CHAPTER 1: INTRODUCTION TO DATA

math 189 : Data Analysis and Inference : winter 2018

Jelena Bradic
http://www.jelenabradic.net
Assistant Professor, Department of Mathematics, University of California, San Diego
jbradic@ucsd.edu

Case study

Data basics

Overview of data collection principles

Observational studies and sampling strategies

Experiments

Examining numerical data

Considering categorical data

Case study: Gender discrimination

Jelena Bradic http://www.jelenabradic.net Assistant Professor, Department of Mathematics, University of California, San Diego jbradic@ucsd.edu

Treating Chronic Fatigue Syndrome

* Objective: Evaluate the effectiveness of cognitive-behavior therapy for chronic fatigue syndrome.
* Participant pool: 142 patients who were recruited from referrals by primary care physicians and consultants to a hospital clinic specializing in chronic fatigue syndrome.
* Actual participants: Only 60 of the 142 referred patients entered the study. Some were excluded because they didn't meet the diagnostic criteria, some had other health issues, and some refused to be a part of the study.

Deale et. al. Cognitive behavior therapy for chronic fatigue syndrome: A randomized controlled trial. The American Journal of Psychiatry 154.3 (1997).

Study design

* Patients randomly assigned to treatment and control groups, 30 patients in each group:
  * Treatment: Cognitive behavior therapy – collaborative, educative, and with a behavioral emphasis. Patients were shown on how activity could be increased steadily and safely without exacerbating symptoms.
  * Control: Relaxation – No advice was given about how activity could be increased. Instead progressive muscle relaxation, visualization, and rapid relaxation skills were taught.

Results

The table below shows the distribution of patients with good outcomes at 6-month follow-up. Note that 7 patients dropped out of the study: 3 from the treatment and 4 from the control group.

		Good outcome		
		Yes	No	Total
Group	Treatment	19	8	27
	Control	5	21	26
	Total	24	29	53

* Proportion with good outcomes in treatment group:

$$19/27 \approx 0.70 \rightarrow 70\%$$

* Proportion with good outcomes in control group:

$$5/26 \approx 0.19 \rightarrow 19\%$$

Understanding the results

Do the data show a “real” difference between the groups?

* Suppose you flip a coin 100 times. While the chance a coin lands heads in any given coin flip is 50%, we probably won’t observe exactly 50 heads. This type of fluctuation is part of almost any type of data generating process.
* The observed difference between the two groups (70 - 19 = 51%) may be real, or may be due to natural variation.
* Since the difference is quite large, it is more believable that the difference is real.
* We need statistical tools to determine if the difference is so large that we should reject the notion that it was due to chance.

Generalizing the results

Are the results of this study generalizable to all patients with chronic fatigue syndrome?

These patients had specific characteristics and volunteered to be a part of this study, therefore they may not be representative of all patients with chronic fatigue syndrome. While we cannot immediately generalize the results to all patients, this first study is encouraging. The method works for patients with some narrow set of characteristics, and that gives hope that it will work, at least to some degree, with other patients.

Case study

Data basics

- Observations and variables
- Types of variables
- Relationships among variables
- Associated and independent variables

Overview of data collection principles

Observational studies and sampling strategies

Experiments

Examining numerical data

Considering categorical data

Case study: Gender discrimination

Jelena Bradic, http://www.jelenabradic.net, Assistant Professor, Department of Mathematics, University of California, San Diego, jbradic@ucsd.edu

Data matrix

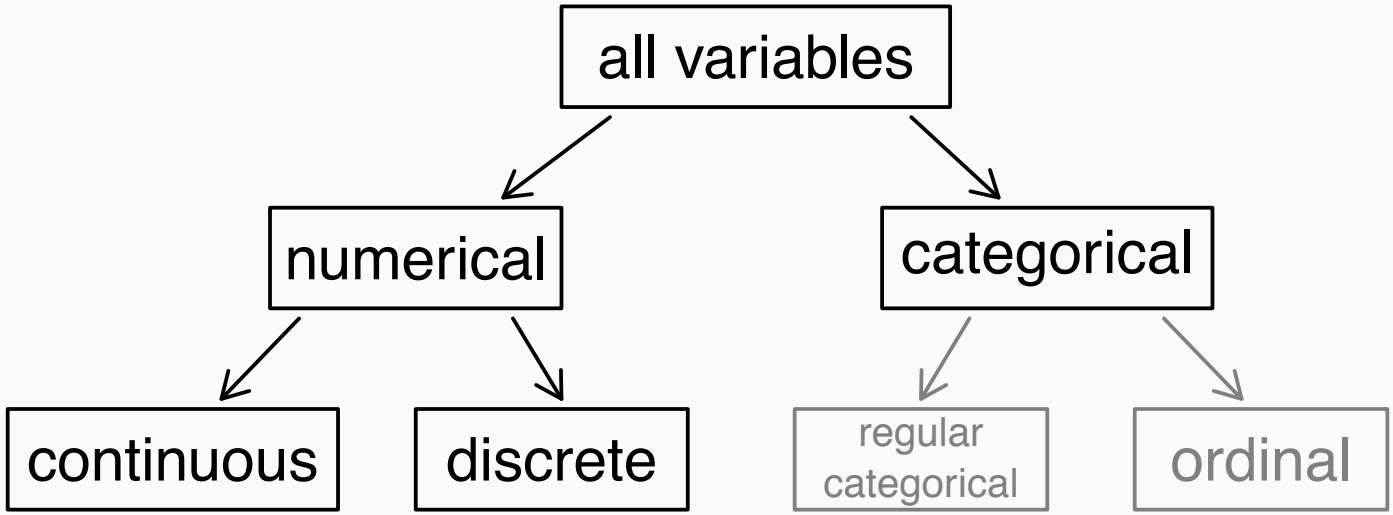
Data collected on students in a statistics class on a variety of variables:

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

Types of variables



Types of variables (cont.)

	gender	sleep	bedtime	countries	dread
1	male	5	12-2	13	3
2	female	7	10-12	7	2
3	female	5.5	12-2	1	4
4	female	7	12-2		2
5	female	3	12-2	1	3
6	female	3	12-2	9	4

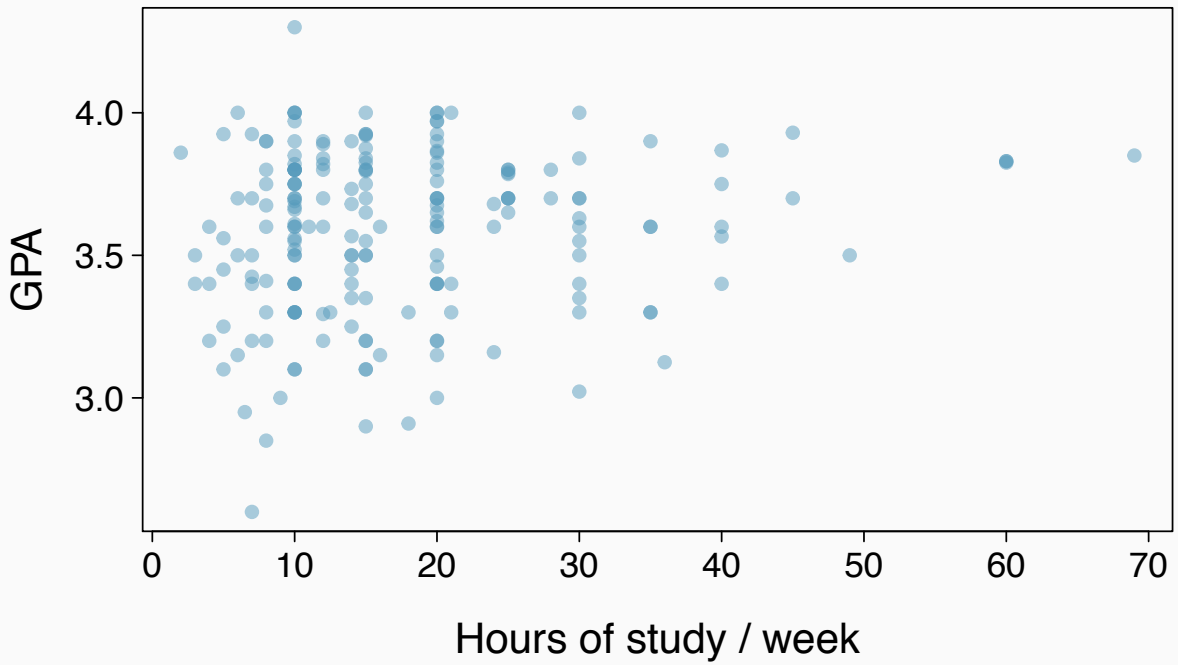
* `gender`: categorical
* `sleep`: numerical, continuous
* `bedtime`: categorical, ordinal
* `countries`: numerical, discrete
* `dread`: categorical, ordinal - could also be used as numerical

Practice

What type of variable is a telephone area code?

* numerical, continuous
* numerical, discrete
* categorical
* categorical, ordinal

Does there appear to be a relationship between number of alcoholic drinks consumed per week and age at first alcohol consumption?

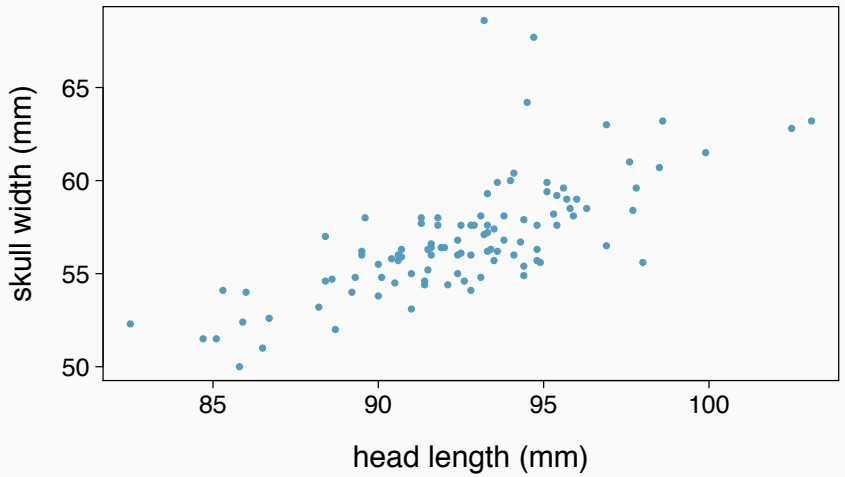


Can you spot anything unusual about any of the data points?

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

Practice

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



* There is no relationship between head length and skull width, i.e. the variables are independent.
* Head length and skull width are positively associated.
* Skull width and head length are negatively associated.
* A longer head causes the skull to be wider.
* A wider skull causes the head to be longer.

Associated vs. independent

* 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.

Case study

Data basics

Overview of data collection principles

- Populations and samples
- Anecdotal evidence
- Sampling from a population
- Explanatory and response variables
- Observational studies and experiments

Observational studies and sampling strategies

Experiments

Examining numerical data

Considering categorical data

Case study: Gender discrimination

Jelena Bradic, http://www.jelenabradic.net, Assistant Professor, Department of Mathematics, University of California, San Diego, jbradic@ucsd.edu

Populations and samples

PHYS ED | AUGUST 29, 2012, 12:01 AM | 21 Comments

Finding Your Ideal Running Form

By GRETCHEN REYNOLDS

David De Lossy/Getty Images

http://well.blogs.nytimes.com/2012/08/29/finding-your-ideal-running-form

Research question: Can people become better, more efficient runners on their own, merely by running?

Population of interest: All people

Sample: Group of adult women who recently joined a running group

Population to which results can be generalized: Adult women, if the data are randomly sampled

Anecdotal evidence and early smoking research

* Anti-smoking research started in the 1930s and 1940s when cigarette smoking became increasingly popular. While some smokers seemed to be sensitive to cigarette smoke, others were completely unaffected.
* Anti-smoking research was faced with resistance based on anecdotal evidence such as “My uncle smokes three packs a day and he’s in perfectly good health”, evidence based on a limited sample size that might not be representative of the population.
* It was concluded that “smoking is a complex human behavior, by its nature difficult to study, confounded by human variability.”
* In time researchers were able to examine larger samples of cases (smokers), and trends showing that smoking has negative health impacts became much clearer.

Brandt, The Cigarette Century (2009), Basic Books.

Census

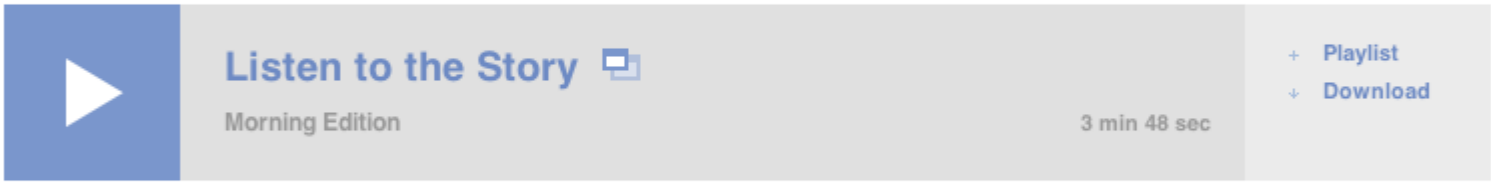
* Wouldn't it be better to just include everyone and "sample" the entire population?
  * This is called a census.
* There are problems with taking a census:
  * It can be difficult to complete a census: there always seem to be some individuals who are hard to locate or hard to measure. And these difficult-to-find 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.

Illegal Immigrants Reluctant To Fill Out Census Form

by PETER O'DOWD

March 31, 2010 4:00 AM

from KJZZ



There is an effort underway to make sure Hispanics are accurately counted in the 2010 Census. Phoenix has some of the country's "hardest-to-count" districts. Some Latinos, especially illegal residents, fear that participating in the count will expose them to immigration raids or government harassment.

http://www.npr.org/templates/story/story.php?storyId=125380052

Exploratory analysis to inference

* 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 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.

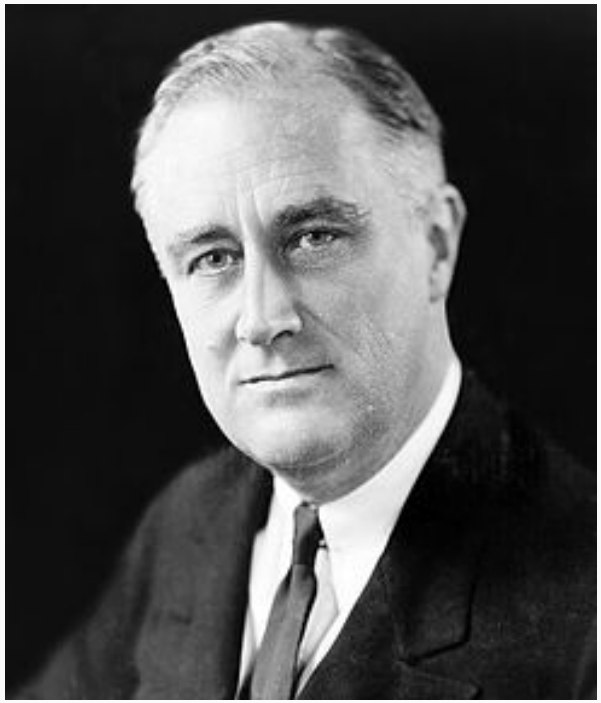
cnn.com, Jan 14, 2012

* Convenience sample: Individuals who are easily accessible are more likely to be included in the sample.

Sampling bias example: Landon vs. FDR

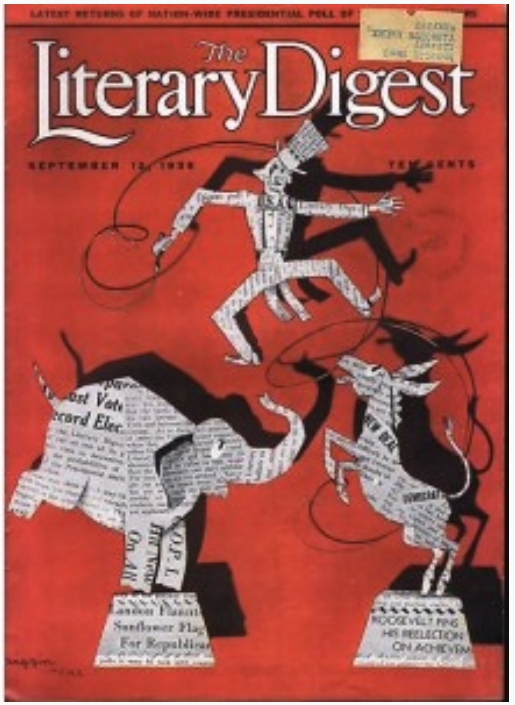
A historical example of a biased sample yielding misleading results:

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 preferable, 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?

* Some of the mailings may have never reached the parents.
* The school district has strong support from parents to move forward with the policy approval.
* It is possible that majority of the parents of high school students disagree with the policy change.
* The survey results are unlikely to be biased because all parents were mailed a survey.

* Only I
* I and II
* I and III
* III and IV
* Only IV

Explanatory and response variables

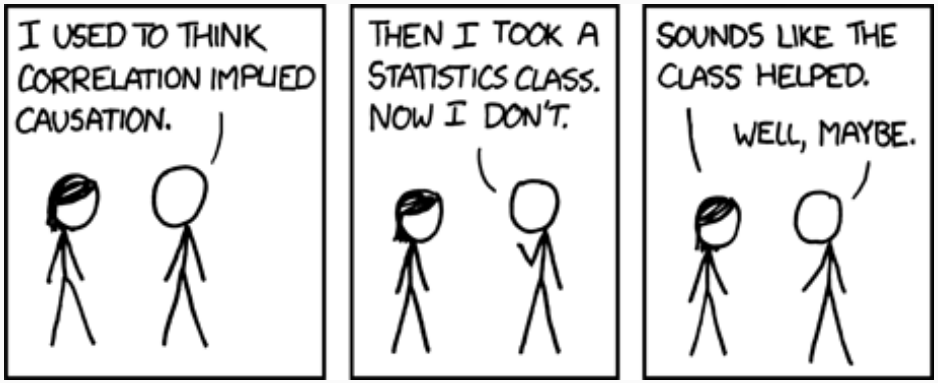
* To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other:

$$\text{explanatory variable} \xrightarrow{\text{might affect}} \text{response variable}$$

* Labeling variables as explanatory and response does not guarantee the relationship between the two is actually causal, even if there is an association identified between the two variables. We use these labels only to keep track of which variable we suspect affects the other.

Observational studies and experiments

* Observational study: Researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely "observe", and can only establish an association between the explanatory and response variables.
* Experiment: Researchers randomly assign subjects to various treatments in order to establish causal connections between the explanatory and response variables.
* If you're going to walk away with one thing from this class, let it be "correlation does not imply causation".



http://xkcd.com/552/

Case study

Data basics

Overview of data collection principles

Observational studies and sampling strategies

- Confounding
- Sampling strategies

Experiments

Examining numerical data

Considering categorical data

Case study: Gender discrimination

New study sponsored by General Mills says that eating breakfast makes girls thinner

Study: Breakfast Helps Girls Stay Slim
I love these studies....and finding out who sponsored them!

By ALEX DOMINGUEZ, Associated Press

Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years.

Girls who ate breakfast of any type had a lower average body mass index, a common obesity gauge, than those who said they didn't. The index was even lower for girls who said they ate cereal for breakfast, according to findings of the study conducted by the Maryland Medical Research Institute. The study received funding from the National Institutes of Health and cereal-maker General Mills.

"Not eating breakfast is the worst thing you can do, that's really the take-home message for teenage girls," said study author Bruce Barton, the Maryland institute's president and CEO.

The fiber in cereal and healthier foods that normally accompany cereal, such as milk and orange juice, may account for the lower body mass index among cereal eaters, Barton said.

The results were gleaned from a larger NIH survey of 2,379 girls in California, Ohio and Maryland who were tracked between ages 9 and 19. Results of the study appear in the September issue of the Journal of the American Dietetic Association.

Nearly one in three adolescent girls in the United States is overweight, according to the association. The problem is particularly troubling because research shows becoming overweight as a child can lead to a lifetime struggle with obesity.

As part of the survey, the girls were asked once a year what they had eaten during the previous three days. The data were adjusted to compensate for factors such as differences in physical activity among the girls and normal increases in body fat during adolescence.

http://www.peertrainer.com/LoungeCommunityThread.aspx?ForumID=1&ThreadID=3118

What type of study is this, observational study or an experiment? "Girls who regularly ate breakfast, particularly one that includes cereal, were slimmer than those who skipped the morning meal, according to a study that tracked nearly 2,400 girls for 10 years. [...] As part of the survey, the girls were asked once a year what they had eaten during the previous three days."

This is an observational study since the researchers merely observed the behavior of the girls (subjects) as opposed to imposing treatments on them.

What is the conclusion of the study?

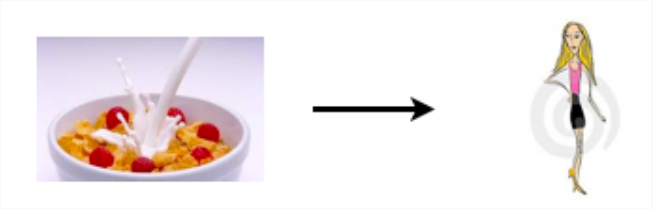
There is an association between girls eating breakfast and being slimmer.

Who sponsored the study?

General Mills.

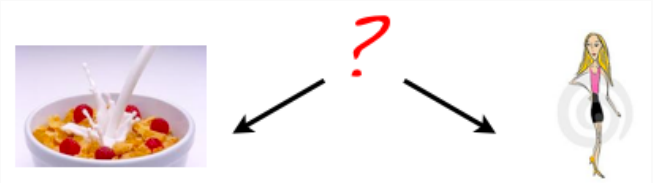
3 POSSIBLE EXPLANATIONS

* Eating breakfast causes girls to be thinner.



* Being thin causes girls to eat breakfast.

* A third variable is responsible for both. What could it be?
  An extraneous variable that affects both the explanatory and the response variable and that make it seem like there is a relationship between the two are called confounding variables.



Images from: http://www.appforhealth.com/wp-content/uploads/2011/08/ipn-cerealfrijo-300x135.jpg,

http://www.dreamstime.com/stock-photography-too-thin-woman-anorexia-model-image2814892.

Prospective vs. retrospective studies

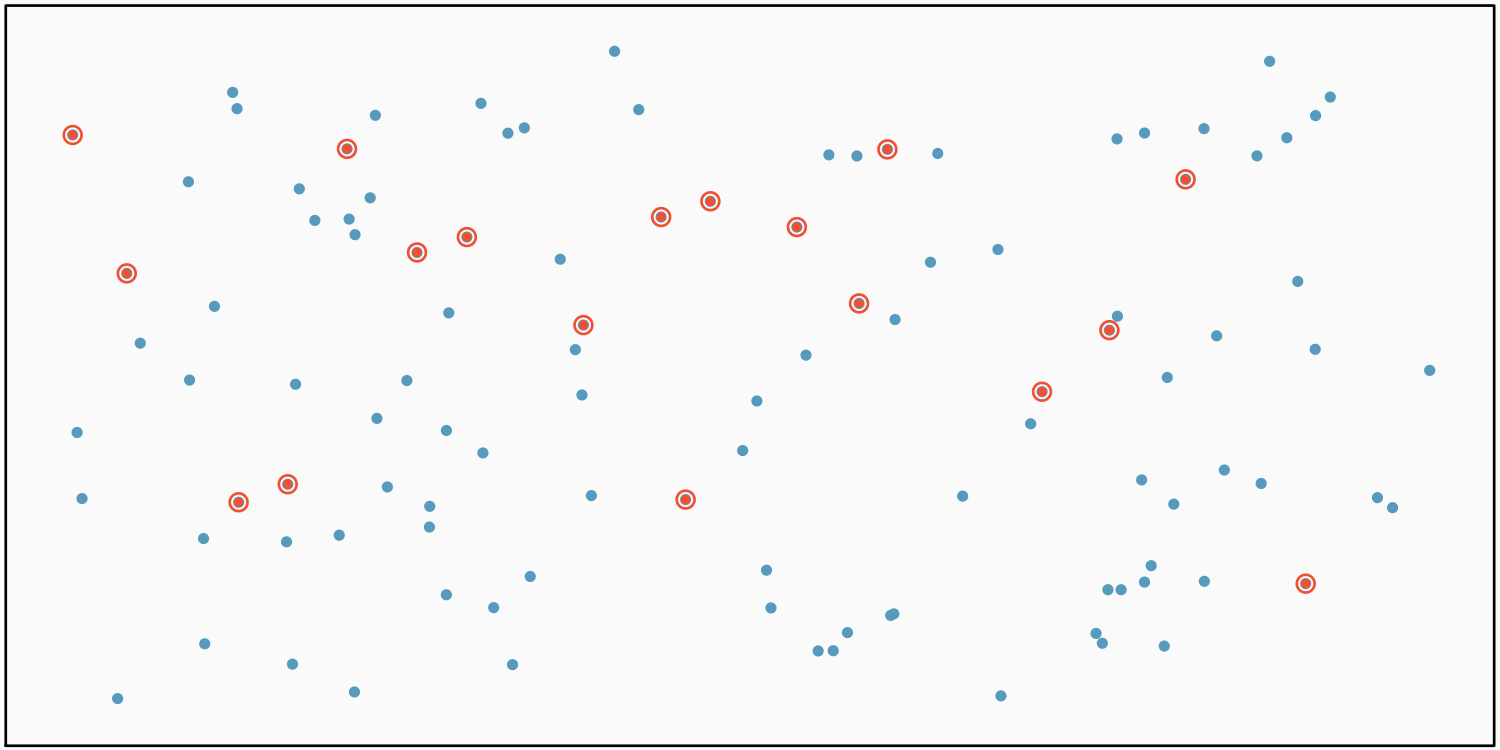
* A prospective study identifies individuals and collects information as events unfold.
  * Example: The Nurses Health Study has been recruiting registered nurses and then collecting data from them using questionnaires since 1976.
* Retrospective studies collect data after events have taken place.
  * Example: Researchers reviewing past events in medical records.

Obtaining good samples

* Almost all statistical methods are based on the notion of implied randomness.
* If observational data are not collected in a random framework from a population, these statistical methods – the estimates and errors associated with the estimates – are not reliable.
* Most commonly used random sampling techniques are simple, stratified, and cluster sampling.

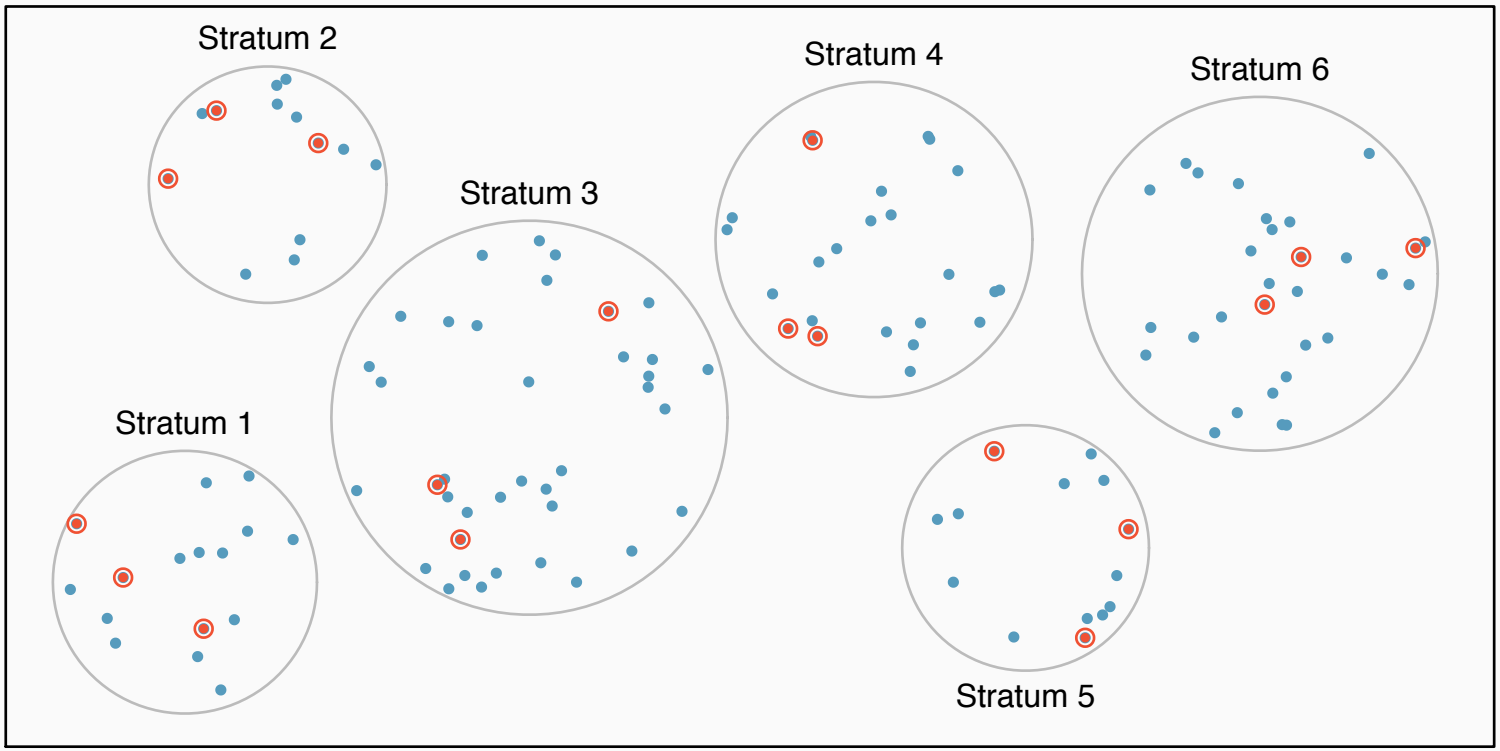
Simple random sample

Randomly select cases from the population, where there is no implied connection between the points that are selected.



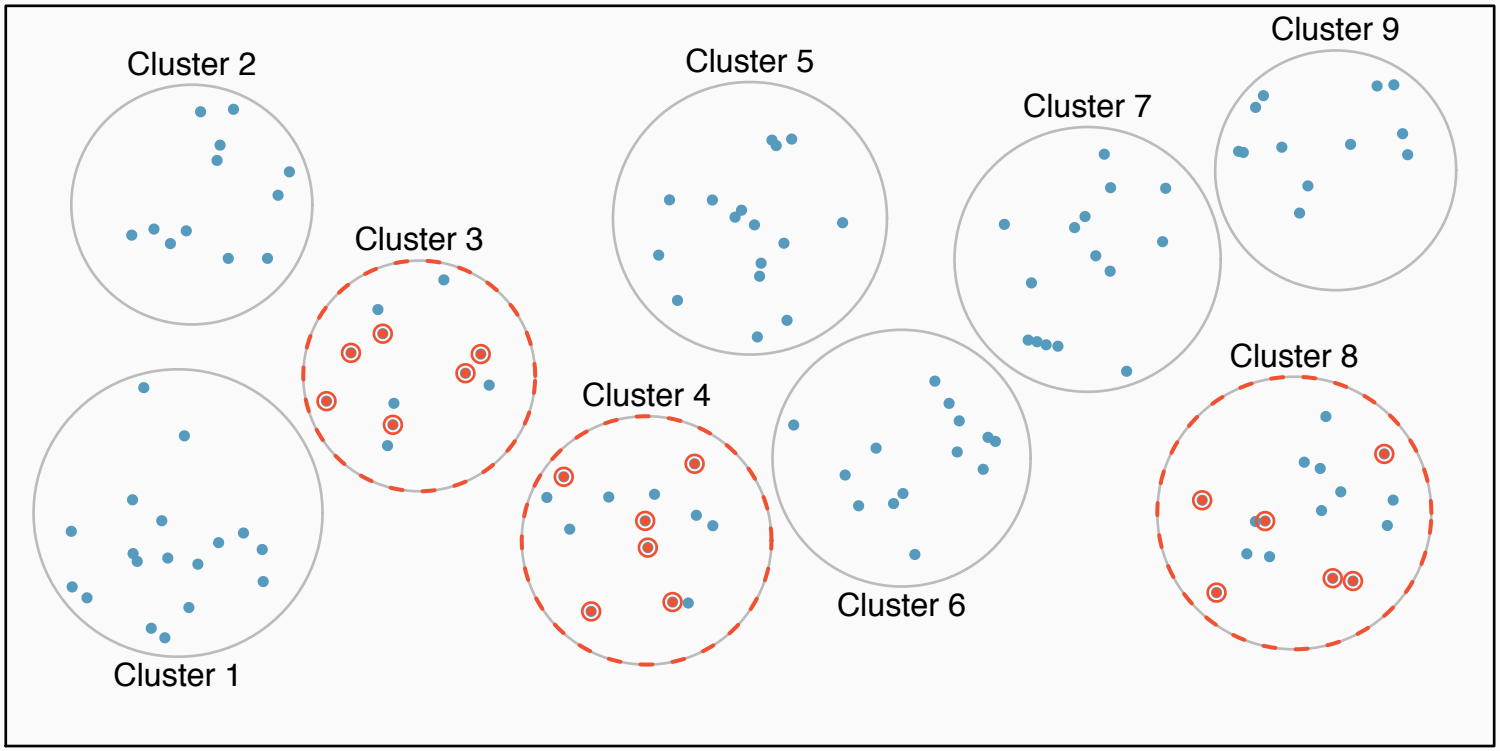
Stratified sample

Strata are made up of similar observations. We take a simple random sample from each stratum.



Cluster sample

Clusters are usually not made up of homogeneous observations, and we take a simple random sample from a random sample of clusters. Usually preferred for economical reasons.



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?

* Simple random sampling
* Cluster sampling
* Stratified sampling
* Blocked sampling

Case study

Data basics

Overview of data collection principles

Observational studies and sampling strategies

Experiments

Examining numerical data

Considering categorical data

Case study: Gender discrimination

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.

More on 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)

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 experimental design terminology...

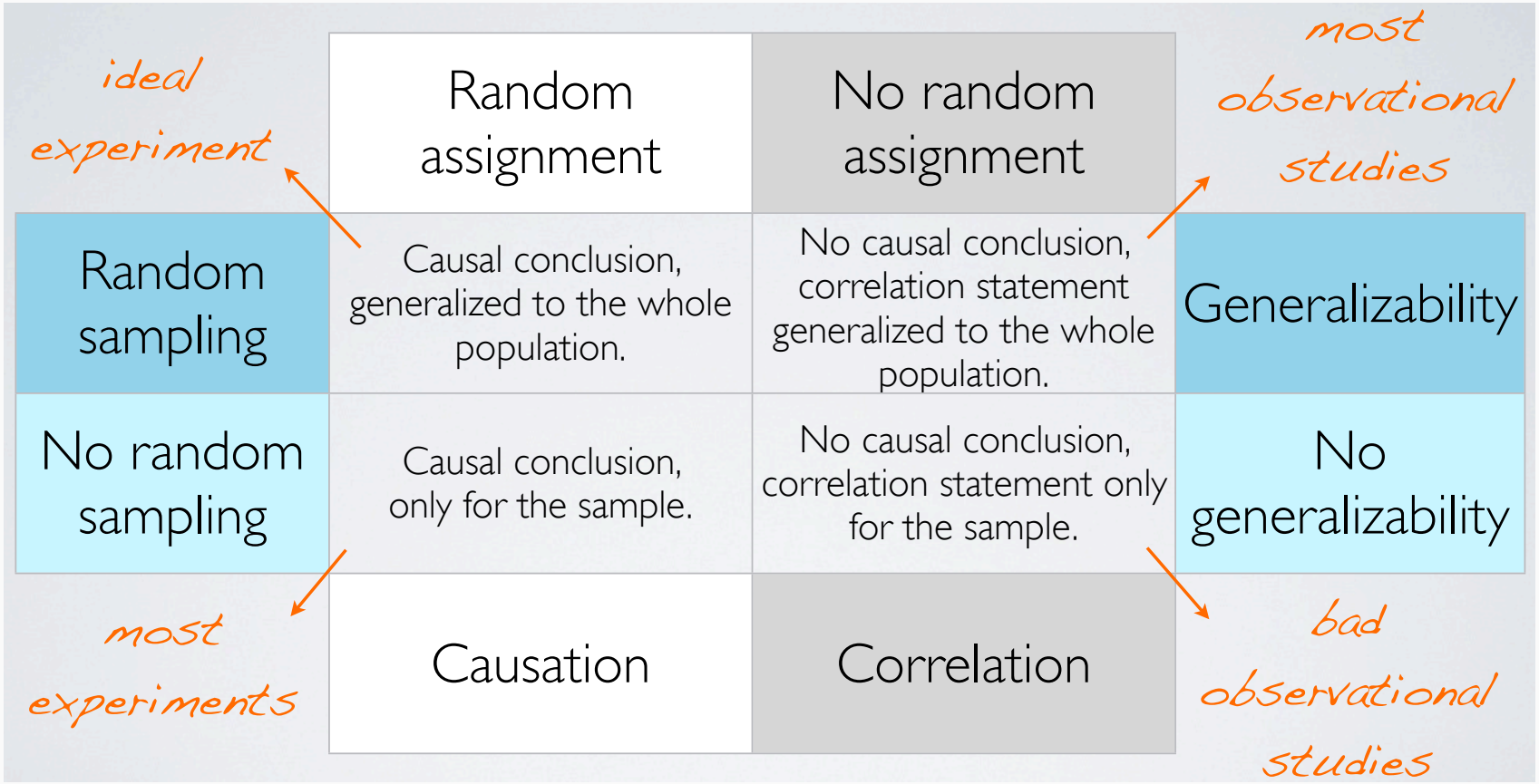
* Placebo: fake treatment, often used as the control group for medical studies
* Placebo effect: experimental units showing improvement simply because they believe they are receiving a special treatment
* Blinding: when experimental units do not know whether they are in the control or treatment group
* 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?

* Experiments take place in a lab while observational studies do not need to.
* In an observational study we only look at what happened in the past.
* Most experiments use random assignment while observational studies do not.
* Observational studies are completely useless since no causal inference can be made based on their findings.

Random assignment vs. random sampling



Case study

Data basics

Overview of data collection principles

Observational studies and sampling strategies

Experiments

Examining numerical data

- Scatterplots for paired data
- Dot plots and the mean
- Histograms and shape
- Variance and standard deviation
- Box plots, quartiles, and the median
- Robust statistics
- Transforming data
- Mapping data

Considering categorical data

Case study: Gender discrimination

Jelena Bradic http://www.jelenabradic.net Assistant Professor, Department of Mathematics, University of California, San Diego jbradic@ucsd.edu

Scatterplot

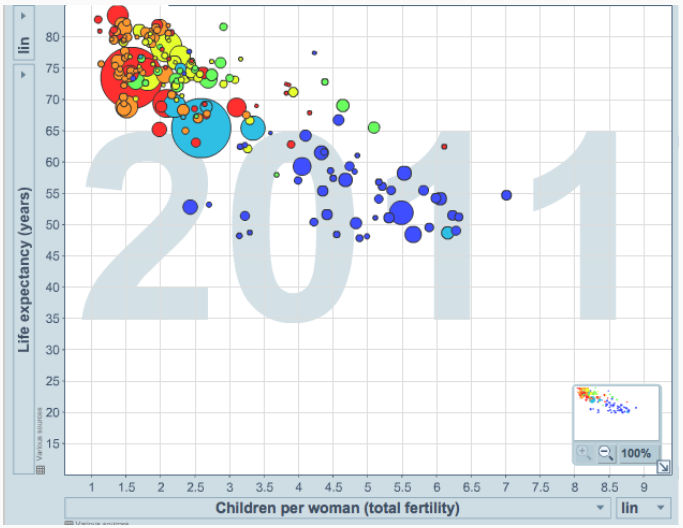
Scatterplots are useful for visualizing the relationship between two numerical variables.

Do life expectancy and total fertility appear to be associated or independent?

They appear to be linearly and negatively associated: as fertility increases, life expectancy decreases.

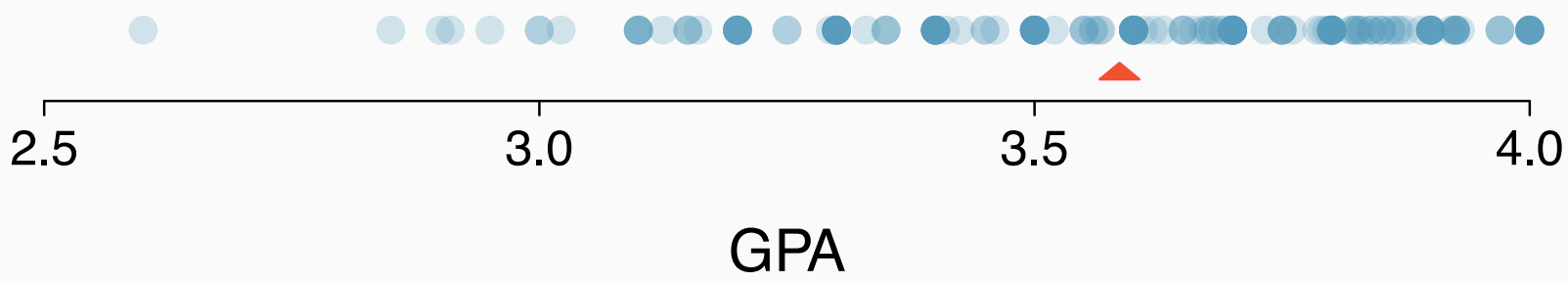
Was the relationship the same throughout the years, or did it change?

The relationship changed over the years.



http://www.gapminder.org/world

Dot plots & mean



* The mean, also called the average (marked with a triangle in the above plot), is one way to measure the center of a distribution of data.
* The mean GPA is 3.59.

Mean

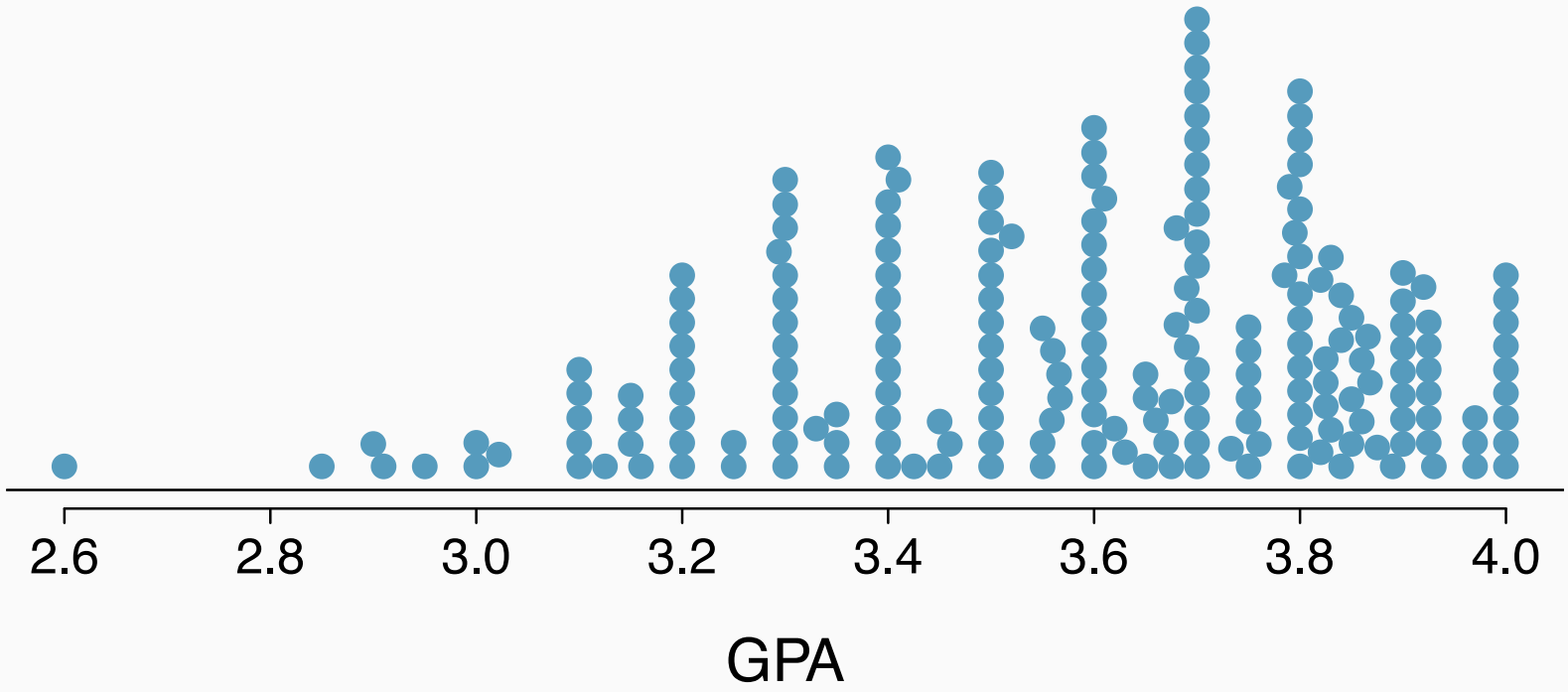
* The sample mean, denoted as $\bar{x}$, can be calculated as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

  where $x_1, x_2, \cdots, x_n$ represent the n observed values.
* The population mean is also computed the same way but is denoted as μ. It is often not possible to calculate μ since population data are rarely available.
* The sample mean is a sample statistic, and serves as a point estimate of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population), it is usually a pretty good estimate.

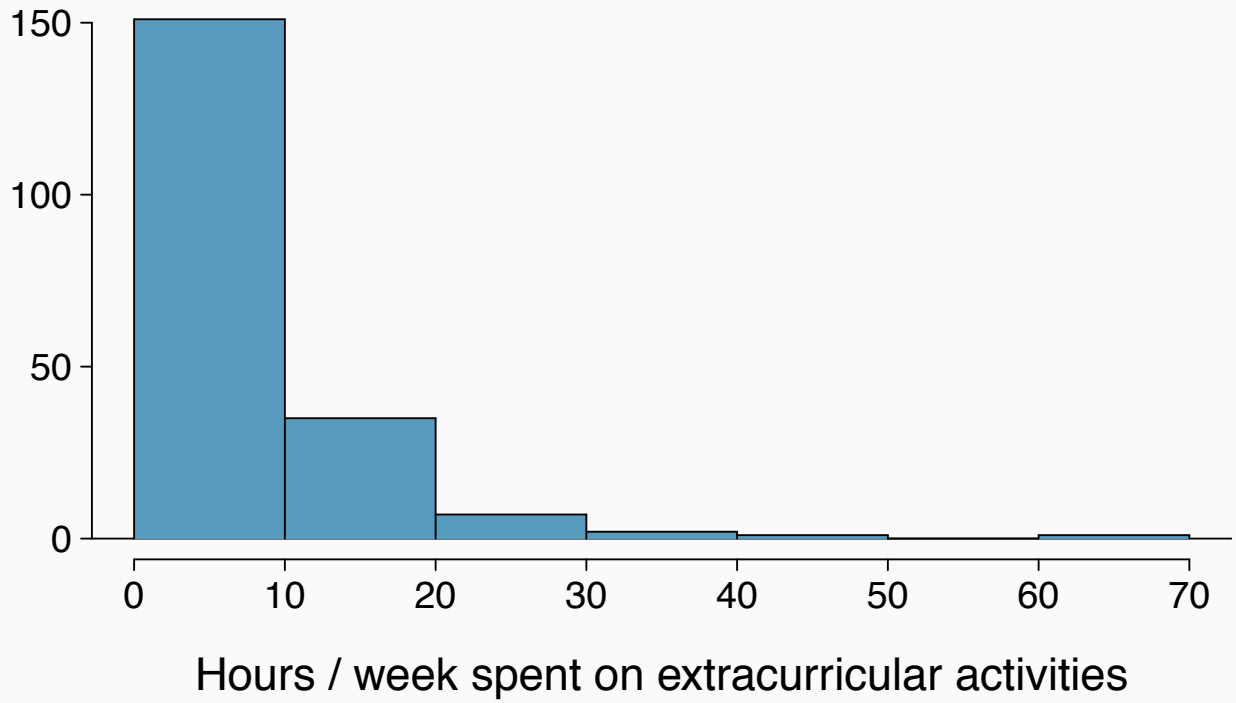
Stacked dot plot

Higher bars represent areas where there are more observations, makes it a little easier to judge the center and the shape of the distribution.



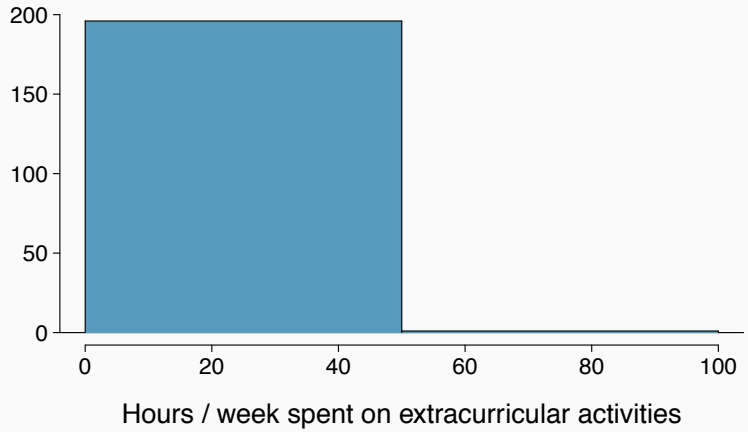
Histograms - Extracurricular hours

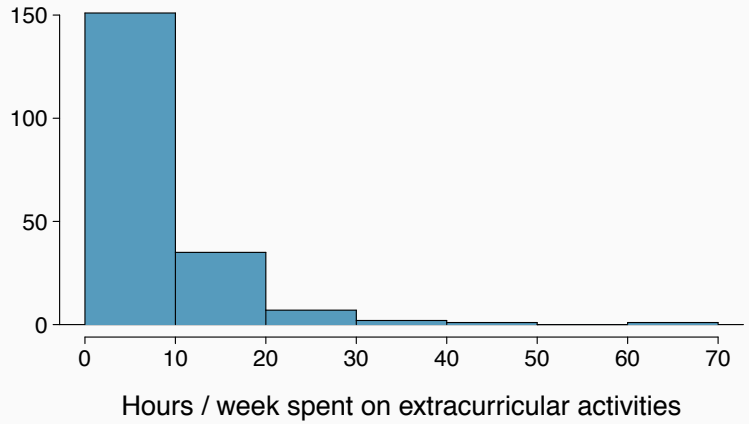
* Histograms provide a view of the data density. Higher bars represent where the data are relatively more common.
* Histograms are especially convenient for describing the shape of the data distribution.
* The chosen bin width can alter the story the histogram is telling.

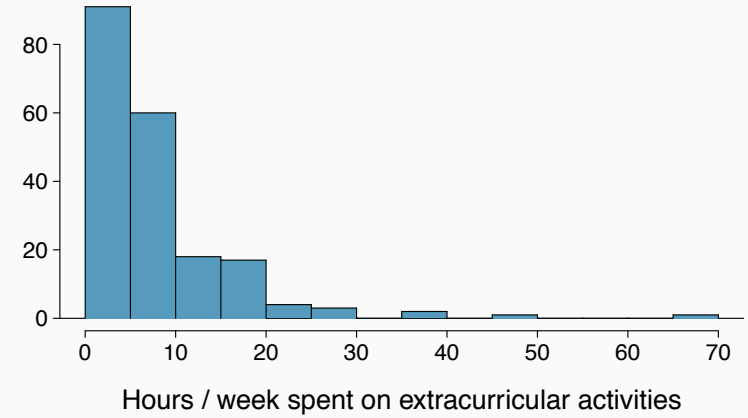


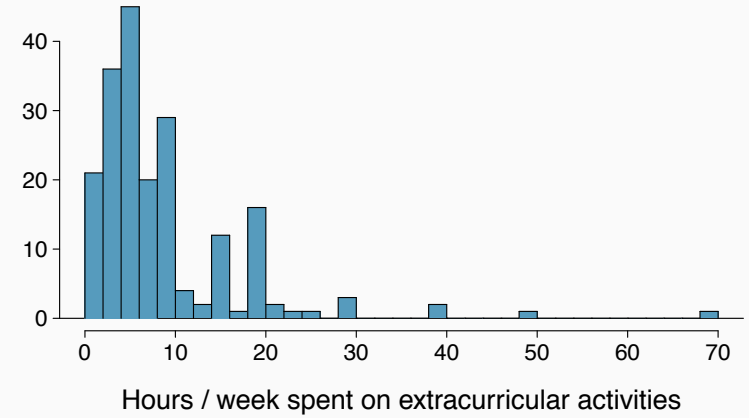
Bin width

Which one(s) of these histograms are useful? Which reveal too much about the data? Which hide too much?



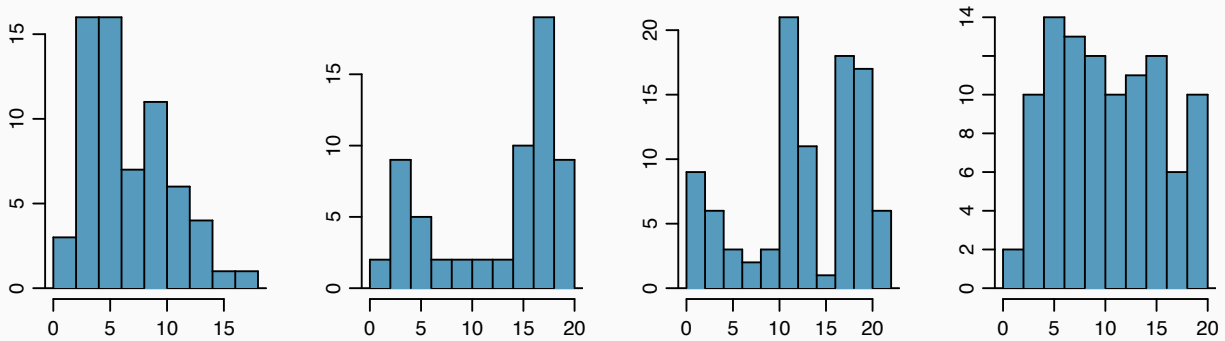






Shape of a distribution: modality

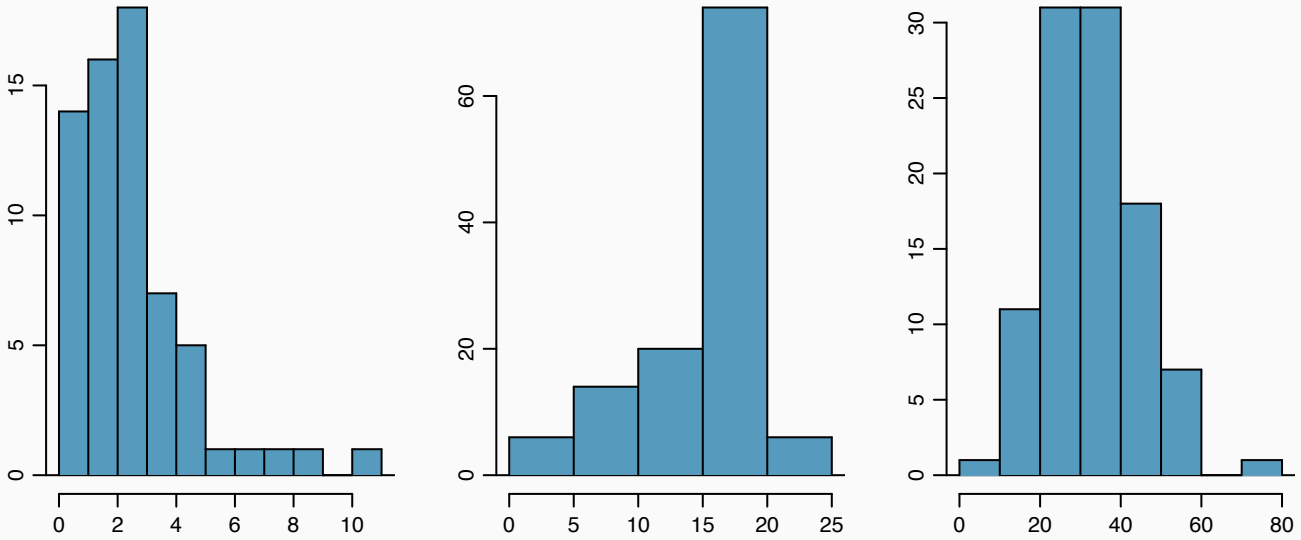
Does the histogram have a single prominent peak (unimodal), several prominent peaks (bimodal/multimodal), or no apparent peaks (uniform)?



Note: In order to determine modality, step back and imagine a smooth curve over the histogram – imagine that the bars are wooden blocks and you drop a limp spaghetti over them, the shape the spaghetti would take could be viewed as a smooth curve.

Shape of a distribution: skewness

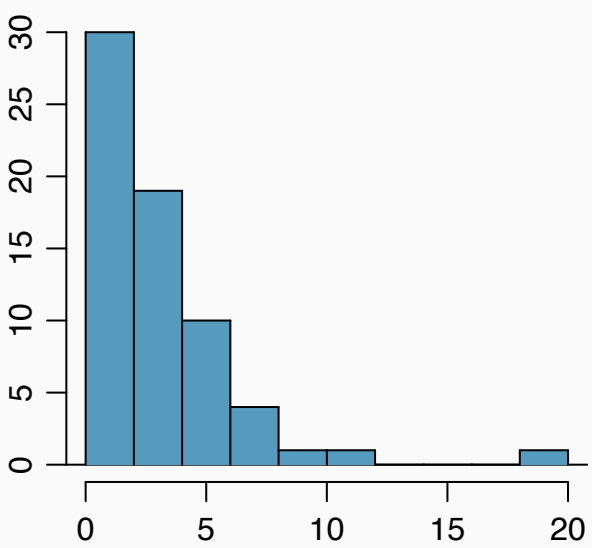
Is the histogram right skewed, left skewed, or symmetric?

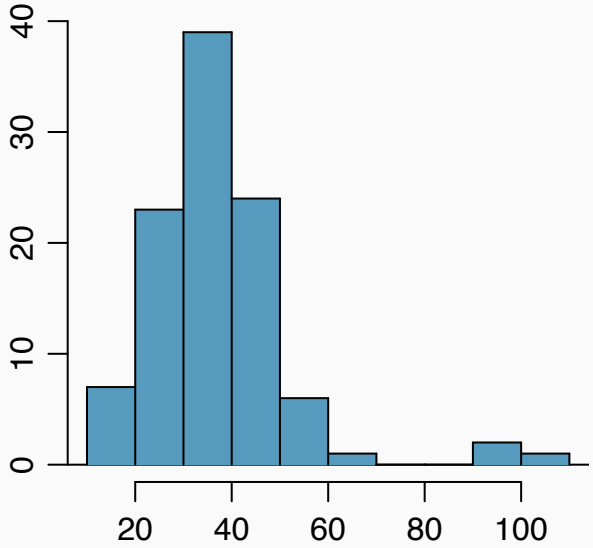


Note: Histograms are said to be skewed to the side of the long tail.

Shape of a distribution: unusual observations

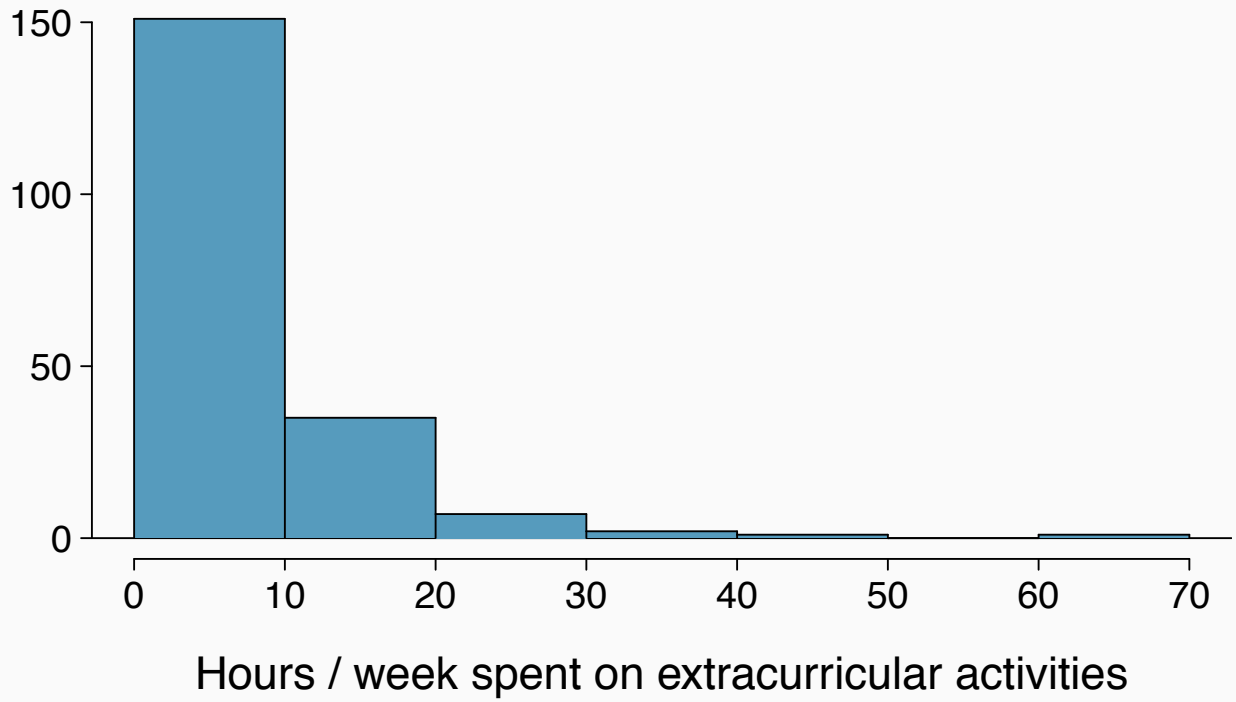
Are there any unusual observations or potential outliers?





Extracurricular activities

How would you describe the shape of the distribution of hours per week students spend on extracurricular activities?

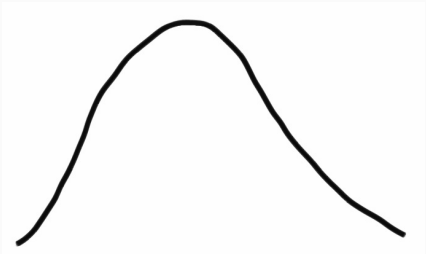


Unimodal and right skewed, with a potentially unusual observation at 60 hours/week.

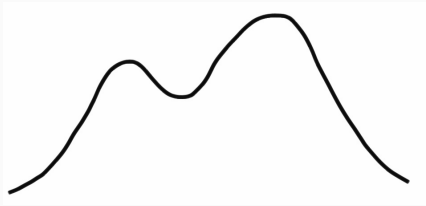
Commonly observed shapes of distributions

* modality

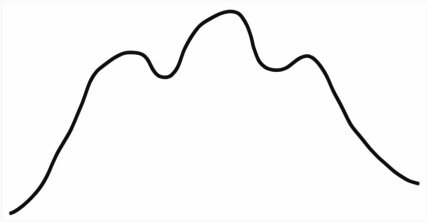
unimodal



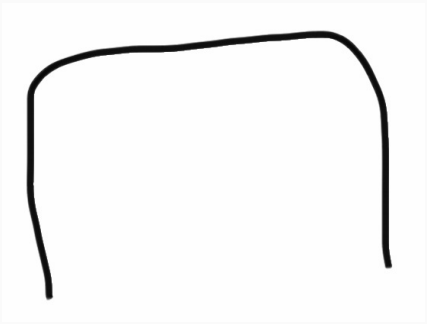
bimodal



multimodal

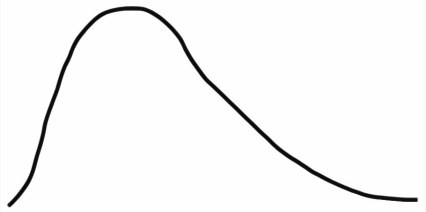


uniform

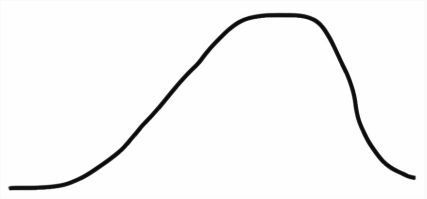


* skewness

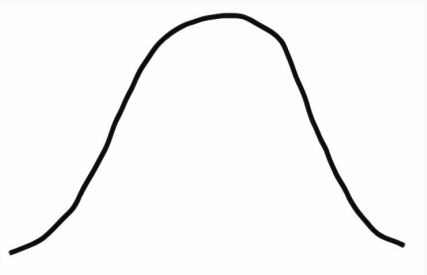
right skew



left skew



symmetric



Practice

Which of these variables do you expect to be uniformly distributed?

* weights of adult females
* salaries of a random sample of people from North Carolina
* house prices
* birthdays of classmates (day of the month)

Application activity: Shapes of distributions

Sketch the expected distributions of the following variables:

* number of piercings
* scores on an exam
* IQ scores

Come up with a concise way (1-2 sentences) to teach someone how to determine the expected distribution of any variable.

Are you typical?

http://www.youtube.com/watch?v=4B2xOvKFFz4

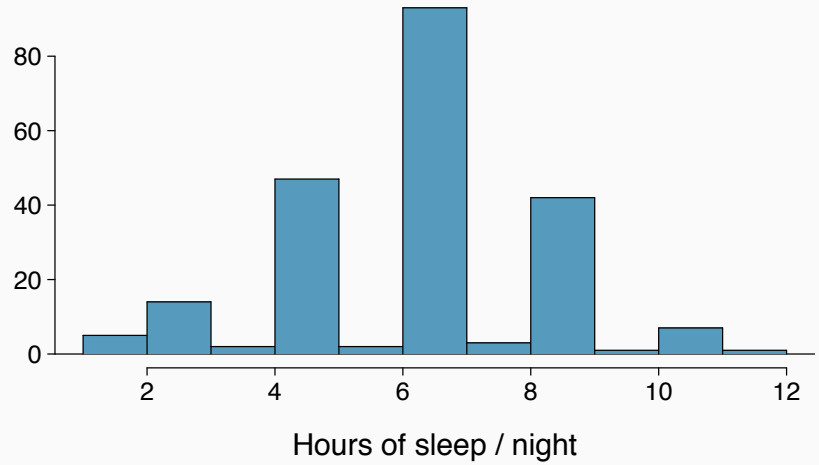
How useful are centers alone for conveying the true characteristics of a distribution?

Variance

Variance is roughly the average squared deviation from the mean.

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

* The sample mean is $\bar{x} = 6.71$, and the sample size is $n = 217$.
* The variance of amount of sleep students get per night can be calculated as:



$$s^2 = \frac{(5-6.71)^2 + (9-6.71)^2 + \cdots + (7-6.71)^2}{217-1} = 4.11 \text{ hours}^2$$

Variance (cont.)

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.

Standard deviation

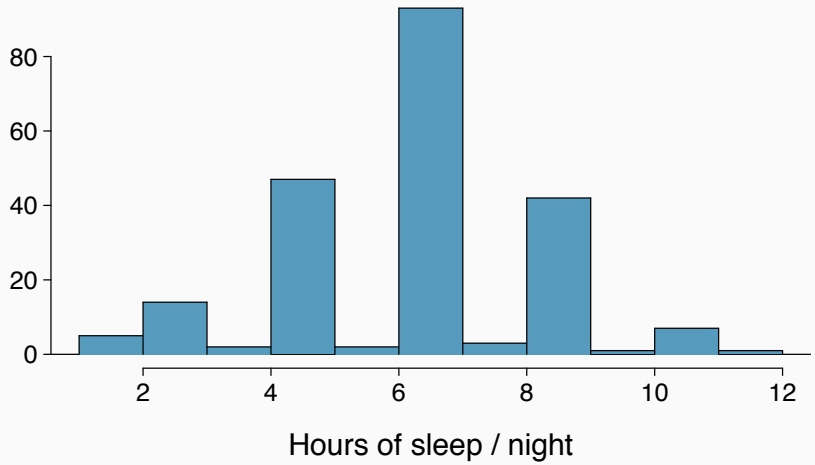
The standard deviation is the square root of the variance, and has the same units as the data.s

$$s = \sqrt{s^2}$$

* The standard deviation of amount of sleep students get per night can be calculated as:

$$s = \sqrt{4.11} = 2.03 \text{ hours}$$

* We can see that all of the data are within 3 standard deviations of the mean.



Median

* The median is the value that splits the data in half when ordered in ascending order.

$$0, 1, 2, 3, 4$$

* If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

* Since the median is the midpoint of the data, 50% of the values are below it. Hence, it is also the 50^{th} percentile.

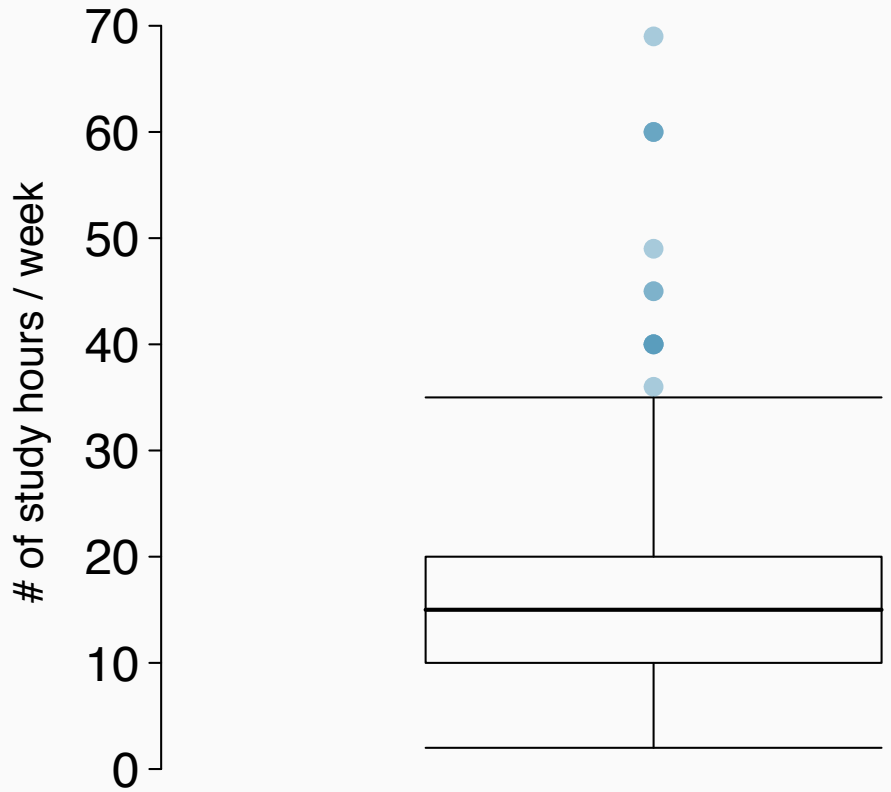
Q1, Q3, and IQR

* The 25^{th} percentile is also called the first quartile, Q1.
* The 50^{th} percentile is also called the median.
* The 75^{th} percentile is also called the third quartile, Q3.
* Between Q1 and Q3 is the middle 50% of the data. The range these data span is called the interquartile range, or the IQR.

$$\text{IQR} = \text{Q3} - \text{Q1}$$

Box plot

The box in a box plot represents the middle 50% of the data, and the thick line in the box is the median.



Anatomy of a box plot

Whiskers and outliers

* Whiskers of a box plot can extend up to 1.5 × IQR away from the quartiles.

$$\text{max upper whisker reach} = \text{Q3} + 1.5 \times \text{IQR}$$

$$\text{max lower whisker reach} = \text{Q1} - 1.5 \times \text{IQR}$$

$$\text{IQR} : 20 - 10 = 10$$

$$\text{max upper whisker reach} = 20 + 1.5 \times 10 = 35$$

$$\text{max lower whisker reach} = 10 - 1.5 \times 10 = -5$$

* A potential outlier is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

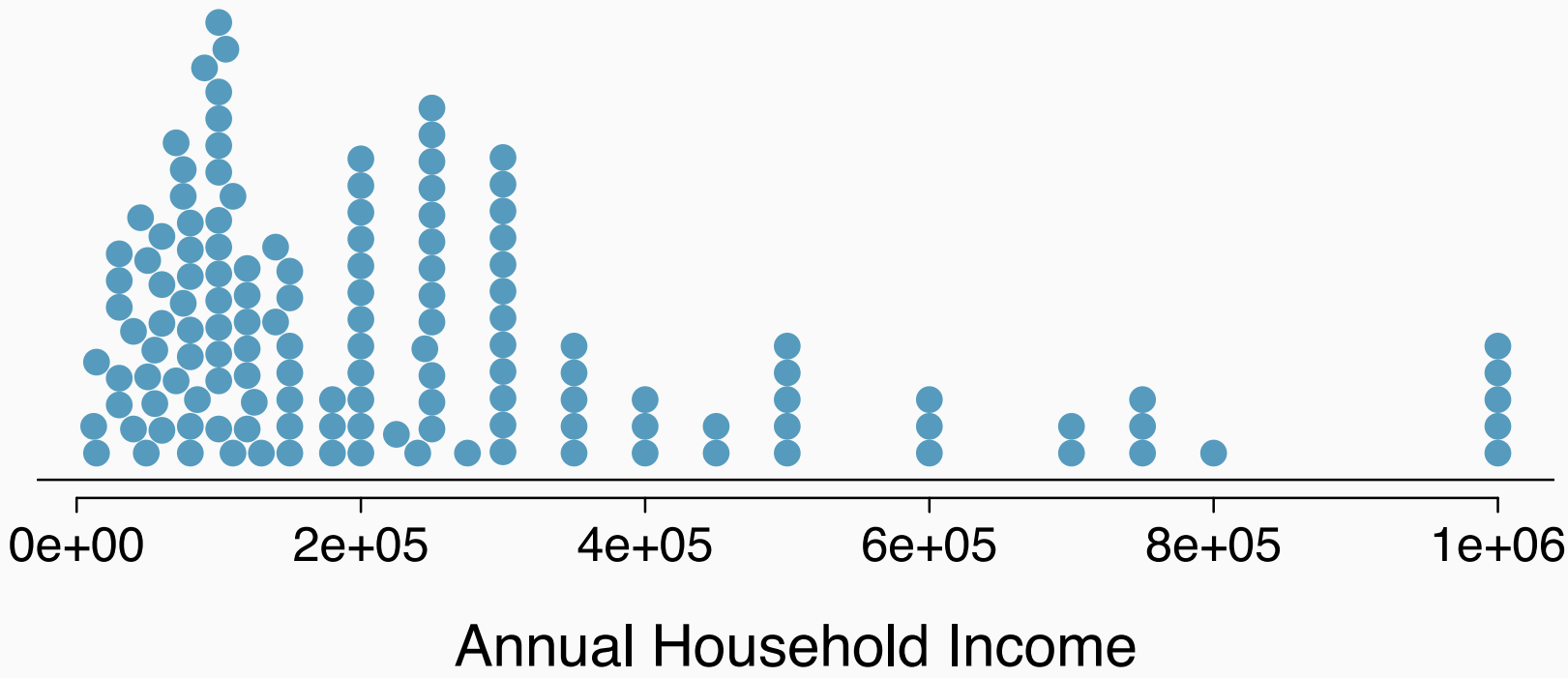
Outliers (cont.)

Why is it important to look for outliers?

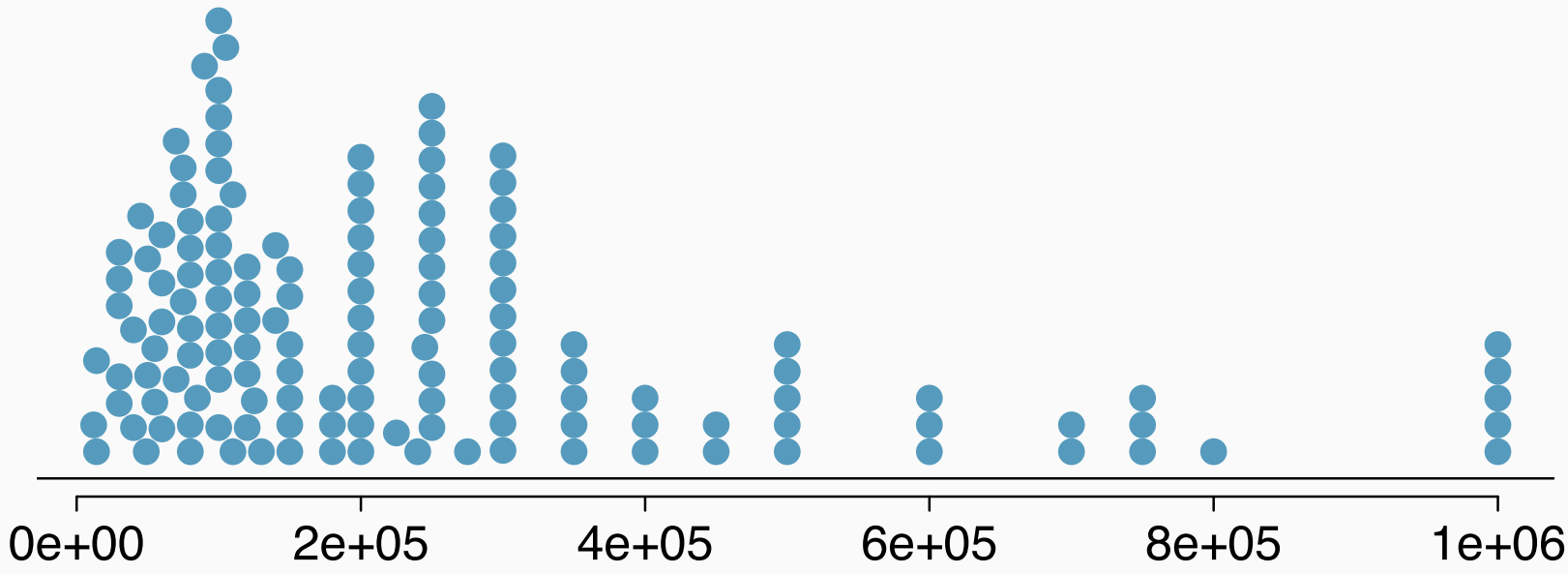
* Identify extreme skew in the distribution.
* Identify data collection and entry errors.
* Provide insight into interesting features of the data.

Extreme observations

How would sample statistics such as mean, median, SD, and IQR of household income be affected if the largest value was replaced with $10 million? What if the smallest value was replaced with $10 million?



Robust statistics



	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original data	190K	200K	245K	226K
move largest to $10 million	190K	200K	309K	853K
move smallest to $10 million	200K	200K	316K	854K

Robust statistics

Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

* for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
* for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

Robust statistics

Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

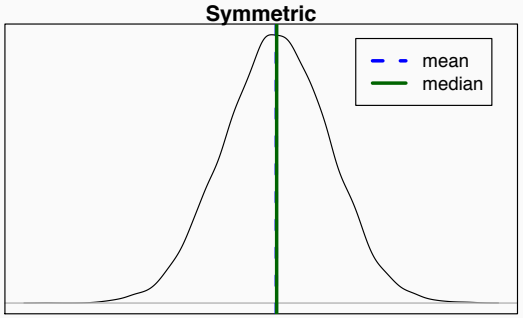
* for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
* for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

If you would like to estimate the typical household income for a student, would you be more interested in the mean or median income?

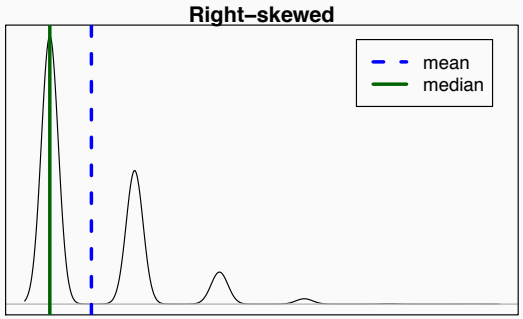
Median

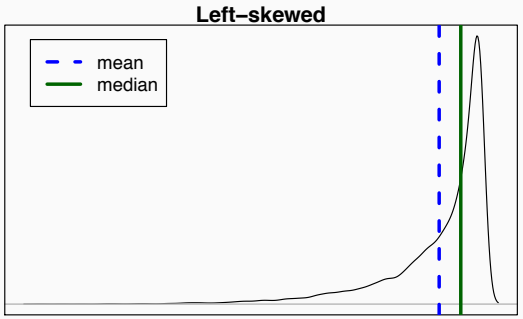
Mean vs. median

* If the distribution is symmetric, center is often defined as the mean: mean $\approx$ median



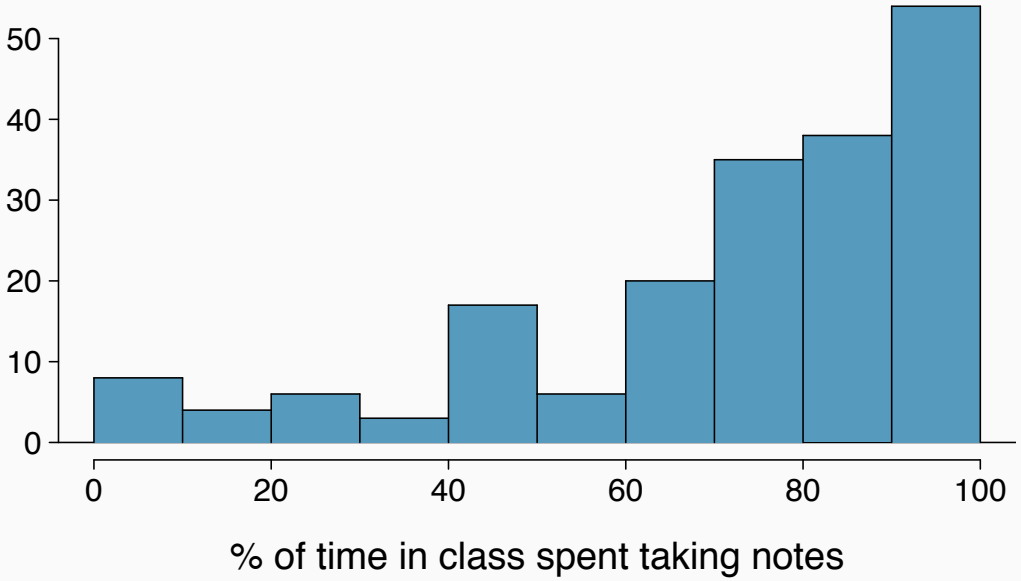
* If the distribution is skewed or has extreme outliers, center is often defined as the median
  * Right-skewed: mean $>$ median
  * Left-skewed: mean $<$ median





Practice

Which is most likely true for the distribution of percentage of time actually spent taking notes in class versus on Facebook, Twitter, etc.?



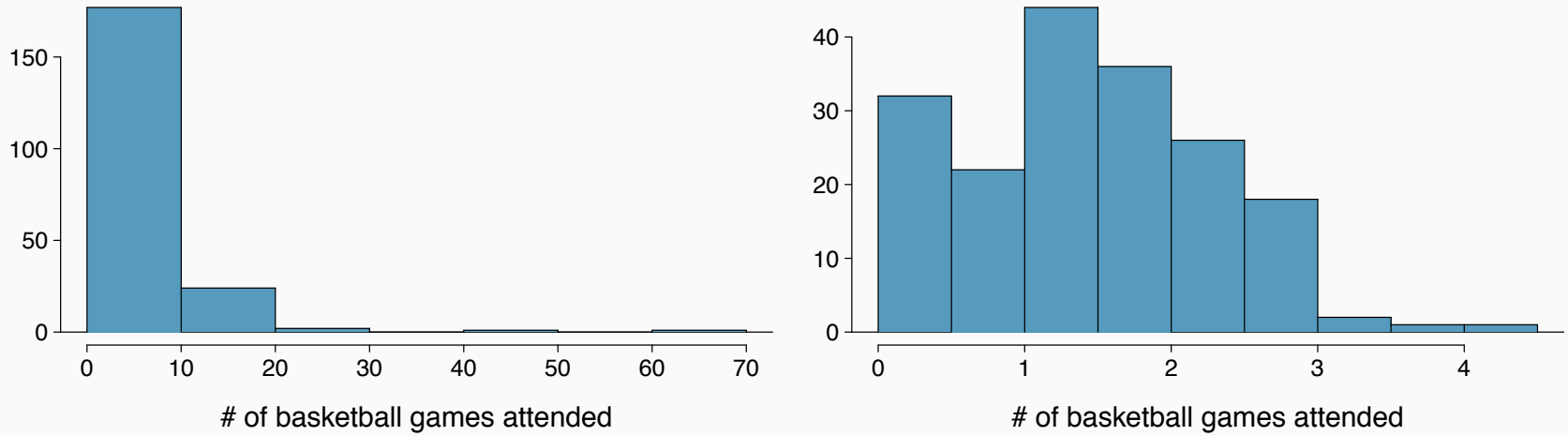
median: 80%
mean: 76%

* mean$>$ median
* mean $<$ median
* mean $\approx$ median
* impossible to tell

Extremely skewed data

When data are extremely skewed, transforming them might make modeling easier. A common transformation is the log transformation.

The histograms on the left shows the distribution of number of basketball games attended by students. The histogram on the right shows the distribution of log of number of games attended.



Pros and cons of transformations

* Skewed data are easier to model with when they are transformed because outliers tend to become far less prominent after an appropriate transformation.

# of games	70	50	25	$\cdots$
log(# of games)	4.25	3.91	3.22	$\cdots$

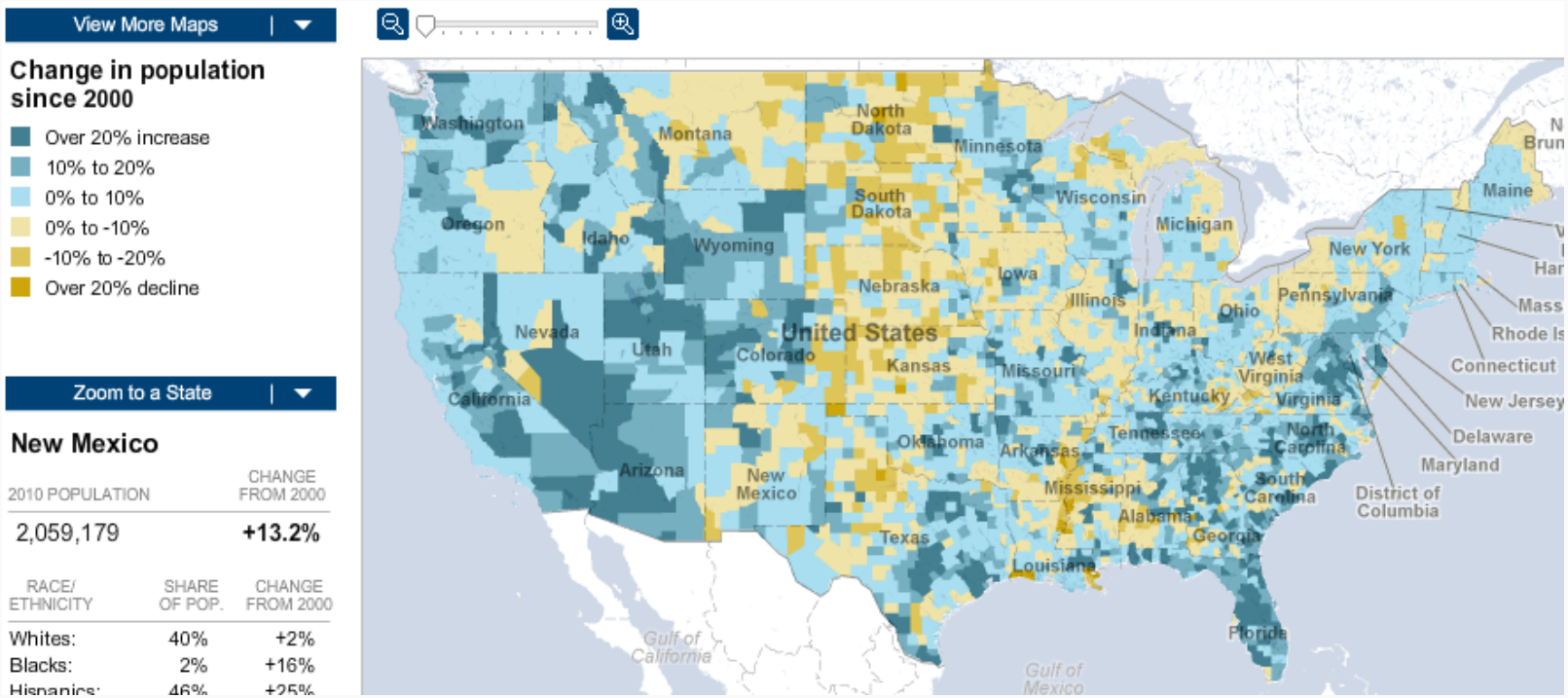
* However, results of an analysis might be difficult to interpret because the log of a measured variable is usually meaningless.

What other variables would you expect to be extremely skewed?

Salary, housing prices, etc.

Intensity maps

What patterns are apparent in the change in population between 2000 and 2010?



http://projects.nytimes.com/census/2010/map

Case study

Data basics

Overview of data collection principles

Observational studies and sampling strategies

Experiments

Examining numerical data

Considering categorical data

- Contingency tables and bar plots
- Row and column proportions
- Segmented bar and mosaic plots
- Pie charts
- Comparing numerical data across groups

Case study: Gender discrimination

Contingency tables

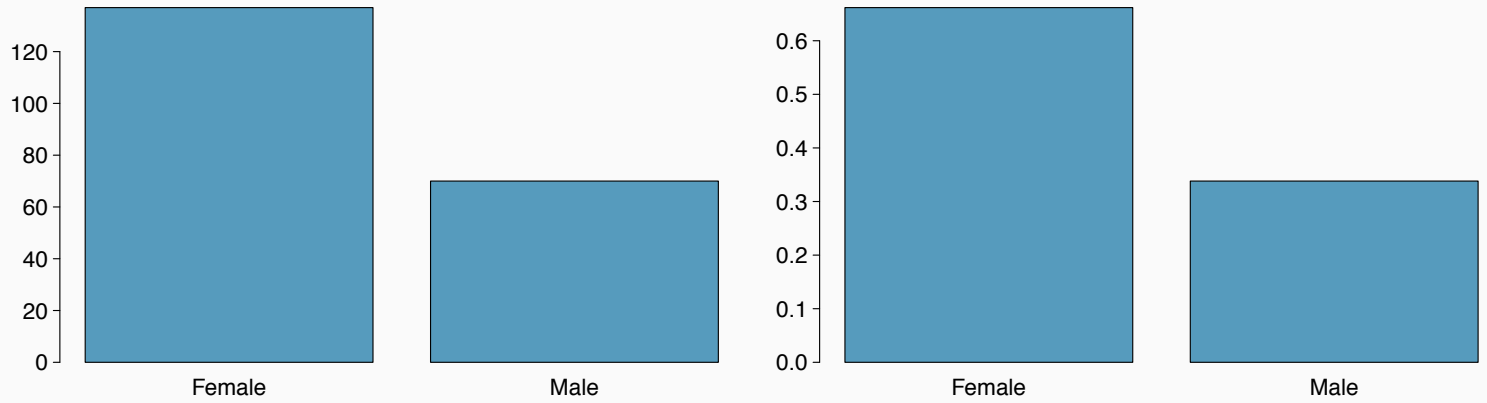
A table that summarizes data for two categorical variables is called a contingency table.

The contingency table below shows the distribution of students' genders and whether or not they are looking for a spouse while in college.

		looking for spouse		
		No	Yes	Total
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

Bar plots

A bar plot is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a relative frequency bar plot.



How are bar plots different than histograms?

Bar plots are used for displaying distributions of categorical variables, while histograms are used for numerical variables. The x-axis in a histogram is a number line, hence the order of the bars cannot be changed, while in a bar plot the categories can be listed in any order (though some orderings make more sense than others, especially for ordinal variables.)

Choosing the appropriate proportion

Does there appear to be a relationship between gender and whether the student is looking for a spouse in college?

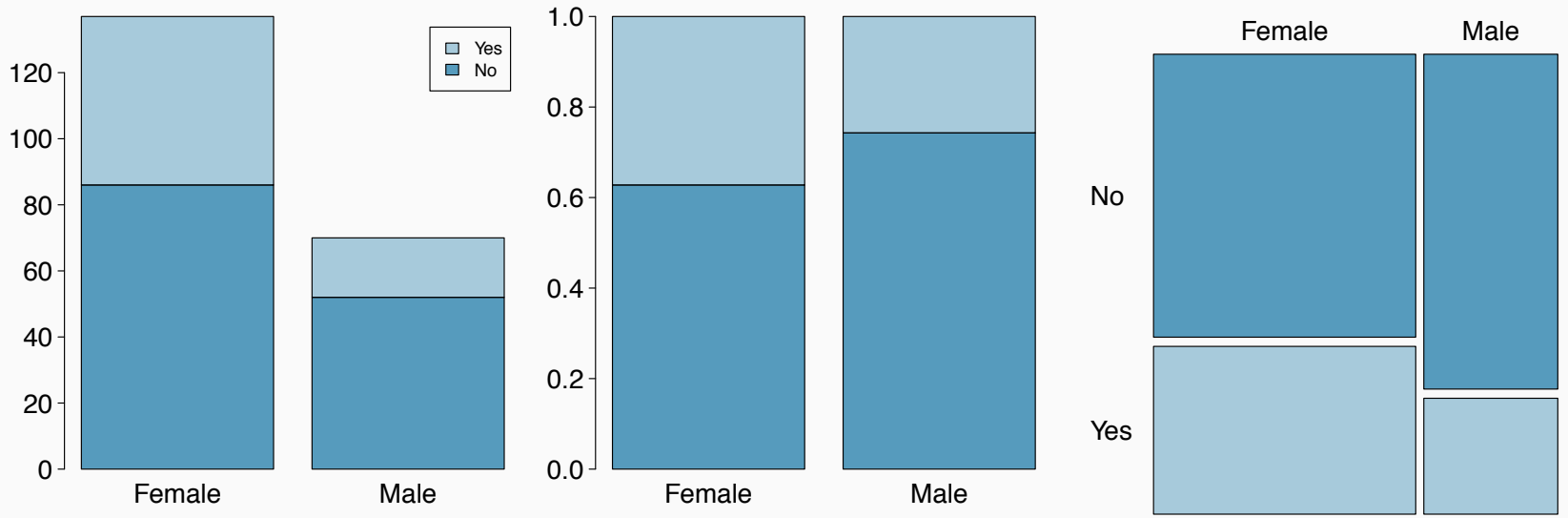
		looking for spouse		
		No	Yes	Total
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

To answer this question we examine the row proportions:

* % Females looking for a spouse: $51/137 \approx 0.37$
* % Males looking for a spouse: $18/70 \approx 0.26$

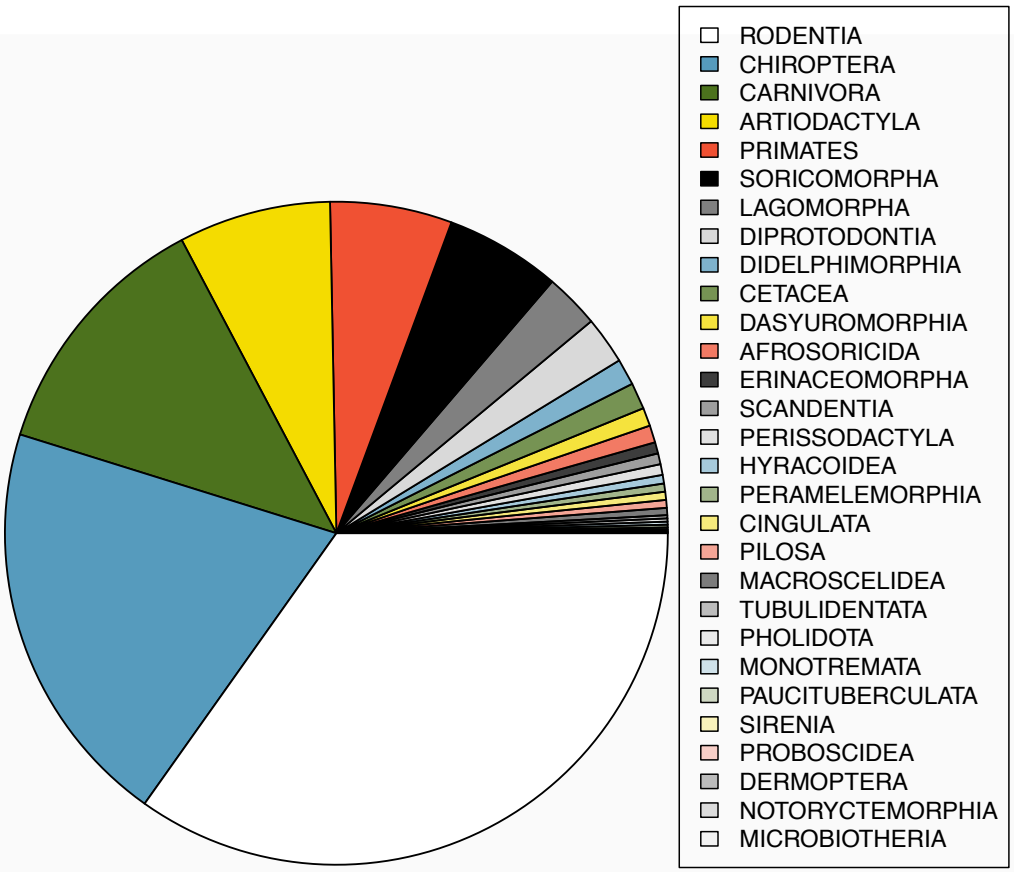
Segmented bar and mosaic plots

What are the differences between the three visualizations shown below?



Pie charts

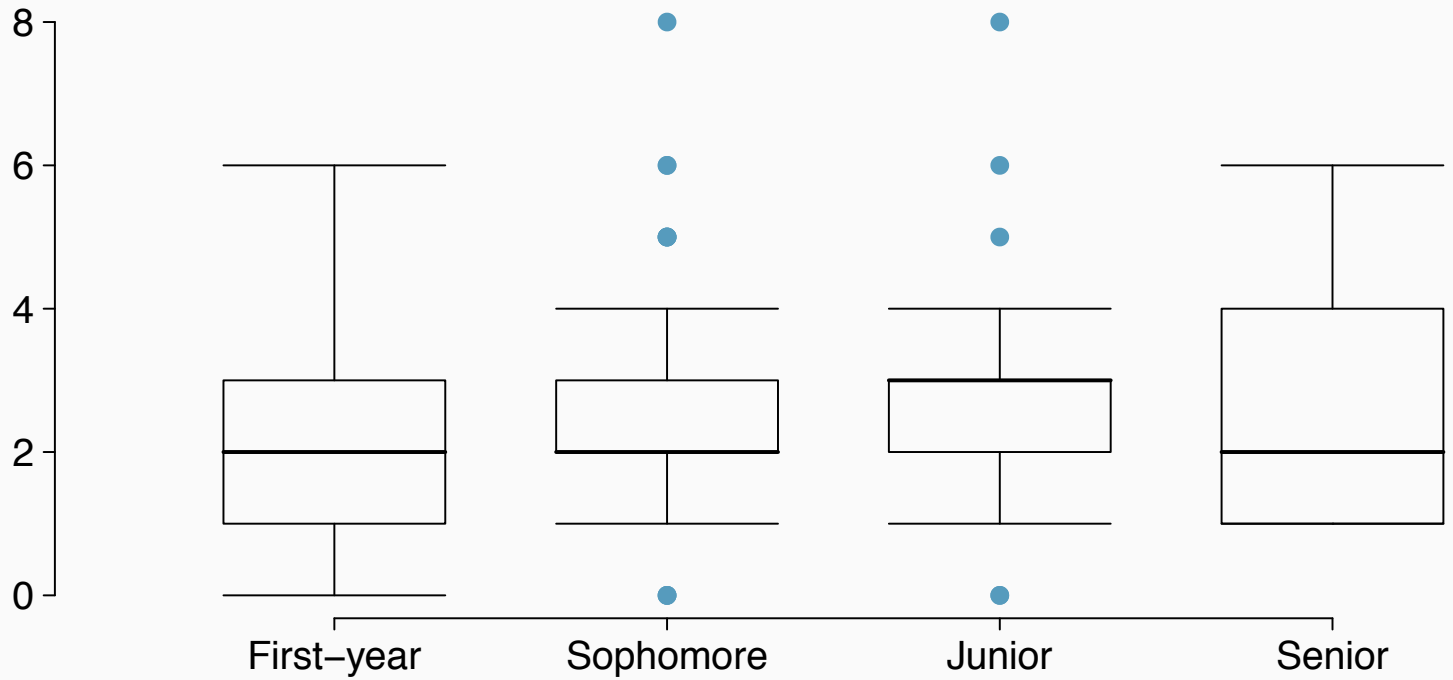
Can you tell which order encompasses the lowest percentage of mammal species?



Data from http://www.bucknell.edu/msw3.

Side-by-side box plots

Does there appear to be a relationship between class year and number of clubs students are in?



Case study

Data basics

Overview of data collection principles

Observational studies and sampling strategies

Experiments

Examining numerical data

Considering categorical data

Case study: Gender discrimination

- Study description and data
- Competing claims
- Testing via simulation
- Checking for independence

Jelena Bradic, http://www.jelenabradic.net, Assistant Professor, Department of Mathematics, University of California, San Diego, jbradic@ucsd.edu

Gender discrimination

* In 1972, as a part of a study on gender discrimination, 48 male bank supervisors were each given the same personnel file and asked to judge whether the person should be promoted to a branch manager job that was described as “routine”.
* The files were identical except that half of the supervisors had files showing the person was male while the other half had files showing the person was female.
* It was randomly determined which supervisors got “male” applications and which got “female” applications.
* Of the 48 files reviewed, 35 were promoted.
* The study is testing whether females are unfairly discriminated against.

Is this an observational study or an experiment?

Experiment

B.Rosen and T. Jerdee (1974), “Influence of sex role stereotypes on personnel decisions”, J.Applied Psychology, 59:9-14.

At a first glance, does there appear to be a relatonship between promotion and gender?

		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

% of males promoted: $21/24 = 0.875$

% of females promoted: $14/24 = 0.583$

Practice

We saw a difference of almost 30% (29.2% to be exact) between the proportion of male and female files that are promoted. Based on this information, which of the below is true?

* If we were to repeat the experiment we will definitely see that more female files get promoted. This was a fluke.
* Promotion is dependent on gender, males are more likely to be promoted, and hence there is gender discrimination against women in promotion decisions. Maybe
* The difference in the proportions of promoted male and female files is due to chance, this is not evidence of gender discrimination against women in promotion decisions. Maybe
* Women are less qualified than men, and this is why fewer females get promoted.

Two competing claims

* "There is nothing going on."
  Promotion and gender are independent, no gender discrimination, observed difference in proportions is simply due to chance. → Null hypothesis
* "There is something going on."
  Promotion and gender are dependent, there is gender discrimination, observed difference in proportions is not due to chance. → Alternative hypothesis

A trial as a hypothesis test

* Hypothesis testing is very much like a court trial.
* H_0: Defendant is innocent
  H_A: Defendant is guilty
* We then present the evidence - collect data.

  * Then we judge the evidence - "Could these data plausibly have happened by chance if the null hypothesis were true?"
    * If they were very unlikely to have occurred, then the evidence raises more than a reasonable doubt in our minds about the null hypothesis.
  * Ultimately we must make a decision. How unlikely is unlikely?

Image from http://www.nwherald.com/_internal/cimg!0/oo1il4sf8zzaqbboq25oevvbg99wpot.

A trial as a hypothesis test (cont.)

* If the evidence is not strong enough to reject the assumption of innocence, the jury returns with a verdict of “not guilty”.
  * The jury does not say that the defendant is innocent, just that there is not enough evidence to convict.
  * The defendant may, in fact, be innocent, but the jury has no way of being sure.
* Said statistically, we fail to reject the null hypothesis.
  * We never declare the null hypothesis to be true, because we simply do not know whether it’s true or not.
  * Therefore we never “accept the null hypothesis”.

A trial as a hypothesis test (cont.)

* In a trial, the burden of proof is on the prosecution.
* In a hypothesis test, the burden of proof is on the unusual claim.
* The null hypothesis is the ordinary state of affairs (the status quo), so it's the alternative hypothesis that we consider unusual and for which we must gather evidence.

Recap: hypothesis testing framework

* We start with a null hypothesis (H_0) that represents the status quo.
* We also have an alternative hypothesis (H_A) that represents our research question, i.e. what we're testing for.
* We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation (today) or theoretical methods (later in the course).
* If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

Simulating the experiment...

... under the assumption of independence, i.e. leave things up to chance.

If results from the simulations based on the chance model look like the data, then we can determine that the difference between the proportions of promoted files between males and females was simply due to chance (promotion and gender are independent).

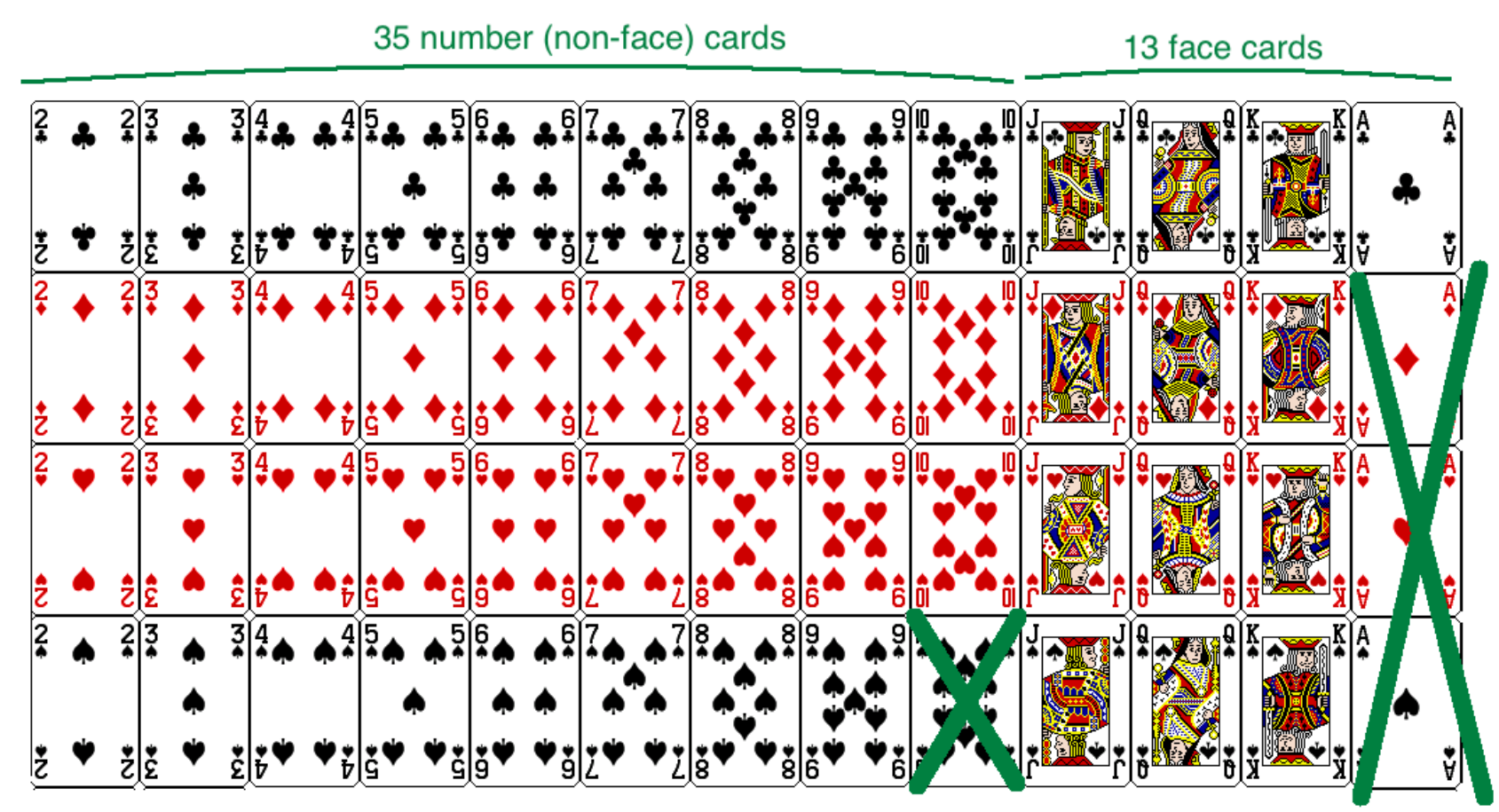
If the results from the simulations based on the chance model do not look like the data, then we can determine that the difference between the proportions of promoted files between males and females was not due to chance, but due to an actual effect of gender (promotion and gender are dependent).

Application activity: simulating the experiment

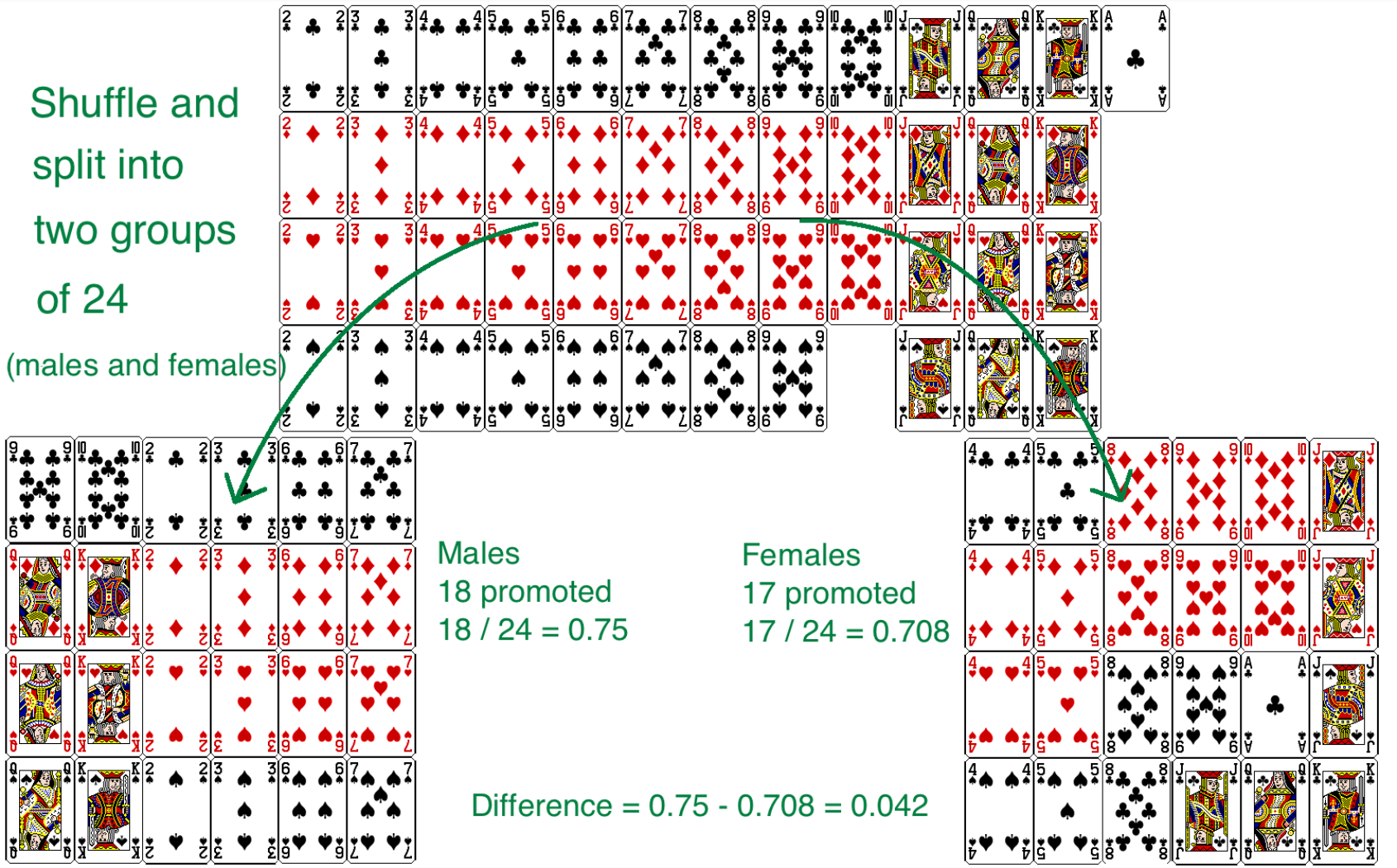
Use a deck of playing cards to simulate this experiment.

* Let a face card represent not promoted and a non-face card represent a promoted. Consider aces as face cards.
  * Set aside the jokers.
  * Take out 3 aces $\rightarrow$ there are exactly 13 face cards left in the deck (face cards: A, K, Q, J).
  * Take out a number card $\rightarrow$ there are exactly 35 number (non-face) cards left in the deck (number cards: 2-10).
* Shuffle the cards and deal them intro two groups of size 24, representing males and females.
* Count and record how many files in each group are promoted (number cards).
* Calculate the proportion of promoted files in each group and take the difference (male - female), and record this value.
* Repeat steps 2 - 4 many times.

Step 1



Step 2 - 4


Shuffle and
split into
two groups
of 24
(males and females)
Males
18 promoted
18 / 24 = 0.75
Females
17 promoted
17 / 24 = 0.708
Difference = 0.75 - 0.708 = 0.042

Practice

Do the results of the simulation you just ran provide convincing evidence of gender discrimination against women, i.e. dependence between gender and promotion decisions?

* No, the data do not provide convincing evidence for the alternative hypothesis, therefore we can't reject the null hypothesis of independence between gender and promotion decisions. The observed difference between the two proportions was due to chance.
* Yes, the data provide convincing evidence for the alternative hypothesis of gender discrimination against women in promotion decisions. The observed difference between the two proportions was due to a real effect of gender.

Simulations using software

These simulations are tedious and slow to run using the method described earlier. In reality, we use software to generate the simulations. The dot plot below shows the distribution of simulated differences in promotion rates based on 100 simulations.