CHAPTER 2

Summarizing Data

Variable
Characteristic of the individual

Numerical variable
Takes numerical values for which arithmetic operations make sense

Scatterplots
Dot plots
Histogram
Box plots

Categorical variable
Places individual into one of several groups or categories

Bar plot
Pie chart

2.1 Examining Numerical Data

Scatterplots for Paired Data

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

A *scatterplot:*

- Shows the relationship between two quantitative variables measured on the same individuals.
- The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis.
- Each individual corresponds to one point on the graph.

How to Make a Scatterplot

1. Decide which variable should go on each axis. If a distinction exists, plot the explanatory variable on the *x* axis and the response variable on the *y* axis.
2. Label and scale your axes.
3. Plot individual data values.

Example

Make a scatterplot of the relationship between body weight and backpack weight for a group of hikers.

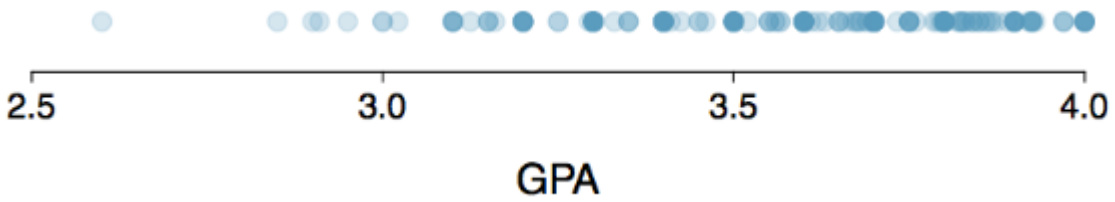
Body weight (lb)	120	187	109	103	131	165	158	116
Backpack weight (lb)	26	30	26	24	29	35	31	28

Dot Plot

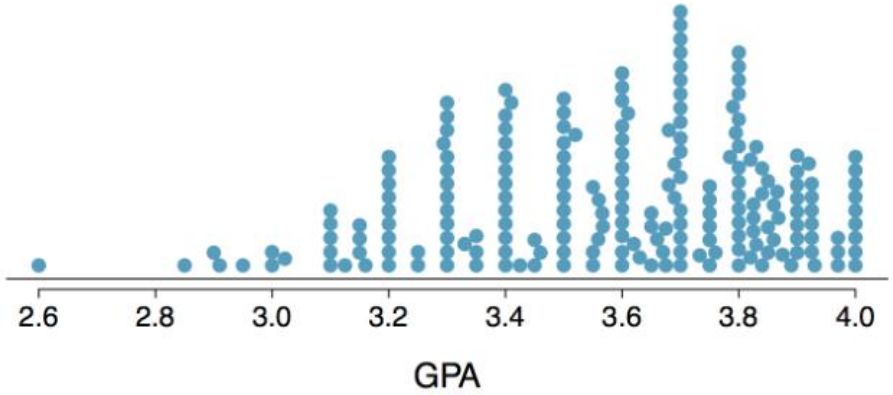
Dot plot is a Useful way for visualizing one numerical variable.

Example:

The distribution of GPAs off 200 students.



Darker colors represent areas where there are more observations.



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

The Mean ($\bar{x}$)

The *mean*, also called the *average*, is one way to measure the center of a *distribution* of data. The *sample mean*, denoted as $\bar{x}$, can be calculated as

$$\bar{x} = \frac{Sum\ of\ Observation}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

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 (Greek letter mu) . It is often not possible to calculate μ since population data are rarely available.

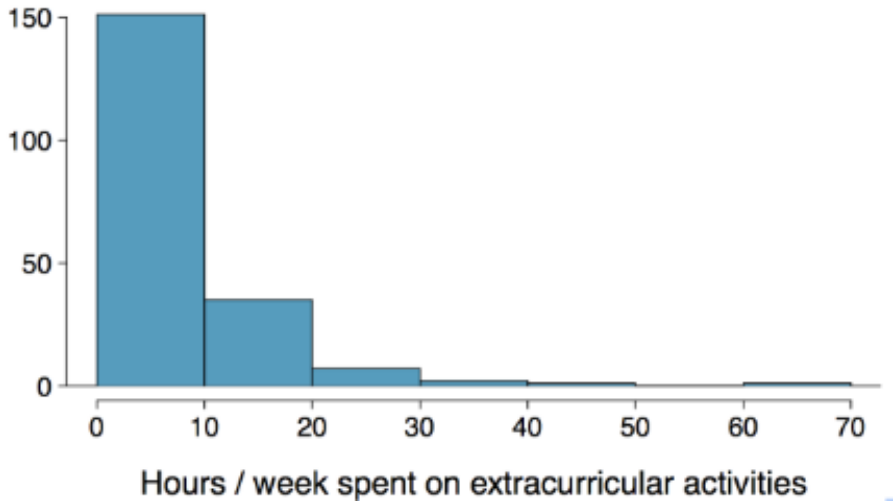
The sample mean is a *sample statistic and* serves as a *point estimate* of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population), it is usually a pretty good estimate.

Histogram and Shape

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.



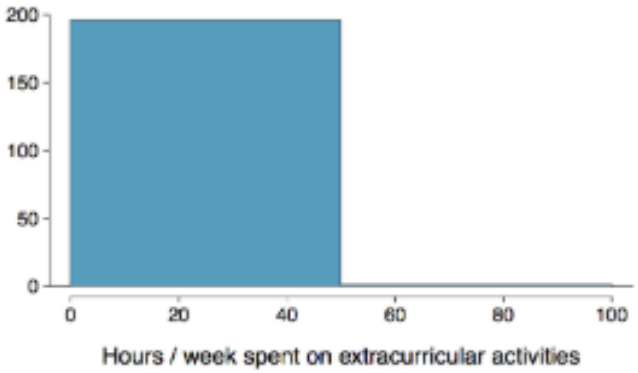
How to make a Histogram

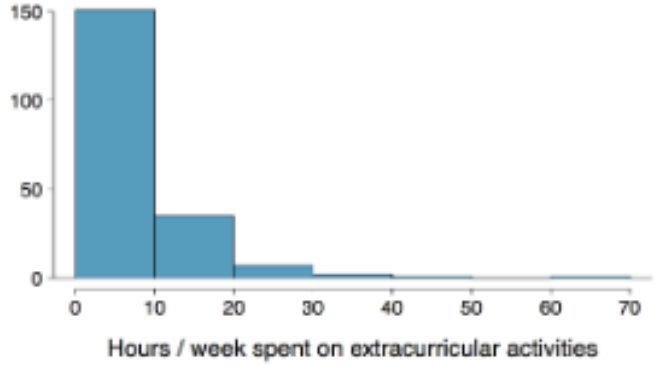
For large datasets and/or quantitative variables that take many values:

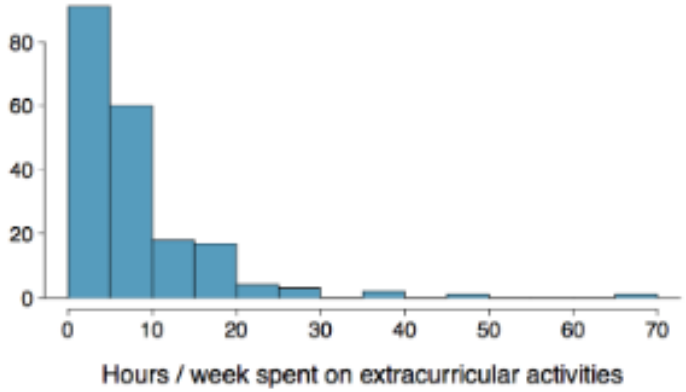
- Divide the possible values into Bins (equal widths). Be sure to specify the bins precisely, so that each individual falls into exactly one class.
- Count how many observations fall into each Bin (may change to percent). Check that the counts add to the number of individuals in the data set.
- Draw a picture representing the distribution**.** Each bar height is equal to the number (percent) of observations in its interval.

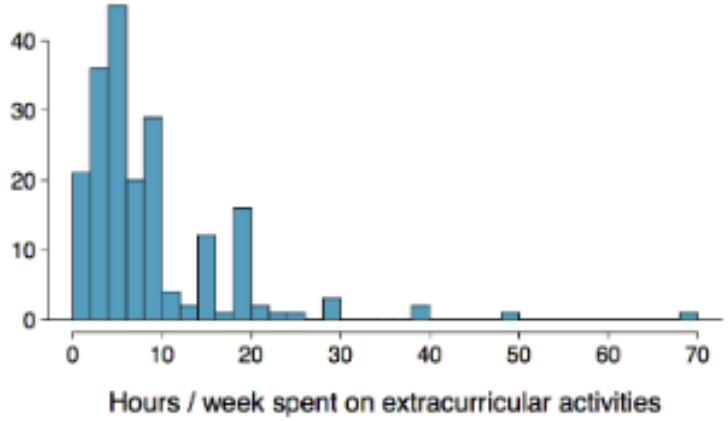
Bin Width

The appearance of a histogram can change when you change the bins width.



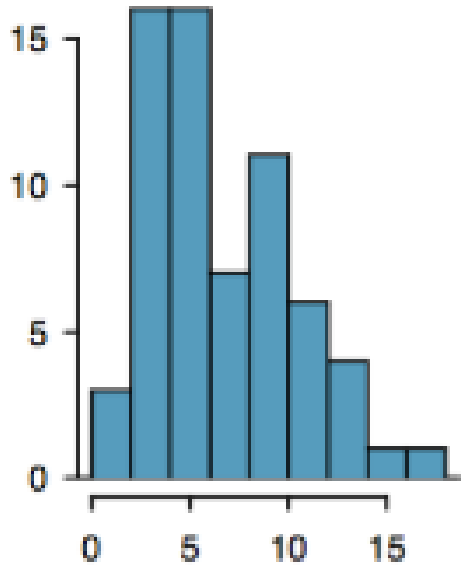




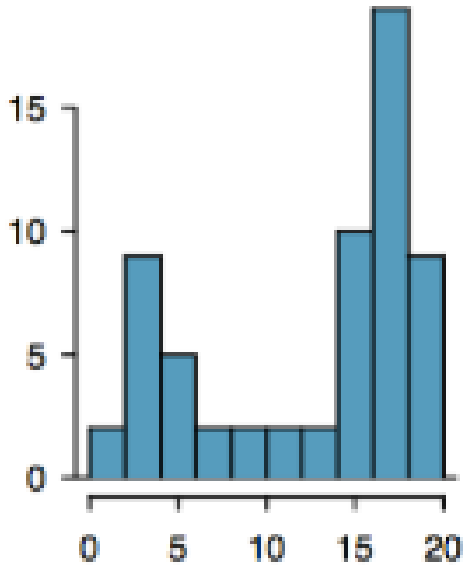


Shape of a Distribution: Modality

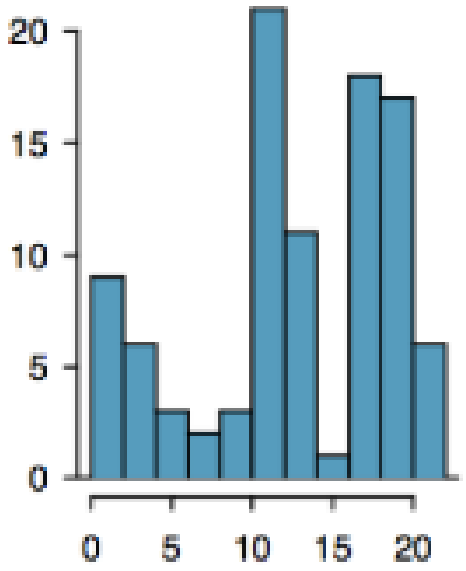
Points at which distributional shapes peak are called **modes**



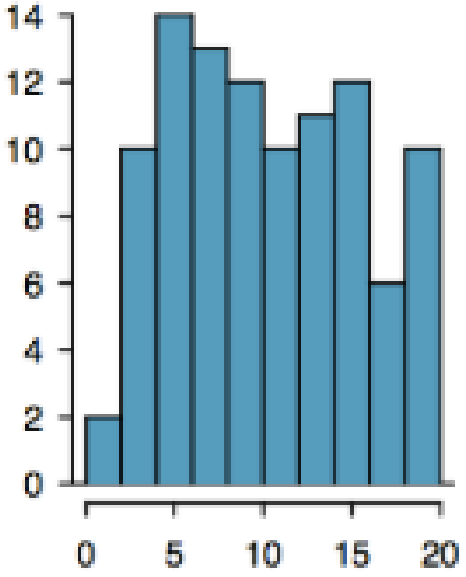
Unimodal:
distribution with one mode.



Bimodal:
distribution with two mode.



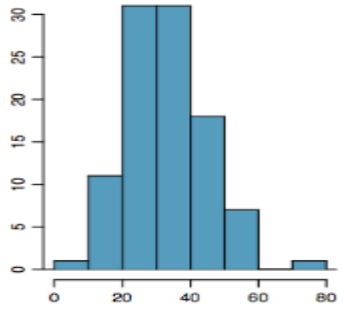
multimodal:
distribution with one mode.



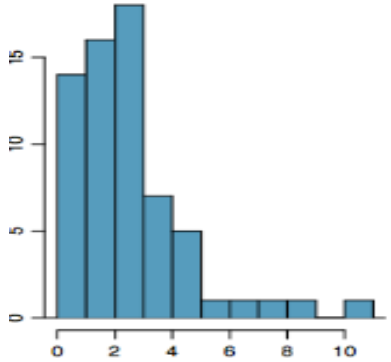
Uniform:
distribution with no apparent mode.

Shape of a Distribution: Skewness

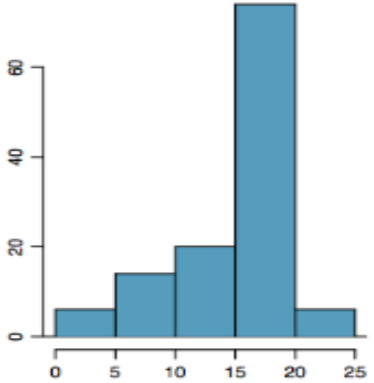
- **Symmetric:**
  - The left and right "tails" are approximately mirror images of each other.



- **Skewed right:**
  - If the right side of the graph (containing the half of the observations with larger values) is much longer than the left side.
  - The distribution has a long right tail.



- **Skewed left:**
  - If the left side of the graph is much longer than the right
  - The distribution has a long left tail.



Variance and Standard Deviation

The most common measure of spread looks at how far each observation is from the mean. This measure is called the *standard deviation.*

The *standard deviation* (***s***) measures the average distance of the observations from their mean. It is calculated by finding an average of the squared distances and then taking the square root. This average squared distance is called the **variance.**

$$\text{Variance} = s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1} = \frac{\sum(x_i-\bar{x})^2}{n-1}$$

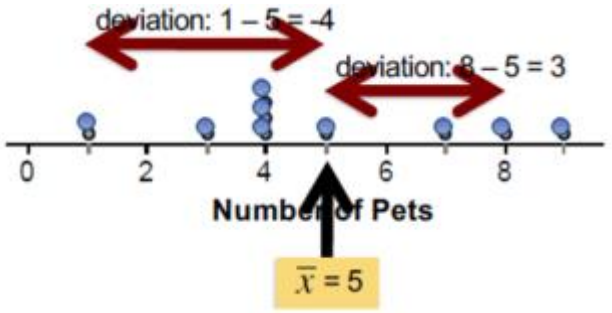
$$\text{Standard deviation} = s_x = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$$

The *population standard deviation* is denoted as σ (Greek letter sigma)

Calculating the Standard Deviation

Consider the following data on the number of pets owned by a group of nine children.

1. Calculate the mean
2. Calculate each deviation
3. Square each deviation
4. Find the "average" squared deviation. Calculate the sum of the squared deviations divided by (n-1). This is called the variance.
5. Calculate the square root of the variance. This is the standard deviation.



x_i	$(x_i - \text{mean})$	$(x_i - \text{mean})^2$
1	$1 - 5 = -4$	$(-4)^2 = 16$
3	$3 - 5 = -2$	$(-2)^2 = 4$
4	$4 - 5 = -1$	$(-1)^2 = 1$
4	$4 - 5 = -1$	$(-1)^2 = 1$
4	$4 - 5 = -1$	$(-1)^2 = 1$
5	$5 - 5 = 0$	$(0)^2 = 0$
7	$7 - 5 = 2$	$(2)^2 = 4$
8	$8 - 5 = 3$	$(3)^2 = 9$
9	$9 - 5 = 4$	$(4)^2 = 16$
	Sum = ?	**Sum = ?**

"Average" squared deviation = 52/(9 – 1) = 6.5. This is the **variance.**

Standard deviation = square root of variance = $\sqrt{6.5} = 2.55$

Practice Problem

Consider again the wind speeds of the hurricanes and tropical storms in August 2005

40 50 65 105 175

Find the variance and the standard deviation.

Median (M)

The median *M* is the midpoint of a distribution, the number such that half of the observations are smaller, and the other half are larger.

To find the median of a distribution:

1. Arrange all observations from smallest to largest.
2. If the number of observations *n* is odd, the median *M* is the center observation in the ordered list. Find M by counting $(n + 1)/2$ up from the beginning or from the end of the order list .
3. If the number of observations *n* is even, the median *M* is the average of the two center observations in the ordered list. Find M by averaging the values that are $n/2$ and $(n/2) + 1$ from the beginning or the end of the list

Examples:

Find the mean and median of the wind speeds (mph) of the hurricanes/ tropical storms in August 2005

40 50 65 105 175

Find the mean and median of the wind speeds (mph) of the hurricanes/ tropical storms in August 2004

40 65 70 105 145 45 135 120

Q1, Q3, and IQR

The 25th percentile is also called the first quartile, *Q1*.

The 50th percentile is also called the median.

The 75th 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*.

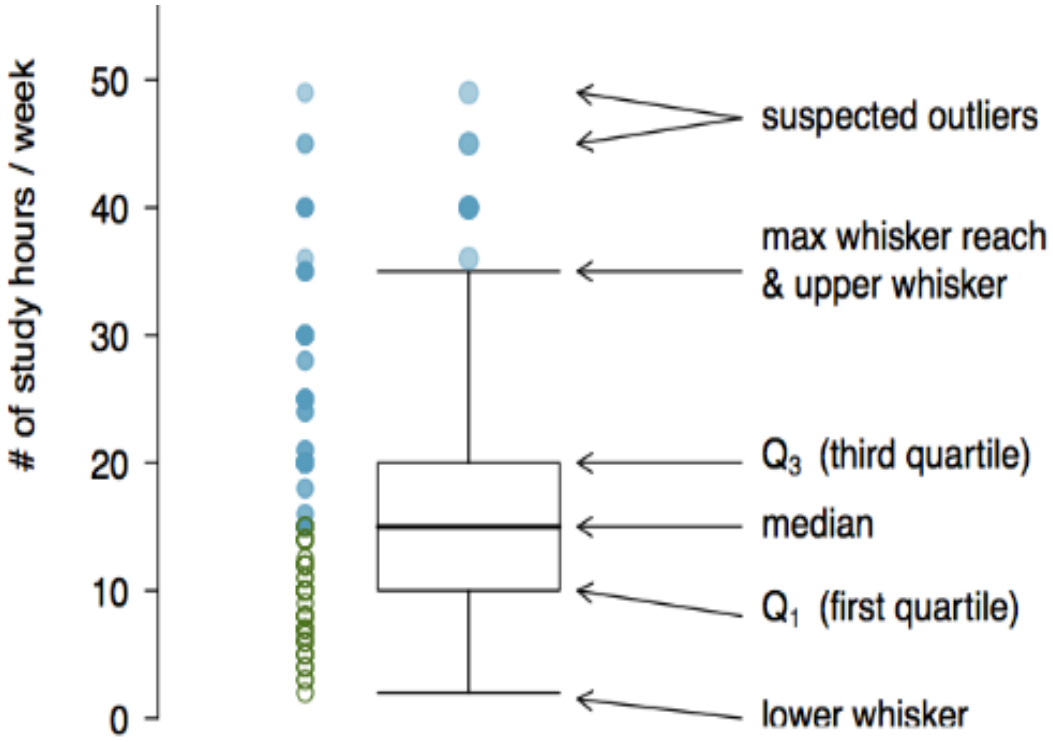
$$IQR = Q3 - Q1$$

Boxplots

The median and quartiles divide the distribution roughly into quarters. This leads to a new way to display quantitative data, the boxplot.

How to make a boxplot:

- Draw and label a number line that includes the range of the distribution.
- Draw a central box from Q_1 to Q_3.
- Note the median M inside the box.
- Extend lines (whiskers) from the box out to the minimum and maximum values that are not outliers.



Whiskers and Outliers

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

max upper whisker reach = Q3 + 1.5 x IQR

max lower whisker reach = Q1 - 1.5 x IQR

A potential *outlier* is defined as an observation beyond the maximum reach of the whiskers. It is an observation that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.

The Five-Number Summary

The minimum and maximum values alone tell us little about the distribution as a whole. Likewise, the median and quartiles tell us little about the tails of a distribution.

To get a quick summary of both center and spread, combine all five numbers.

The *five-number summary* of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest.

$$\textit{Minimum} \quad Q_1 \quad M \quad Q_3 \quad \textit{Maximum}$$

Example

Here are the males' responses to the question about how fast they have driven a car, as given in Case Study 1.1, except now the data are in numerical order. To make them easier to count, the data are arranged in rows of ten numbers. Calculate the Five-Number Summary.

55	60	80	80	80	80	85	85	85	85
90	90	90	90	90	92	94	95	95	95
95	95	95	10	100	100	100	100	100	100
100	100	101	102	105	105	105	105	105	105
105	105	109	110	110	110	110	110	110	110
115	115	120	120	120	120	120	120	120	120
120	120	124	125	125	125	125	125	125	130
130	140	140	140	140	145	150			

Practice Problem

Consider the commuting times (in minutes) of 20 randomly selected New York workers

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

Draw the boxplot (labeling your plot) for these data and use the 1.5 × IQR rule to determine if 85 minutes is an outlier.

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

Example:

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

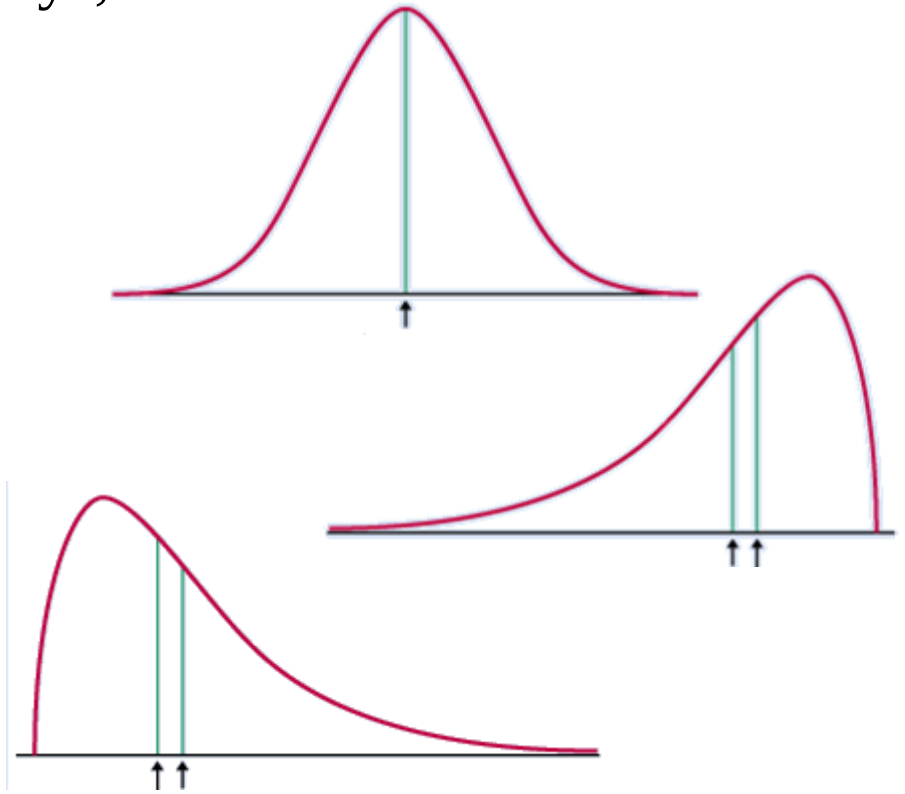
Mean vs. Median

The mean and median measure center in different ways, and both are useful.

The mean and median of a roughly **symmetric** distribution are close together.

If the distribution is exactly **symmetric,** the mean and median are exactly the same.

In a **skewed** distribution, the mean is usually farther out in the long tail than is the median.



2.2 Considering Categorical Data

Numerical summaries- Counts (frequency)

Simplest way to summarize one categorical variable.

Frequency (count): How many observational units are in each category.

- What is the variable?
- What are its levels (possible values for the categorical variable)?
- What is the frequency of the bones in spine?

One-Way Table

Location	Number of Bones
Head/Neck	29
Chest	25
Spine	26
Shoulder/Arms/Heads	64
Hips/Legs/Feet	62
Total	206

Frequency and Relative Frequency

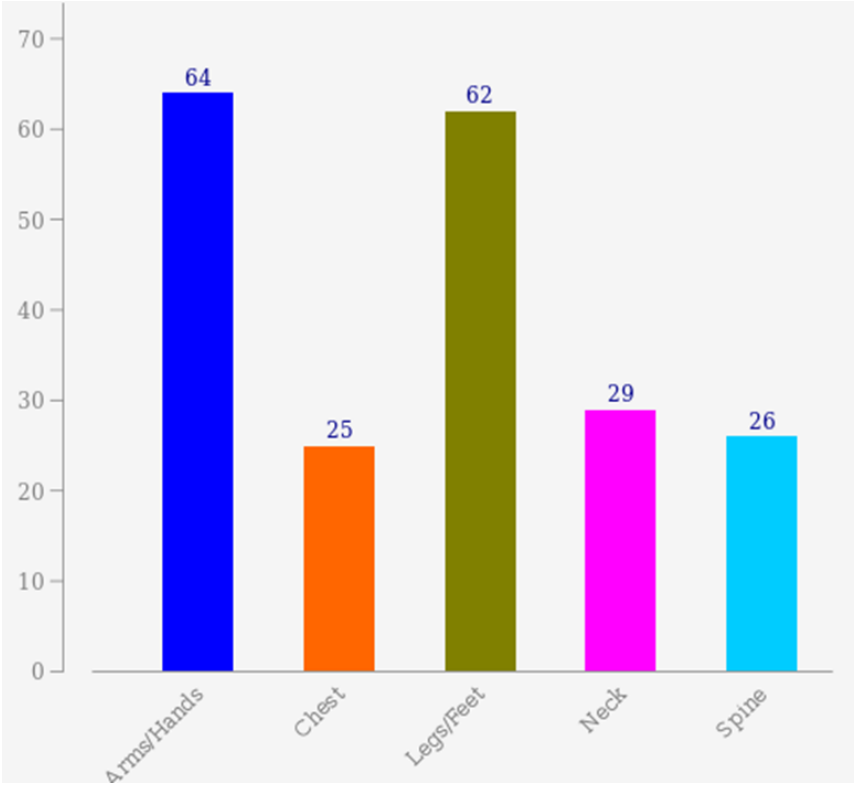
Often, we're not so much interested in how many things there are, but in what percent of the total they are

Relative frequencies: are the frequencies relative to a category total or relative to all data values (Proportion or percentage).

	One-Way Table	
Location	Number of Bones	Relative Frequency
Head/Neck	29	29/206=.14 (14%)
Chest	25	25/206=.12 (12%)
Spine	26	26/206=.13 (13%)
Shoulder/Arms/Heads	64	64/206=.31 (31%)
Hips/Legs/Feet	62	62/206 =.30 (30%)
Total	206	206/206=1.00 (100%)

Visual summaries for categorical Variables

1. *Bar graphs*
   - Useful for summarizing a one or two categorical variables.
   - Very useful when making comparison between two categorical variables.
   - represents the categories as bars whose heights are represented by the count (or percent) of the categories

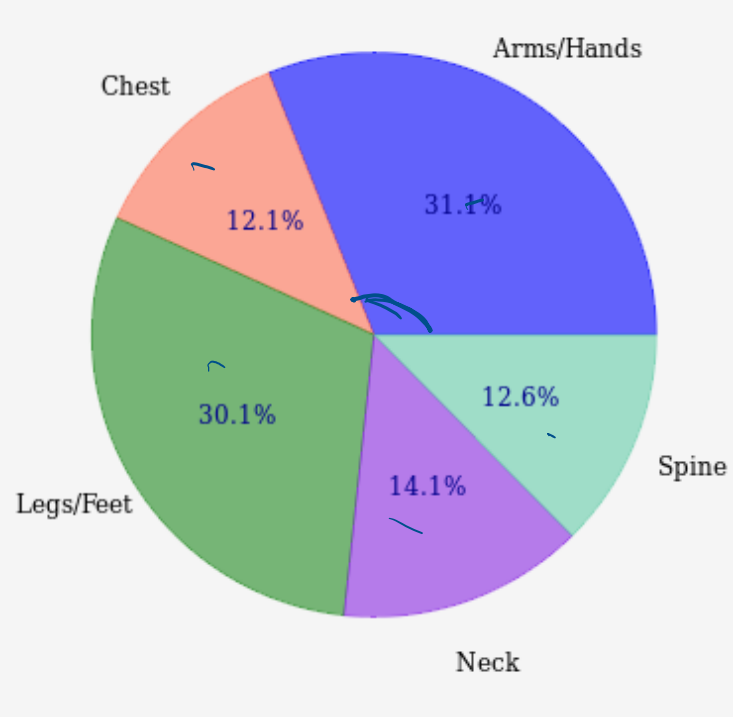


Visual summaries for categorical Variables

The distribution of a categorical variables lists the categories and the count or percent of individuals who fall into each category.

Two simple visual summaries are used for categorical data:

2. *Pie charts*
   - Useful for summarizing a single categorical variable if there are not too many categories.
   - display the variables as a “pie” whose slices are sized by the counts (or percent) of the categories.



Contingency Tables (Two-Way Table)

Consider the question that if there is any relation between the gender of the high school students and how often they use seatbelts while driving. When studying the relationship between two categorical variables, we often display them in a table such as this:

	Gender and Seatbelt Use by 12th Grader While Driving					
	Always	Most times	Sometimes	Rarely	Never	Total
Female	915	276	167	84	25	1467
Male	771	302	247	165	90	1575
Total	1686	578	414	249	115	3042

This table is called *contingency table* because it cover all contingencies for the combinations of the two variables. Because two variables are used to create the table , it is also called *two-way table.*

Contingency Tables (Two-Way Table)

A two-way table gives counts for each combination of values of the two categorical variables.

Gender and Seatbelt Use by 12th Grader While Driving						
	Always	Most times	Sometimes	Rarely	Never	Total
Female	915	276	167	84	25	1467
Male	771	302	247	165	90	1575
Total	1686	578	414	249	115	3042

- *Cell:* is each row category and column category combination in the table.
- When one variable is designated as the explanatory and the other variable as the response, it is customary to define:
  - Row : explanatory variable
  - Column: response variable

Row and Column Proportions

- Row percentages are the percentages within a given row in the contingency table. Row percentages are based on the total number of observations in the row.

	Gender and Seatbelt Use by 12th Grader While Driving					
	Always	Most times	Sometimes	Rarely	Never	Total
Female	.624	.188	.114	.057	.017	1.000
Male	.489	.192	.157	.105	.057	1.000
Total	.554	.190	.136	.082	.038	1.000

- *Column percentages:* are the percentages within a given column of the contingency table. Column percentages are based on the total number of observations in the column.

	Gender and Seatbelt Use by 12th Grader While Driving					
	Always	Most times	Sometimes	Rarely	Never	Total
Female	.543	.478	.403	.337	.217	.482
Male	.457	.522	.597	.663	.783	.518
Total	1.000	1.000	1.000	1.000	1.000	1.000

Example

a) What does 0.624 represent in the row proportion table?

b) What does 0.543 represent in the column proportion table?

Row percentages

	Gender and Seatbelt Use by 12th Grader While Driving					
	Always	Most times	Sometimes	Rarely	Never	Total
Female	.624	.188	.114	.057	.017	1.000
Male	.489	.192	.157	.105	.057	1.000
Total	.554	.190	.136	.082	.038	1.000

Column percentages

	Gender and Seatbelt Use by 12th Grader While Driving					
	Always	Most times	Sometimes	Rarely	Never	Total
Female	.543	.478	.403	.337	.217	.482
Male	.457	.522	.597	.663	.783	.518
Total	1.000	1.000	1.000	1.000	1.000	1.000

Example

(a) Compare the education level of White Orange County residents to their Asian counterpart, as measured by the proportion who have received a bachelor's degree or higher. Which group would you say is more educated? Explain.

(b) If you randomly select an Orange County resident, what are the chances that this person is African-American?

(d) What is the relative frequency of non-white Orange County residents? Your calculated relative frequency is relative to what population?

Table 2.2: Education by race in Orange County, California

Sex By Educational Attainment For The Population 25 Years And Over

California: Orange County

	Less than HS	HS/GED/alt	Some college	BS or higher	Total
White	151,483	233,044	407,794	476,181	1,268,502
Black/African-American	2,281	5,889	12,451	11,435	32,056
Asian	50,068	51,432	82,312	182,963	366,775
Hispanic/Latino	225,563	124,882	114,067	65,401	529,913
Other	120,580	68,746	64,114	36,669	290,109
Total	549,975	483,993	680,738	772,649	2,487,355

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.

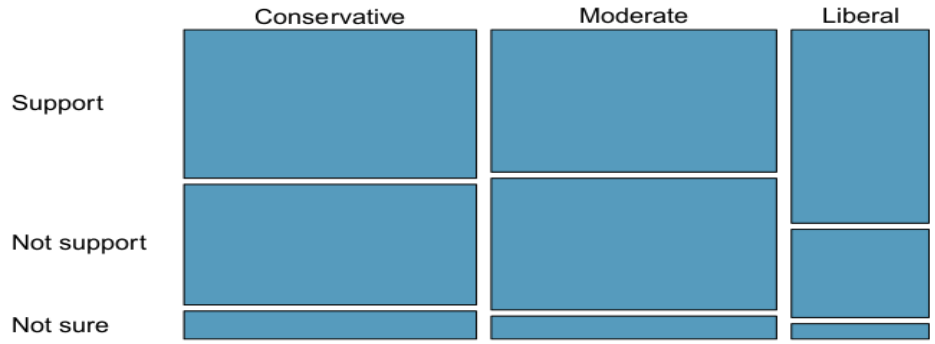
Mosaic Plots

A *mosaic plot* is a visualization technique suitable for contingency tables that resembles a standardized stacked bar plot with the benefit that we still see the relative group sizes of the primary variable as well.

Example:

A random sample of registered voters from Tampa, FL were asked if they support the DREAM Act, a proposed law which would provide a path to citizenship for people brought illegally to the US as children. The survey also collected information on the political ideology of the respondents. Based on the mosaic plot shown below, do views on the DREAM Act and political ideology appear to be independent? Explain your reasoning.

Views on the DREAM Act and political affiliation appear to be dependent or independent?



Example

Does there appear to be relation 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