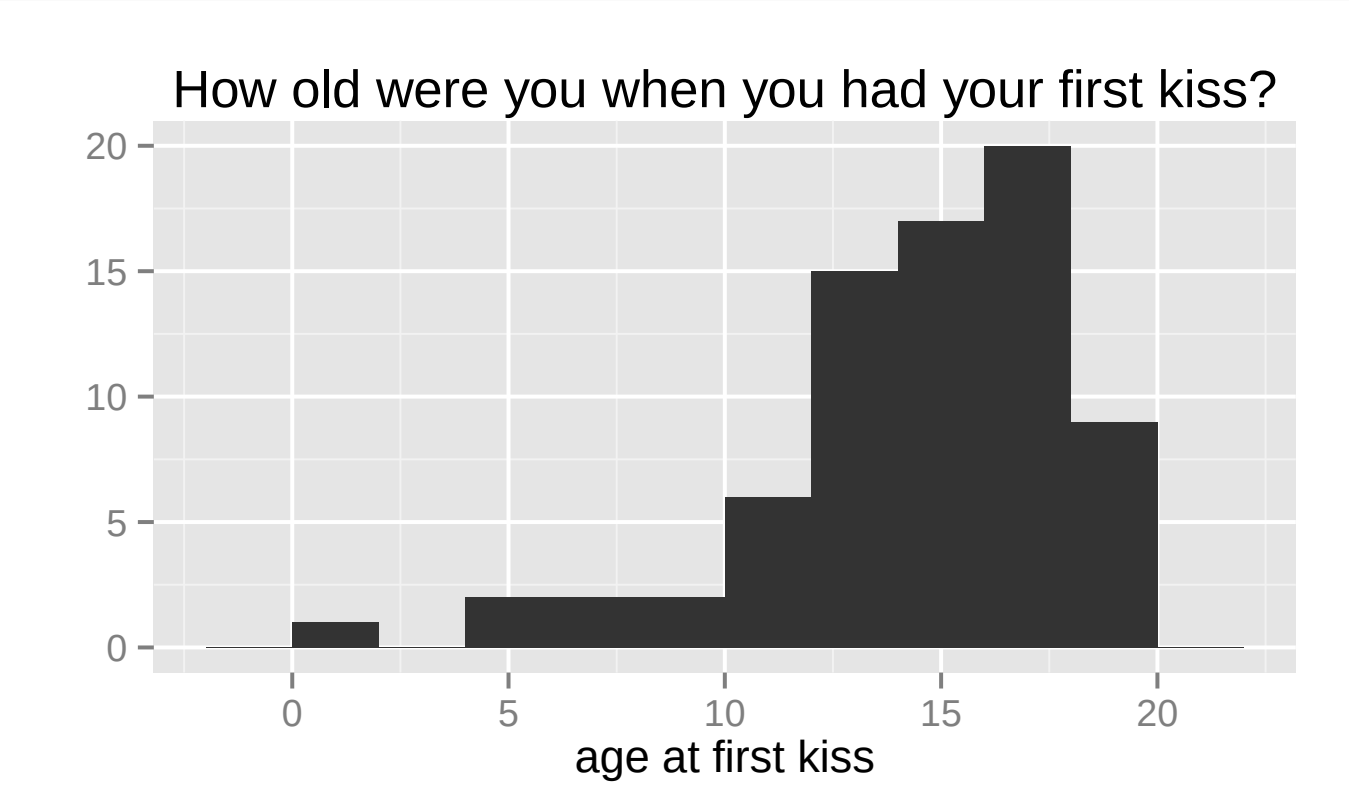
From a class survey...

Do you see anything out of the ordinary?



From a class survey...

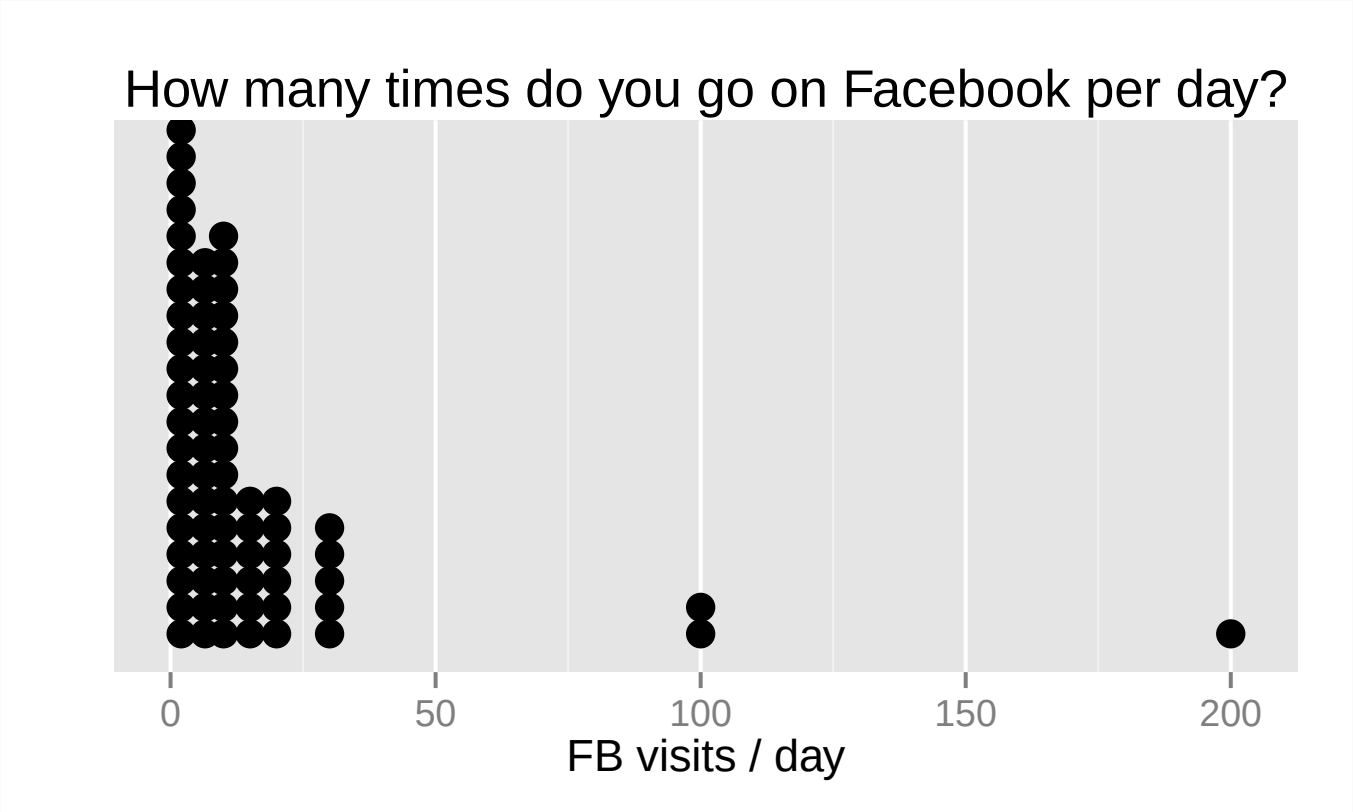
Do you see anything out of the ordinary?



Some people reported very low ages, which might suggest the survey question wasn't clear: romantic kiss or any kiss?

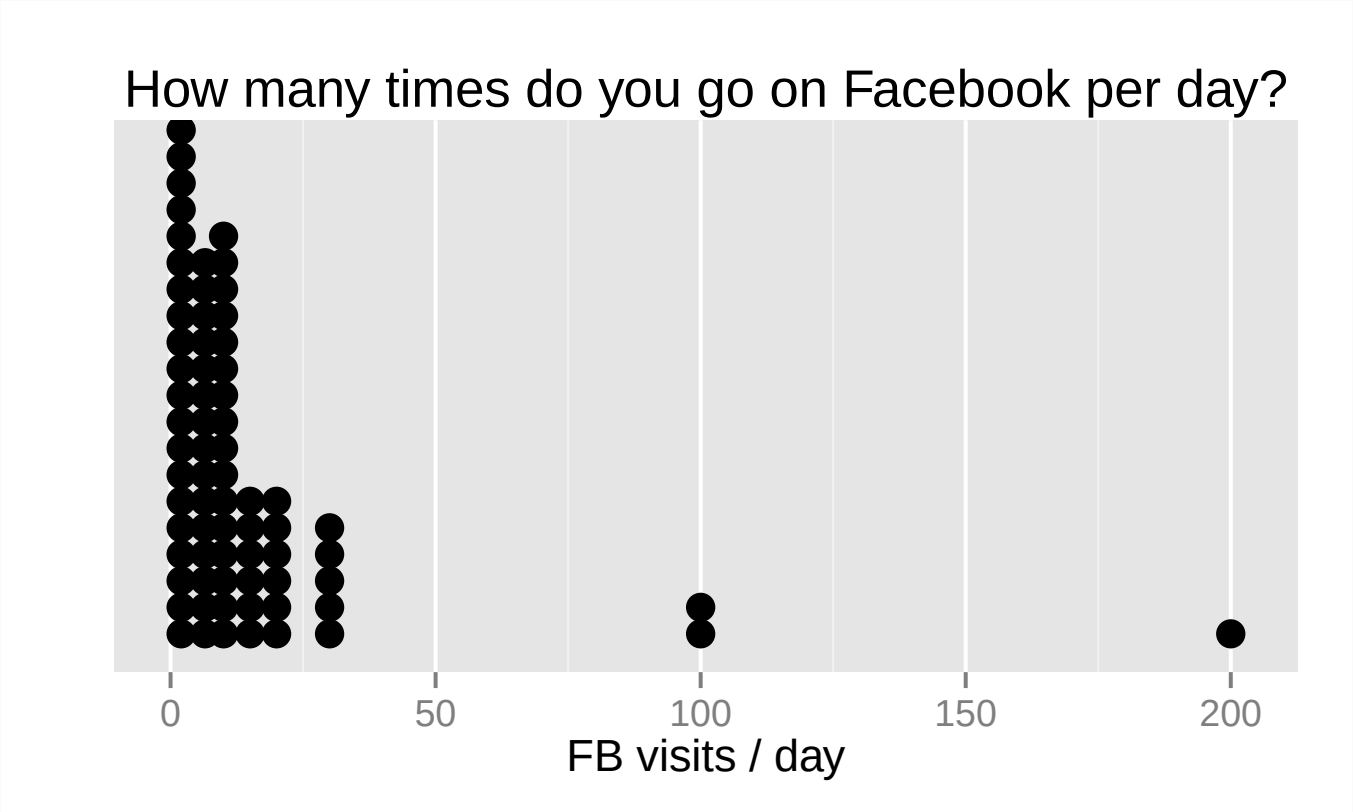
From a class survey...

How are people reporting lower vs. higher values of FB visits?



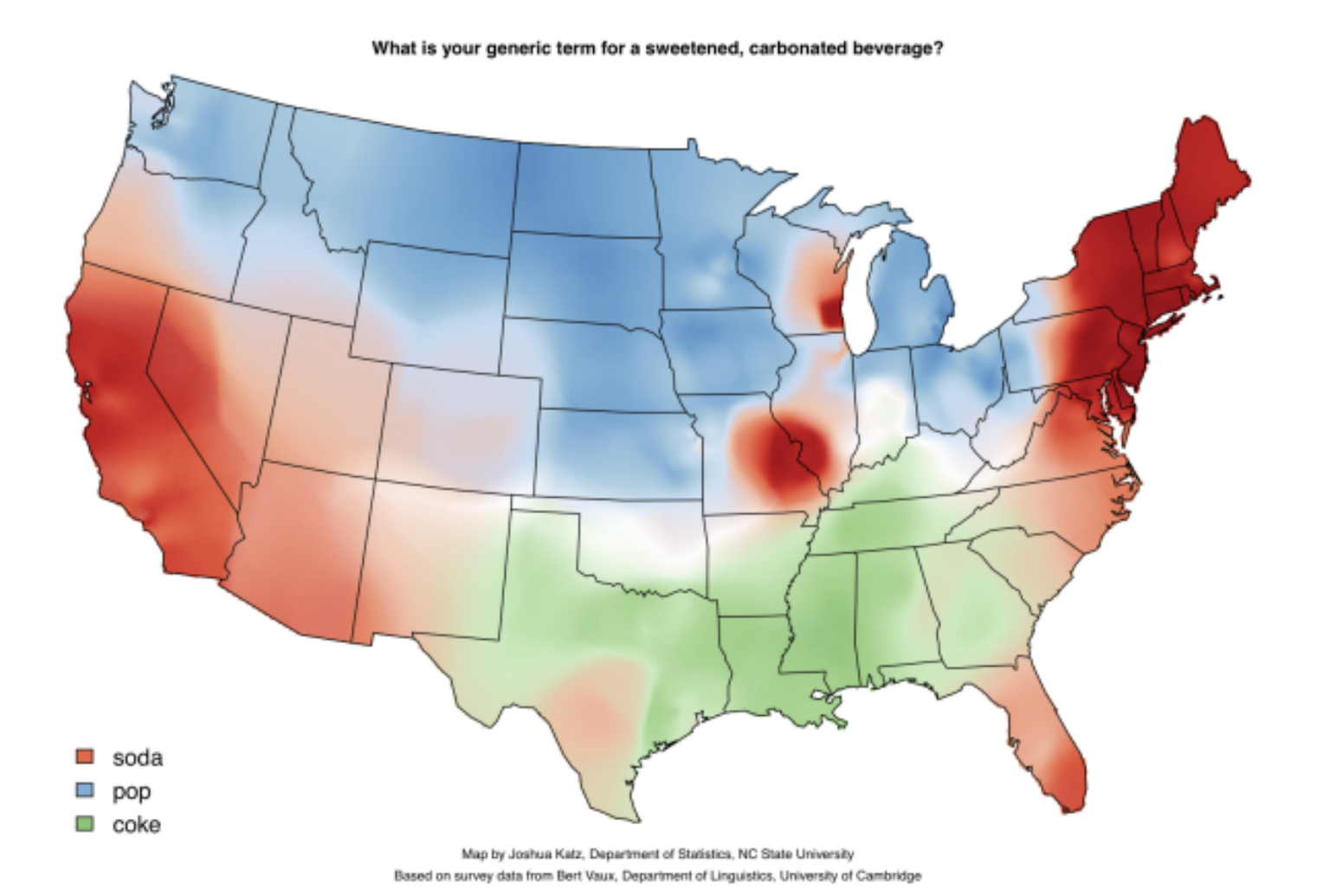
From a class survey...

How are people reporting lower vs. higher values of FB visits?



Finer scale for lower numbers.

Describe the spatial distribution of preferred sweetened carbonated beverage drink.



What is missing in this visualization?

When describing numerical distributions discuss shape, center, spread, and unusual observations

Describing distributions of numerical variables

- Shape: skewness, modality
- Center: an estimate of a typical observation in the distribution (mean, median, mode, etc.)
    - Notation: μ: population mean, $\bar{x}$: sample mean
- Spread: measure of variability in the distribution (standard deviation, IQR, range, etc.)
- Unusual observations: observations that stand out from the rest of the data that may be suspected outliers

Your Turn

Which of these is most likely to have a roughly symmetric distribution?

(a) salaries of a random sample of people from NY

(b) weights of adult females

(c) scores on an well-designed exam

(d) last digits of phone numbers

Your Turn

Which of these is most likely to have a roughly symmetric distribution?

(a) salaries of a random sample of people from NY

(b) **weights of adult females**

(c) scores on an well-designed exam

(d) last digits of phone numbers

Mean vs. median

Your Turn

How do the mean and median of the following two datasets compare?

Dataset 1: 30, 50, 70, 90
Dataset 2: 30, 50, 70, 1000

(a) $\bar{x}_1 = \bar{x}_2$, $median_1 = median_2$

(b) $\bar{x}_1 < \bar{x}_2$, $median_1 = median_2$

(c) $\bar{x}_1 < \bar{x}_2$, $median_1 < median_2$

(d) $\bar{x}_1 > \bar{x}_2$, $median_1 < median_2$

(e) $\bar{x}_1 > \bar{x}_2$, $median_1 = median_2$

Mean vs. median

Your Turn

How do the mean and median of the following two datasets compare?

Dataset 1: 30, 50, 70, 90
Dataset 2: 30, 50, 70, 1000

(a) $\bar{x}_1 = \bar{x}_2$, $median_1 = median_2$

(b) $\bar{x}_1 < \bar{x}_2$, $median_1 = median_2$

(c) $\bar{x}_1 < \bar{x}_2$, $median_1 < median_2$

(d) $\bar{x}_1 > \bar{x}_2$, $median_1 < median_2$

(e) $\bar{x}_1 > \bar{x}_2$, $median_1 = median_2$

Standard deviation and variance

- Most commonly used measure of variability is the standard deviation, which roughly measures the average deviation from the mean
  - Notation: σ: population standard deviation, s: sample standard deviation
- Calculating the standard deviation, for a population (rarely, if ever) and for a sample:

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

- Square of the standard deviation is called the variance.

More on SD

Why divide by $n - 1$ instead of n when calculating the sample standard deviation?

More on SD

Why divide by $n - 1$ instead of n when calculating the sample standard deviation?

Lose a "degree of freedom" for using an estimate (the sample mean, $\bar{x}$), in estimating the sample variance/standard deviation.

More on SD

Why divide by $n - 1$ instead of n when calculating the sample standard deviation?

Lose a "degree of freedom" for using an estimate (the sample mean, $\bar{x}$), in estimating the sample variance/standard deviation.

Why do we use the squared deviation in the calculation of variance?

More on SD

Why divide by $n - 1$ instead of n when calculating the sample standard deviation?

Lose a "degree of freedom" for using an estimate (the sample mean, $\bar{x}$), in estimating the sample variance/standard deviation.

Why do we use the squared deviation in the calculation of variance?

- To get rid of negatives so that observations equally distant from the mean are weighed equally.
- To weigh larger deviations more heavily.

Range and IQR

Our Turn

For the given data set: 7, 6, 5, 5, 9, 10, 11, 10, 9

Calculate

- Range
- Median
- The three quartiles
- Interquartile range (IQR)
- Draw a boxplot

Robust statistics are not easily affected by outliers and extreme skew

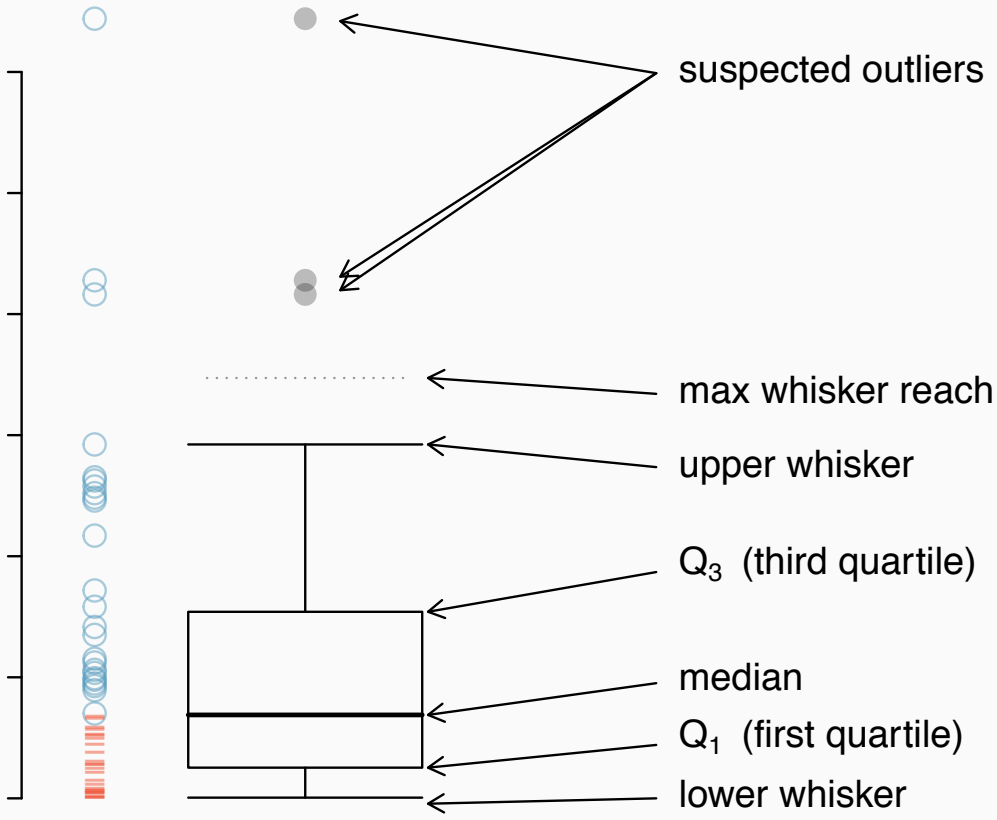
Robust statistics

- Mean and standard deviation are easily affected by extreme observations since the value of each data point contributes to their calculation.
- Median and IQR are more robust.
- Therefore we choose median&IQR (over mean&SD) when describing skewed distributions.

Use box plots to display quartiles, median, and outliers

Box plot

A box plot visualizes the median, the quartiles, and suspected outliers. An outlier is defined as an observation more than 1.5×IQR away from the quartiles.



Aplication Exercise

1.1 Distributions of numerical variables

Summary

Summary of main ideas

1. Always start your exploration with a visualization
2. When describing numerical distributions discuss shape, center, spread, and unusual observations
3. Robust statistics are not easily affected by outliers and extreme skew
4. Use box plots to display quartiles, median, and outliers