Chapter 2: Exploratory Data Analysis - Summarizing data

Wen-Han Hwang
(Slides primarily developed by Mine Çetinkaya-Rundel from OpenIntro.)

Institute of Statistics
National Tsing Hua University
Taiwan

OpenIntro

Outline

1. What is Exploratory Data Analysis?
2. Examining numerical data
3. Considering categorical data
4. Case study: Gender discrimination

What is Exploratory Data Analysis?

Introduction to Exploratory Data Analysis

- EDA is an analytical approach to **understand the data** by summarizing their main characteristics, often visually.
- It helps in **detecting outliers**, **testing assumptions**, and **checking for patterns or anomalies** in the data.
- The goal is to gain insights and inform further data analysis and inference.

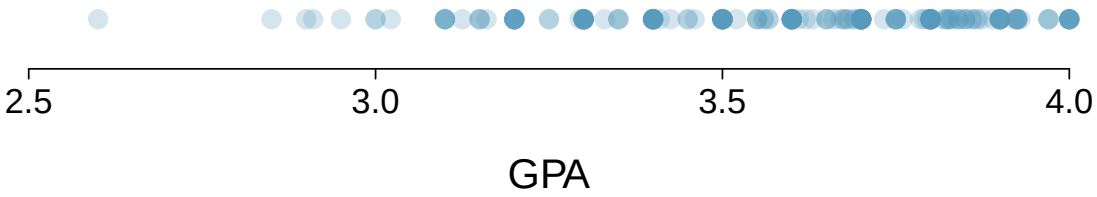
Techniques in EDA

- **Summary Statistics:** Key measures like Mean, Median, Mode, Range, Standard Deviation, Quartiles, and the Correlation coefficient provide a quick snapshot of data characteristics.
- **Visualization Techniques:**
  - **Histograms** reveal the distribution of continuous data, showing how often each value appears.
  - **Bar Charts** compare discrete data quantities across different categories.
  - **Pie Charts** illustrate the proportional representation of different categories within a whole.
  - **Box Plots** highlight the median, quartiles, and potential outliers in data distributions.
  - **Scatter Plots** explore relationships between two variables, helping to identify trends, clusters, and outliers.

Examining numerical data

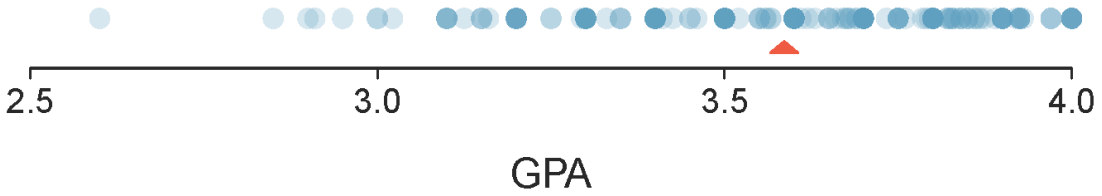
Dot plots

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? Make sure to say something about the center, shape, and spread of the distribution.

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

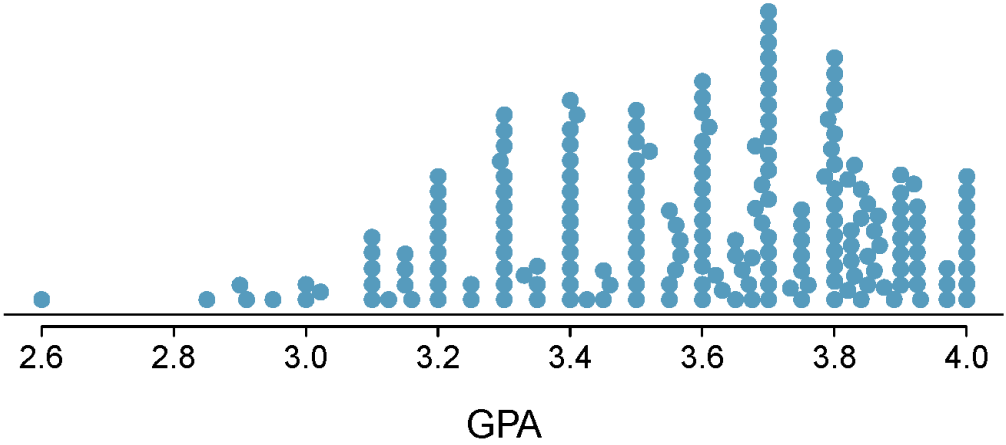
- The sample mean, denoted as $\bar{x}$, can be calculated as

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

  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.

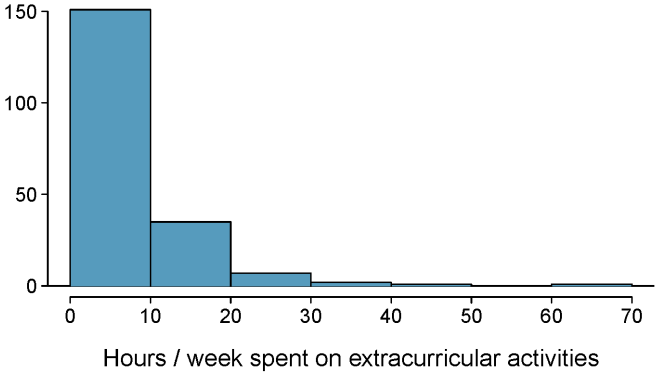
Stacked dot plot

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



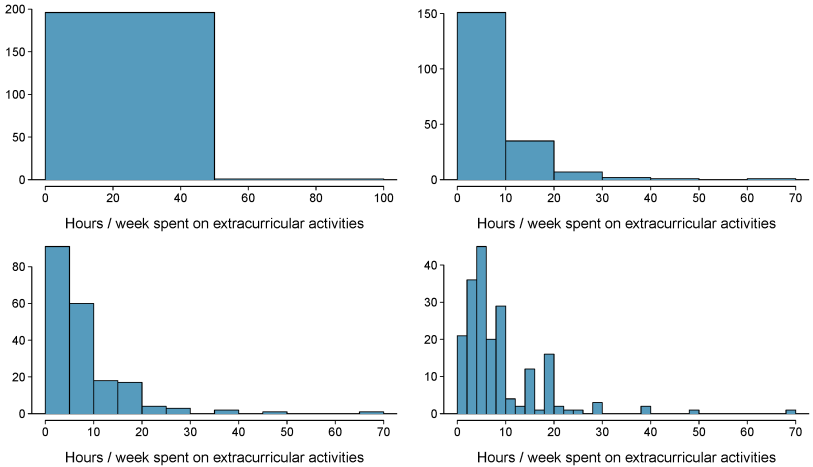
Histograms - Extracurricular hours

- 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.



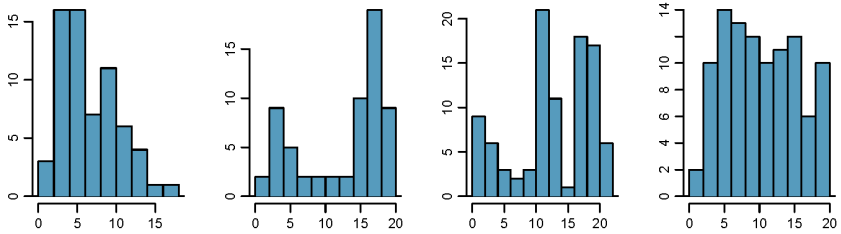
Bin width

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



Shape of a distribution: modality

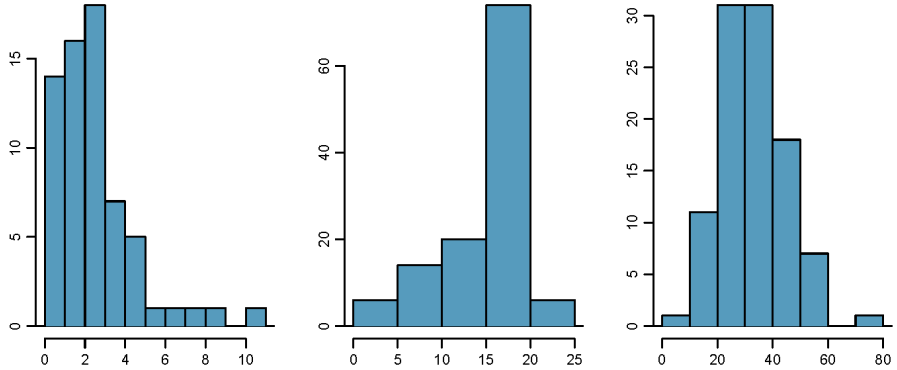
Does the histogram have a single prominent peak (unimodal), several prominent peaks (bimodal/multimodal), or no apparent peaks (uniform)?



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

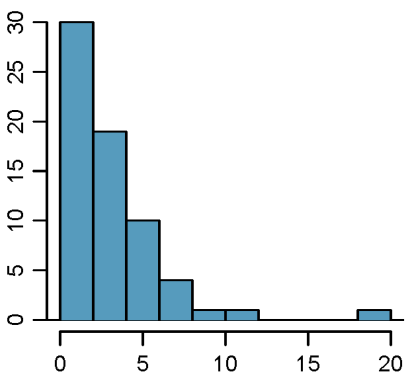
Is the histogram right skewed, left skewed, or symmetric?

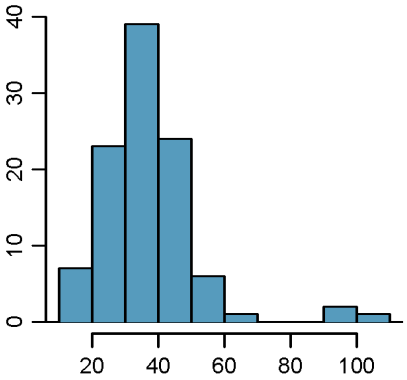


Note: Histograms are said to be skewed to the side of the long tail.

Shape of a distribution: unusual observations

Are there any unusual observations or potential outliers?

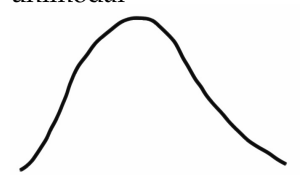




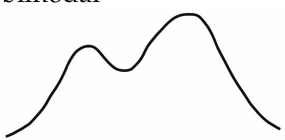
Commonly observed shapes of distributions

- modality

unimodal



bimodal



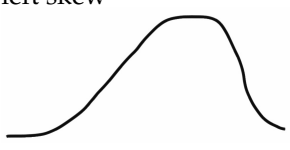
multimodal

uniform

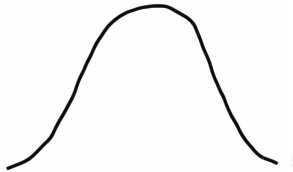
- skewness

right skew

left skew

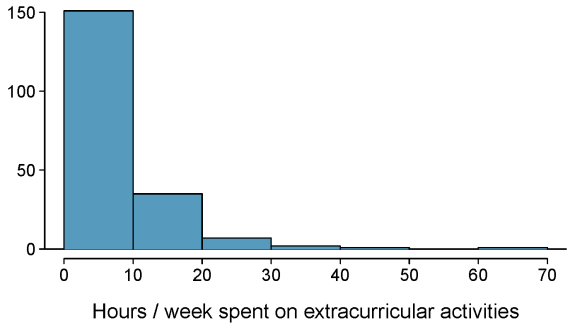


symmetric



Extracurricular activities

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



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 Taipei
(c) house prices
(d)

Are you typical?

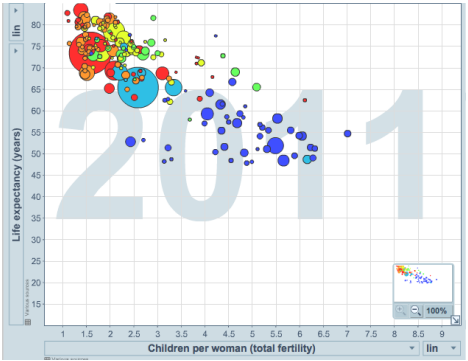
http://www.youtube.com/watch?v=4B2xOvKFFz4

Earth's Most Typical Person

- **Background:** CBS News and National Geographic's analysis revealed "Earth's most typical person" based on global demographic data.
-
  - Age: 28 years old — matching the global median age.
  - Ethnicity: Han Chinese, the largest demographic group identified.
  - Additional Characteristics:
    - Right-handed, speaks Mandarin.
    - Christian, owns a cell phone.
    - Does not have a car or bank account.
    - Earns less than $12,000 a year.
- Represents the largest pool found, with 9 million Han Chinese men sharing similar characteristics.

Scatterplots for paired data

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



`http://www.gapminder.org/world`

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?

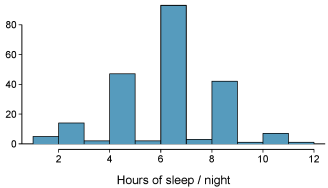
The relationship changed over the years.

Variance

Variance is roughly the average squared deviation from the mean.

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

- **Data:** Hours of sleep per night for a group of students.
- **Sample Size:** $n = 217$ students.
- **Sample Mean:** $\bar{x} = 6.71$ hours.
- **Variance Calculation:** Shows the spread of sleep hours among the students.



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

Variance (cont.)

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

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

Standard deviation

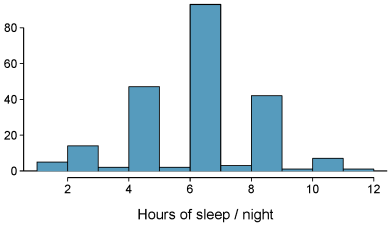
The standard deviation is the square root of the variance, and has the same units as the data

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

- 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.

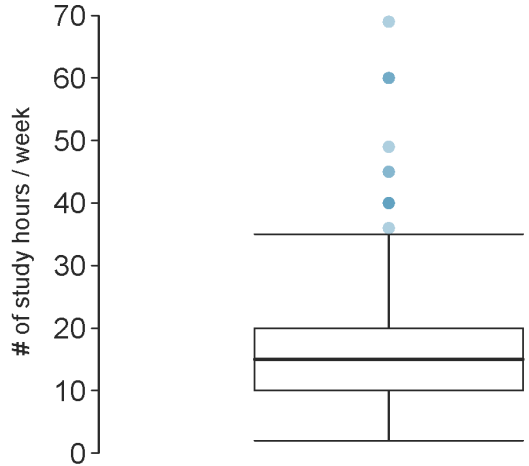
Q1, Q3, and IQR

- 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.

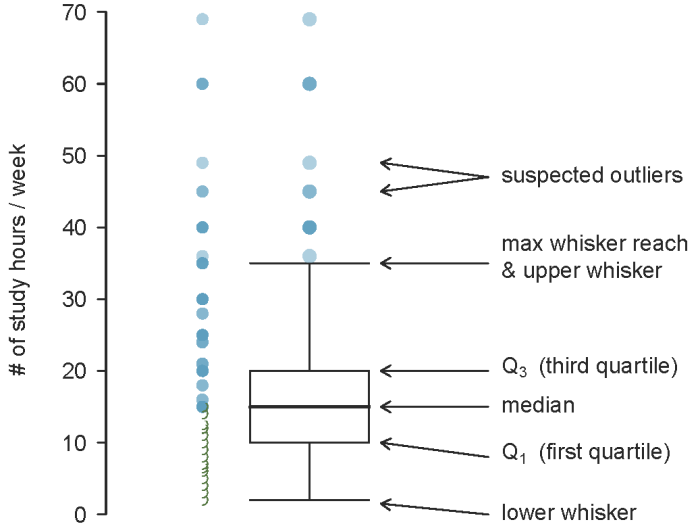
$$IQR = Q3 - Q1$$

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.

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

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

- 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.

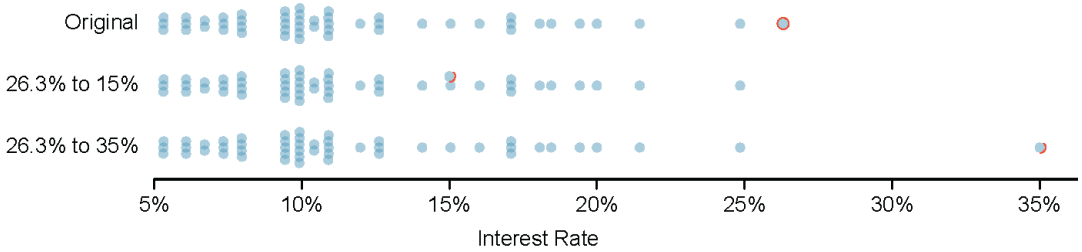
Outliers (cont.)

Why is it important to look for outliers?

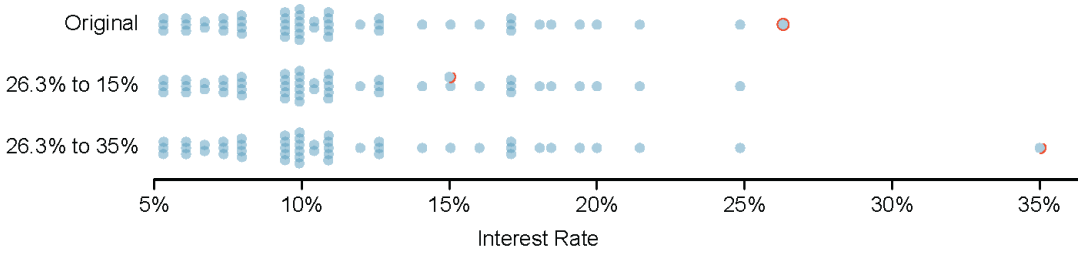
- Identify extreme skew in the distribution.
- Identify data collection and entry errors.
- Provide insight into interesting features of the data.

Extreme observations

How would sample statistics such as mean, median, SD, and IQR of the interest rate data set affected by the observation, 26.3%? What would have happened if this loan had instead been only 15%? What would happen to these summary statistics if the observation at 26.3% had been even larger, say 35%?



Robust statistics



	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original interest rate data	9.93%	5.76%	11.57%	5.05%
move 26.3% $\rightarrow$15%	9.93%	5.76%	11.34%	4.61%
move 26.3% $\rightarrow$35%	9.93%	5.76%	11.74%	5.68%

Robust statistics

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

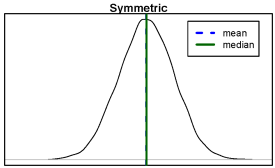
- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

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

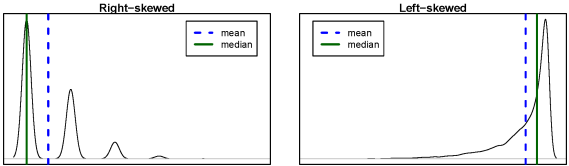
Median

Mean vs. median

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

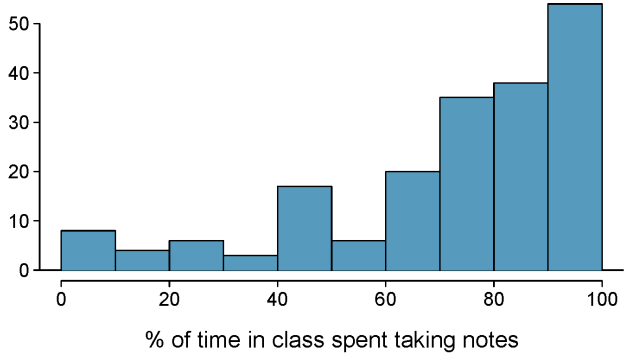


- If the distribution is skewed or has extreme outliers, center is often defined as the median
  - Right-skewed: mean $>$ median
  - Left-skewed: mean $<$ median



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

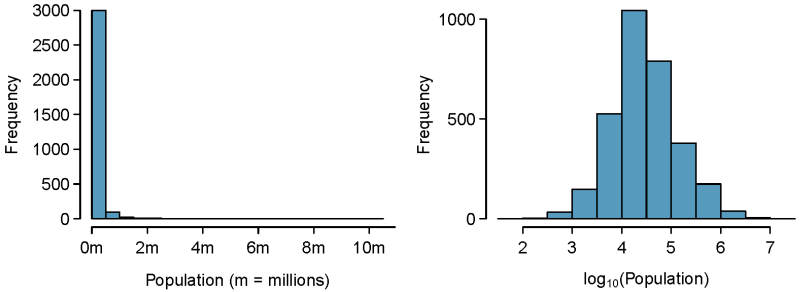
(c) mean $\approx$ median

(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.

The histogram on the left shows the populations of all US counties. The histogram on the right shows the $\log_{10}$-transformed county populations.



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.

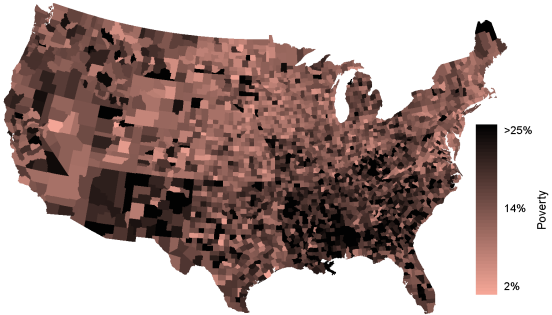
# of games	70	50	25	$\cdots$
log(# of games)	4.25	3.91	3.22	$\cdots$

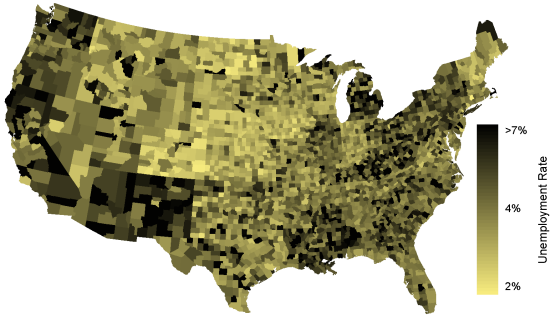
- However, results of an analysis in log units of the measured variable might be difficult to interpret.

What other variables would you expect to be extremely skewed?

Intensity maps

The left image shows the intensity map of the poverty rate (percent) and the right shows the map of the unemployment rate (percent).





Considering categorical data

Contingency tables

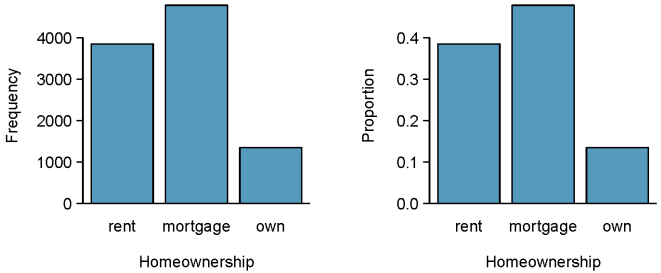
A table that summarizes data for two categorical variables is called a contingency table.

The contingency table below shows the distribution of app type and homeownership.

		homeownership			
		rent	mortgage	own	Total
app type	individual	3496	3839	1170	8505
	joint	362	950	183	1495
	Total	3858	4789	1353	10000

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.



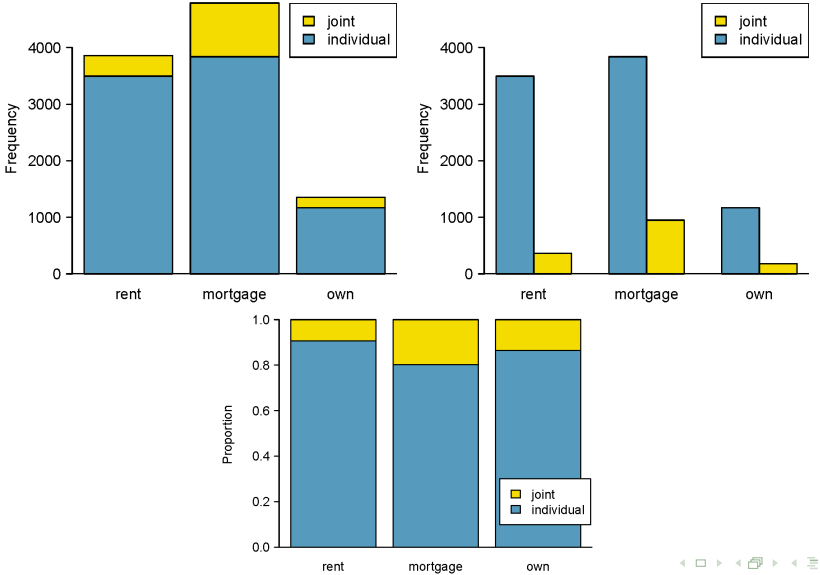
How are bar plots different than histograms?

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.)

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?

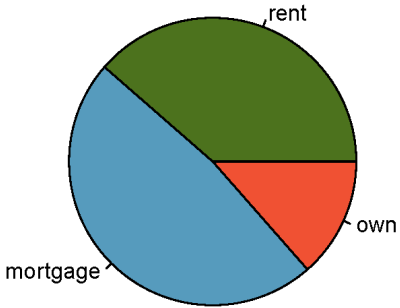


Mosaic plots

What is the difference between the two visualizations shown below?

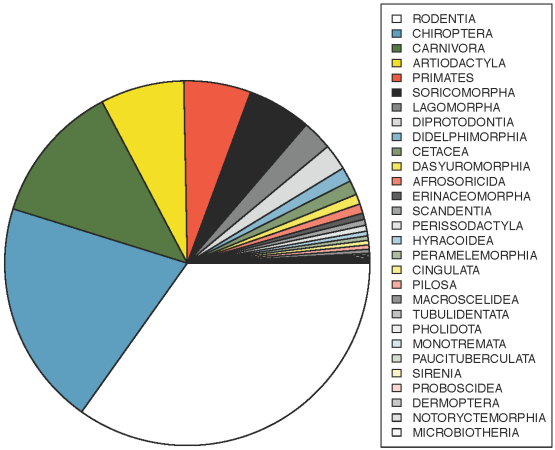
Pie charts

Pie charts can be useful for giving a high-level overview to show how a set of cases break down.



Pie charts

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

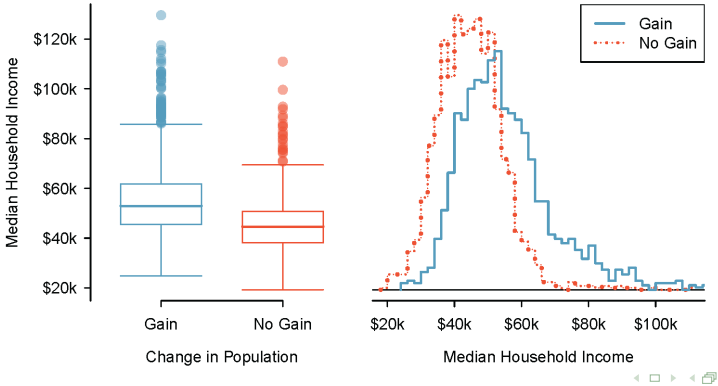


Data from `http://www.bucknell.edu/msw3`.

Side-by-side box plots

There were 1,454 counties where the population increased from 2010 to 2017, and there were 1,672 counties with no gain. A random sample of 100 counties from the first group and 50 from the second group.

The side-by-side box plot is a traditional tool for comparing across groups.



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?

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

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

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.
- 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.
- 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. $\rightarrow$ 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. $\rightarrow$ 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?

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.)

- **Burden of Proof:**
  - In trials, the *prosecution* carries the burden of proof.
  - In hypothesis testing, the burden is on the *test statistic*.
- **Hypothesis Overview:**
  - The *null hypothesis* represents the normal condition or the status quo.
  - The *alternative hypothesis* is viewed as the exceptional claim, requiring evidence to support it.

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).

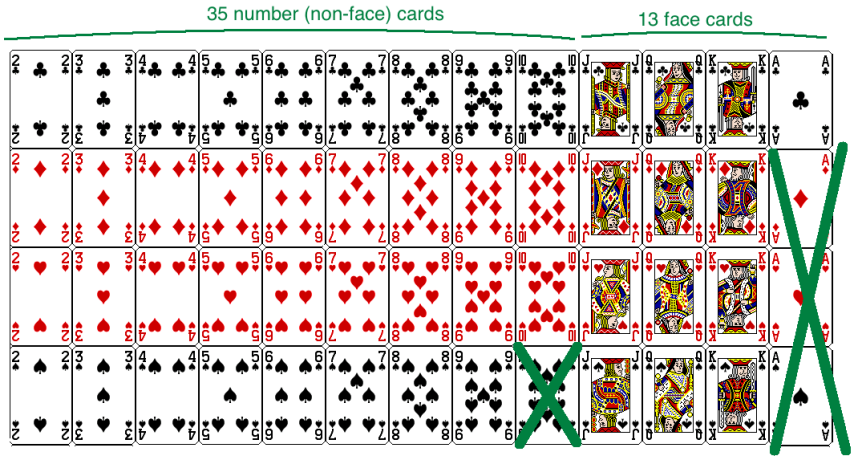
If the results from the simulations based on the chance model do not look like the data, then we can determine that the difference between the proportions of promoted files between males and females was not due to chance, but due to an actual effect of gender (promotion and gender are dependent).

Application activity: simulating the experiment

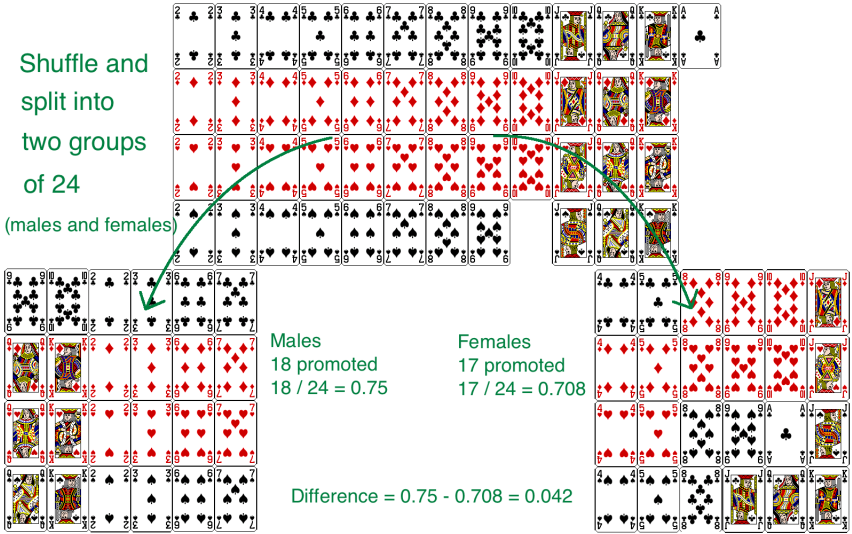
- Use a deck of playing cards to simulate this experiment.

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 $\rightarrow$ 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).
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

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.