Chapter 2: Summarizing data

STA 101, Summer I 2021, Duke University

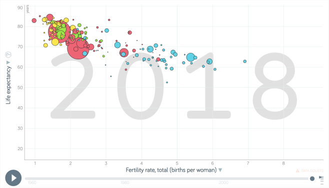
Examining numerical data

Scatterplot

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

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

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



http://www.gapminder.org/world

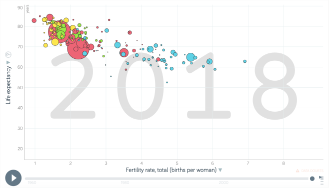
Scatterplot

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?



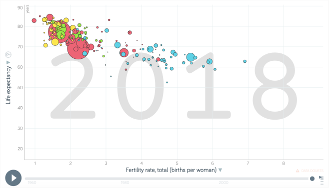
http://www.gapminder.org/world

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.



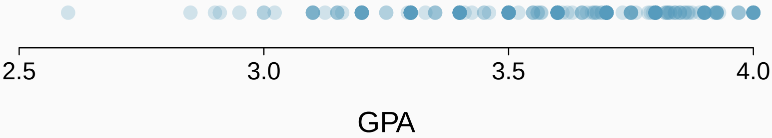
Was the relationship the same throughout the years, or did it change?

The relationship changed over the years.

http://www.gapminder.org/world

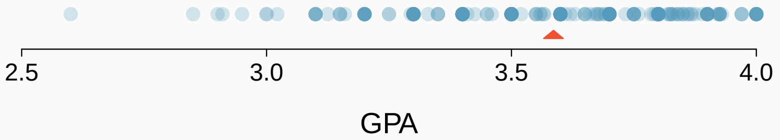
Dot plots

Useful for visualizing one numerical variable. Darker colors represent areas where there are more observations.



How would you describe the distribution of GPAs in this data set?

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.

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.

Mean

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

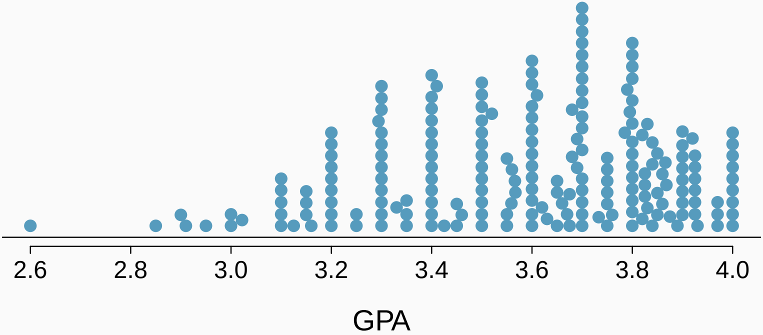
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

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.

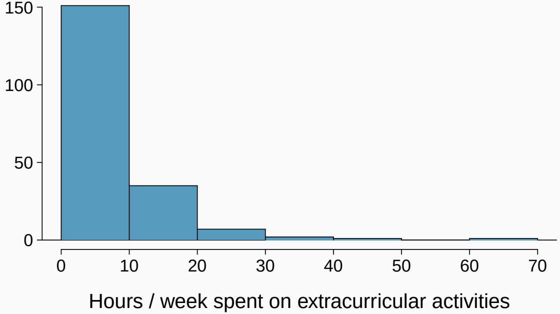
Stacked dot plot

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



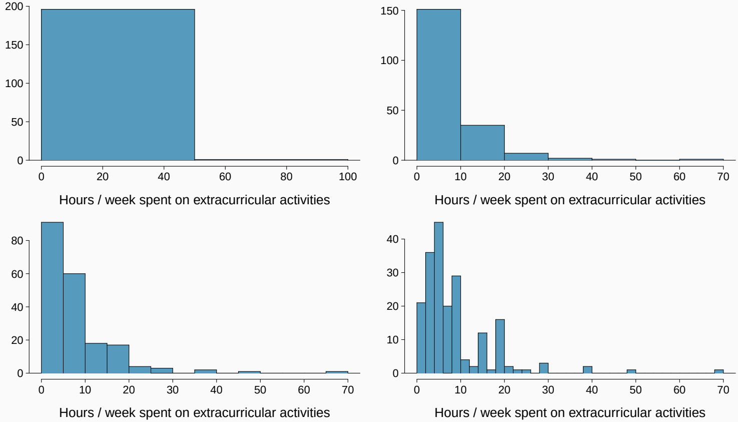
Histograms - Extracurricular hours

- 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.



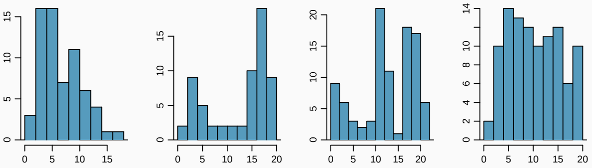
Bin width

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



Shape of a distribution: modality

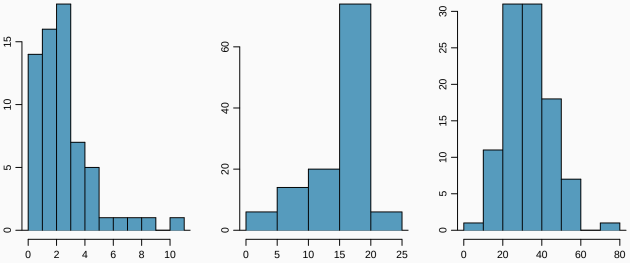
Does the histogram have a single prominent peak (*unimodal*), several prominent peaks (*bimodal/multimodal*), or no apparent peaks (*uniform*)?



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

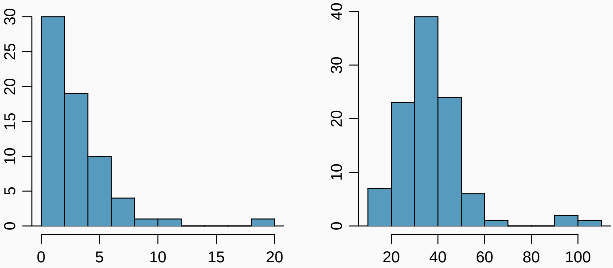
Is the histogram *right skewed*, *left skewed*, or *symmetric*?



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

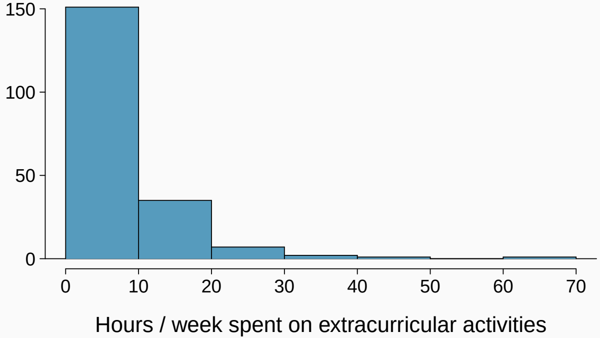
Shape of a distribution: unusual observations

Are there any unusual observations or potential *outliers*?



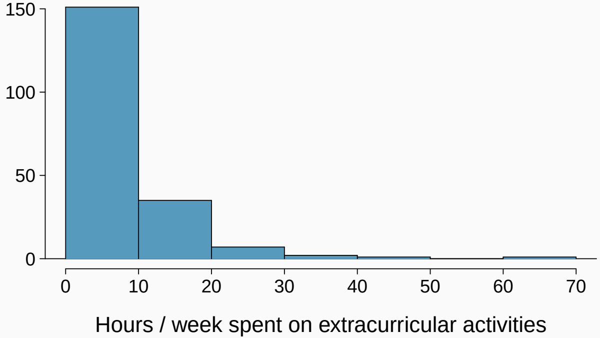
Extracurricular activities

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



Extracurricular activities

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

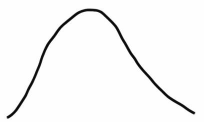


Unimodal and right skewed, with a potentially unusual observation at 60 hours/week.

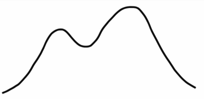
Commonly observed shapes of distributions

- modality

unimodal

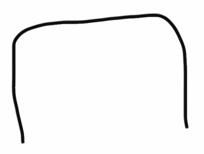


bimodal



multimodal

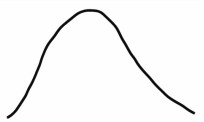
uniform



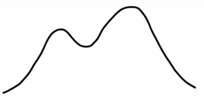
Commonly observed shapes of distributions

- modality

unimodal

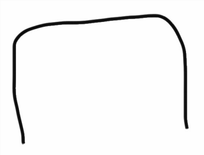


bimodal



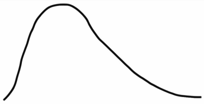
multimodal

uniform

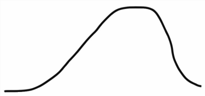


- skewness

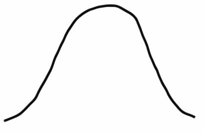
right skew



left skew



symmetric



Practice

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

(a) weights of adult females

(b) salaries of a random sample of people from North Carolina

(c) house prices

(d) birthdays of classmates (day of the month)

Practice

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

(a) weights of adult females

(b) salaries of a random sample of people from North Carolina

(c) house prices

(d) *birthdays of classmates (day of the month)*

Are you typical?

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

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.

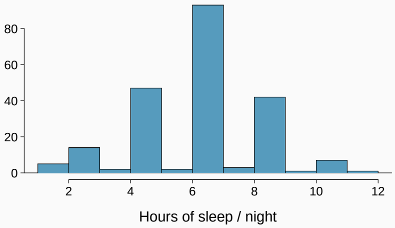
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

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 $n = 217$.

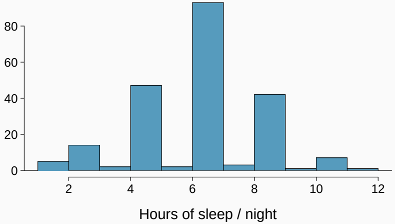


Variance

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:



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

Variance (cont.)

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

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

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

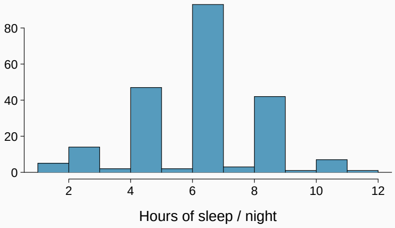
Standard deviation

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

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

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

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



Standard deviation

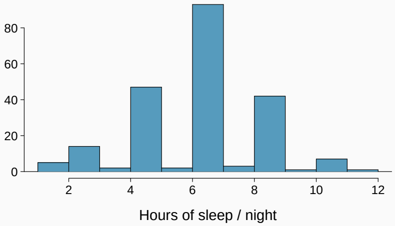
The *standard deviation* is the square root of the variance, and has the same units as the data

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

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

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

- 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*.

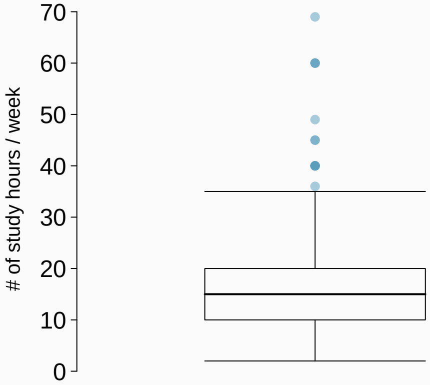
Q1, Q3, and IQR

- 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*.

$$IQR = Q3 - Q1$$

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 $1.5 \times IQR$ away from the quartiles.

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

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

Whiskers and outliers

- *Whiskers*
  of a box plot can extend up to $1.5 \times IQR$ away from the quartiles.

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

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

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

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

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

Whiskers and outliers

- *Whiskers*
  of a box plot can extend up to $1.5 \times IQR$ away from the quartiles.

$$\text{max upper whisker reach} = Q3 + 1.5 \times IQR$$
$$\text{min lower whisker reach} = Q1 - 1.5 \times IQR$$

$$\text{IQR} : 20 - 10 = 10$$
$$\text{max upper whisker reach} = 20 + 1.5 \times 10 = 35$$
$$\text{max lower whisker reach} = 10 - 1.5 \times 10 = -5$$

- 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?

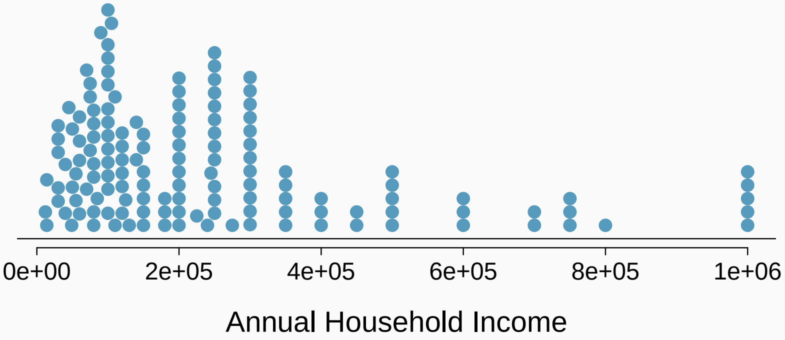
Outliers (cont.)

Why is it important to look for outliers?

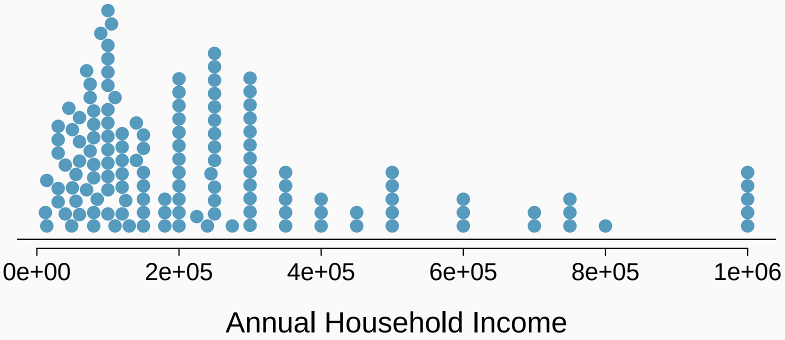
- *Identify extreme skew in the distribution.*
- *Identify data collection and entry errors.*
- *Provide insight into interesting features of the data.*

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?

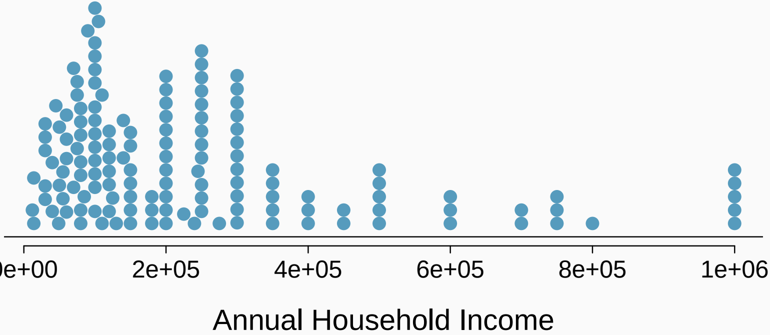


Robust statistics



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



	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?

Robust statistics

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

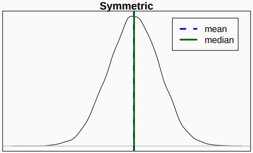
- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

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

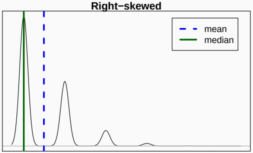
Median

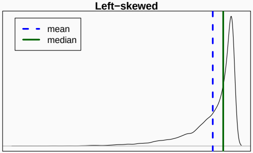
Mean vs. median

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



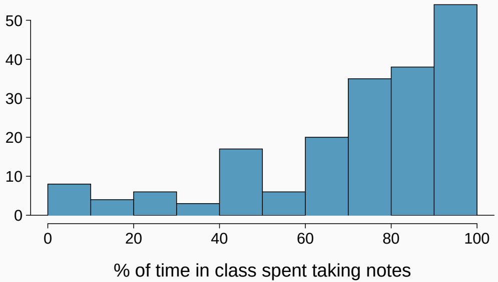
- If the distribution is skewed or has extreme outliers, center is often defined as the median
    - Right-skewed: mean $>$ median
    - Left-skewed: mean $<$ median





Practice

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



(a) mean> median

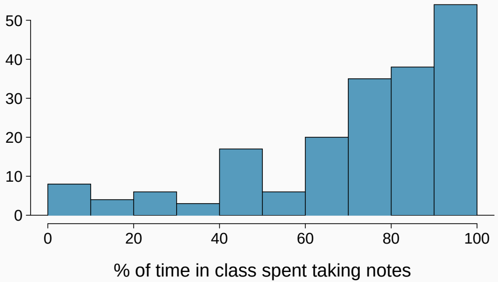
(b) mean < median

(c) mean ≈ median

(d) impossible to tell

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%

(a) mean> median

(b) *mean < median*

(c) mean ≈ median

(d) impossible to tell

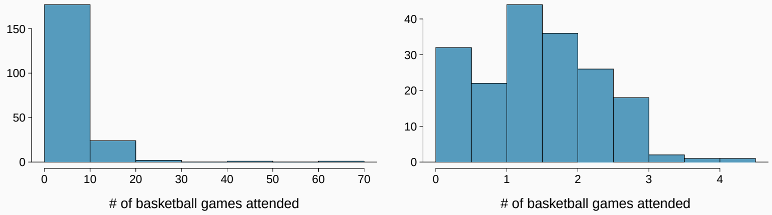
Extremely skewed data

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

Extremely skewed data

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

The histogram 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 in log units of the measured variable might be difficult to interpret.

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 in log units of the measured variable might be difficult to interpret.

What other variables would you expect to be extremely skewed?

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 in log units of the measured variable might be difficult to interpret.

What other variables would you expect to be extremely skewed?

Salary, housing prices, etc.

Considering categorical data

Contingency tables

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

Contingency tables

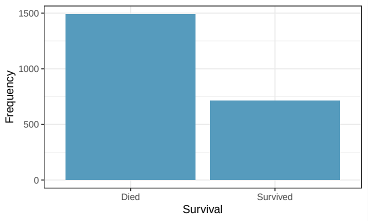
A table that summarizes data for two categorical variables is called a *contingency table*.

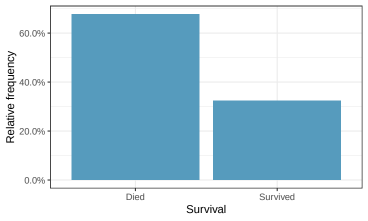
The contingency table below shows the distribution of survival and ages of passengers on the Titanic.

		Survival		
		Died	Survived	Total
Age	Adult	1438	654	2092
	Child	52	57	109
	Total	1490	711	2201

Bar plots

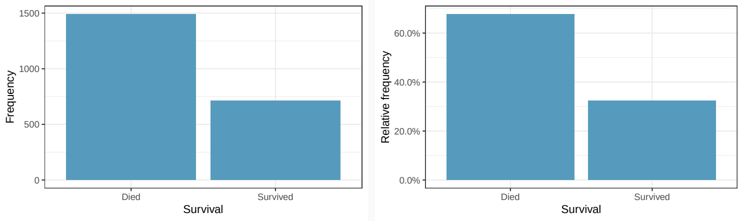
A *bar plot* is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a *relative frequency bar plot*.





Bar plots

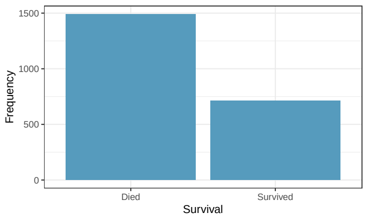
A *bar plot* is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a *relative frequency bar plot*.

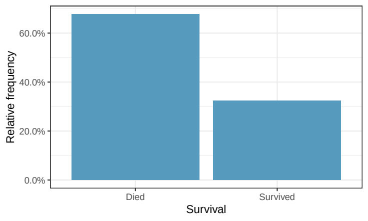


How are bar plots different than histograms?

Bar plots

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, 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. 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 age and survival for passengers on the Titanic?

		Survival		
		Died	Survived	Total
Age	Adult	1438	654	2092
	Child	52	57	109
	Total	1490	711	2201

Choosing the appropriate proportion

Does there appear to be a relationship between age and survival for passengers on the Titanic?

		Survival		
		Died	Survived	Total
Age	Adult	1438	654	2092
	Child	52	57	109
	Total	1490	711	2201

To answer this question we examine the row proportions:

Choosing the appropriate proportion

Does there appear to be a relationship between age and survival for passengers on the Titanic?

		Survival		
		Died	Survived	Total
Age	Adult	1438	654	2092
	Child	52	57	109
	Total	1490	711	2201

To answer this question we examine the row proportions:

- % Adults who survived: 654 / 2092 ≈ 0.31

Choosing the appropriate proportion

Does there appear to be a relationship between age and survival for passengers on the Titanic?

		Survival		
		Died	Survived	Total
Age	Adult	1438	654	2092
	Child	52	57	109
	Total	1490	711	2201

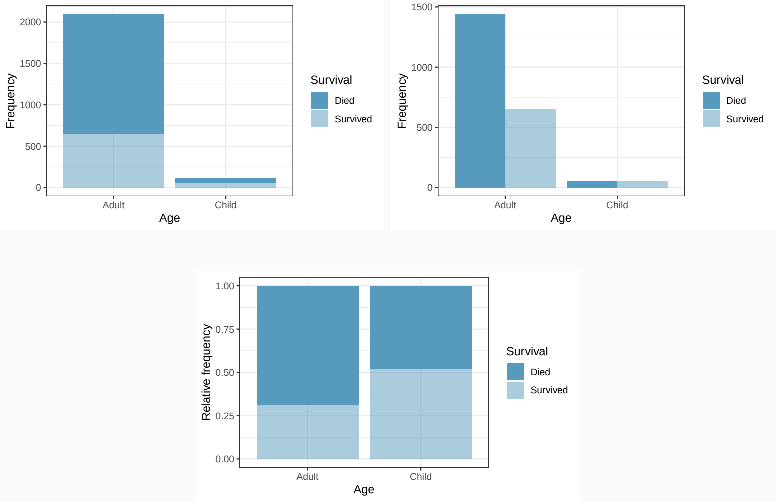
To answer this question we examine the row proportions:

- % Adults who survived: 654 / 2092 ≈ 0.31
- % Children who survived: 57 / 109 ≈ 0.52

Bar plots with two variables

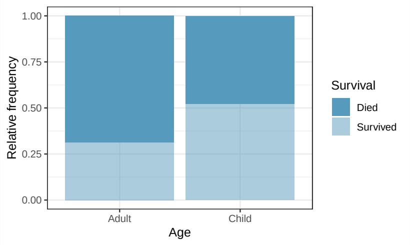
- *Stacked bar plot:* Graphical display of contingency table information, for counts.
- *Side-by-side bar plot:* Displays the same information by placing bars next to, instead of on top of, each other.
- *Standardized stacked bar plot*: Graphical display of contingency table information, for proportions.

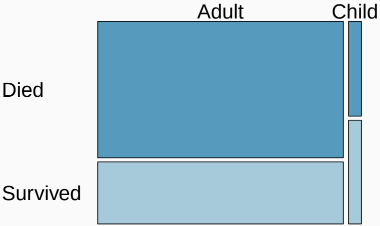
What are the differences between the three visualizations shown below?



Mosaic plots

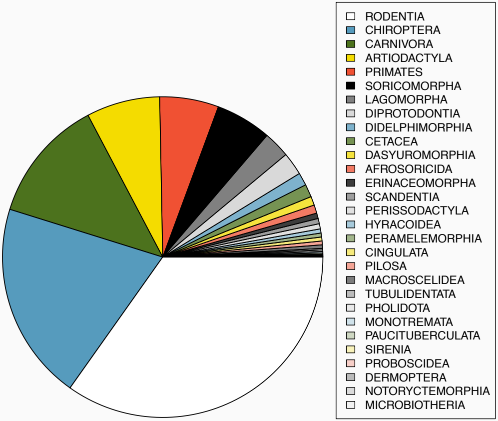
What is the difference between the two visualizations shown below?





Pie charts

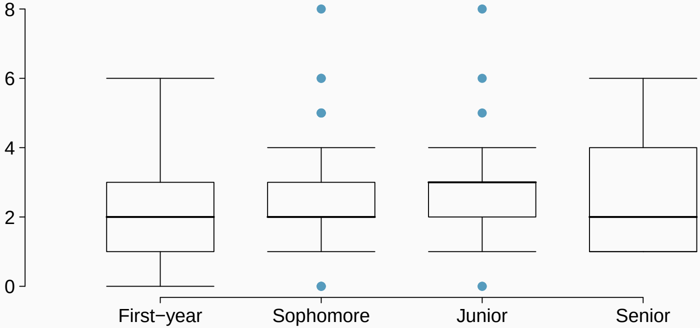
Can you tell which order encompasses the lowest percentage of mammal species?



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: Gender discrimination

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?

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?

Data

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

Data

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?

(a) If we were to repeat the experiment we will definitely see that more female files get promoted. This was a fluke.
(b) Promotion is dependent on gender, males are more likely to be promoted, and hence there is gender discrimination against women in promotion decisions.
(c) 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.
(d) Women are less qualified than men, and this is why fewer females get promoted.

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?

(a) If we were to repeat the experiment we will definitely see that more female files get promoted. This was a fluke.

(b) Promotion is dependent on gender, males are more likely to be promoted, and hence there is gender discrimination against women in promotion decisions. *Maybe*

(c) 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*

(d) Women are less qualified than men, and this is why fewer females get promoted.

Two competing claims

1. "There is nothing going on."
   Promotion and gender are *independent*, no gender discrimination, observed difference in proportions is simply due to chance. → *Null hypothesis*

Two competing claims

1. "There is nothing going on."
   Promotion and gender are *independent*, no gender discrimination, observed difference in proportions is simply due to chance. → *Null hypothesis*
2. "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.

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).

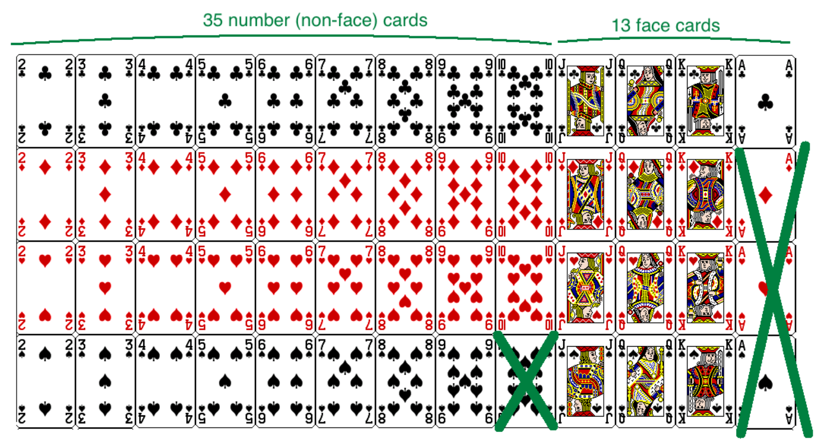
If the results from the simulations based on the chance model do not look like the data, then we can determine that the difference between the proportions of promoted files between males and females was not due to chance, but *due to an actual effect of gender* (promotion and gender are dependent).

Application activity: simulating the experiment

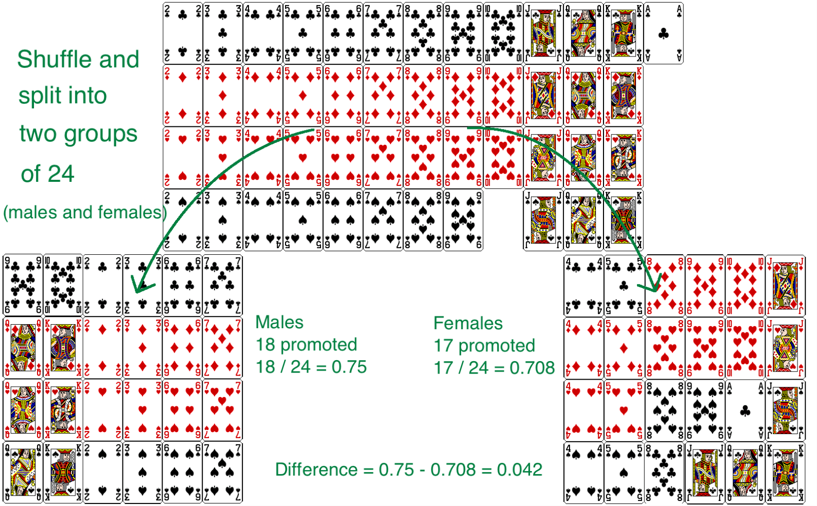
Use a deck of playing cards to simulate this experiment.

1. 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 → there are exactly 13 face cards left in the deck (face cards: A, K, Q, J).
   - Take out a number card → there are exactly 35 number (non-face) cards left in the deck (number cards: 2-10).
2. Shuffle the cards and deal them intro two groups of size 24, representing males and females.
3. Count and record how many files in each group are promoted (number cards).
4. Calculate the proportion of promoted files in each group and take the difference (male - female), and record this value.
5. Repeat steps 2 - 4 many times.

Step 1

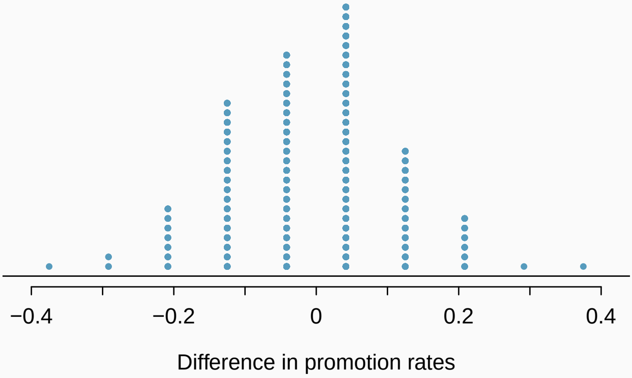


Step 2 - 4


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

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.



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?

(a) 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.

(b) 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.

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?

(a) 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.

(b) *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.*