OpenIntro

Chapter 2: Summarizing data

OpenIntro Statistics, 4th Edition

Slides developed by Mine Çetinkaya-Rundel of OpenIntro.

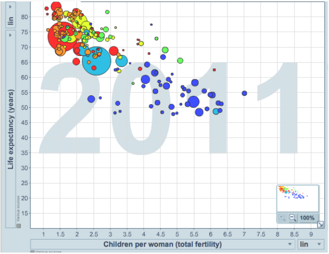
Examining numerical data

Scatterplot

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

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

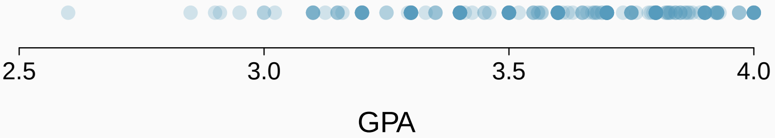
Was the relationship the same throughout the years, or did it change?



http://www.gapminder.org/world

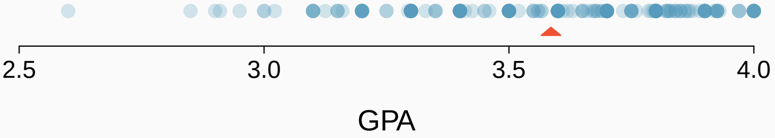
Dot plots

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

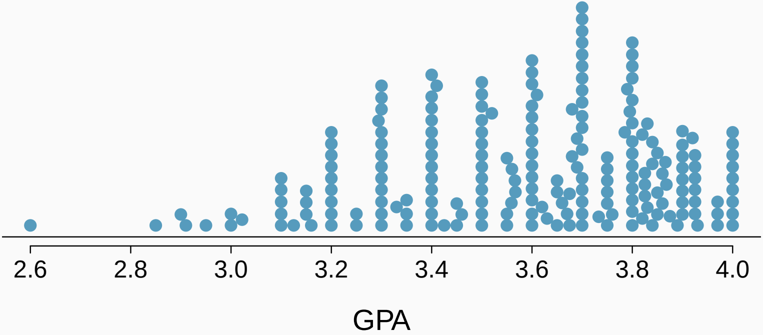
- The *sample mean*, denoted as $\bar{x}$, can be calculated as

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

  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.

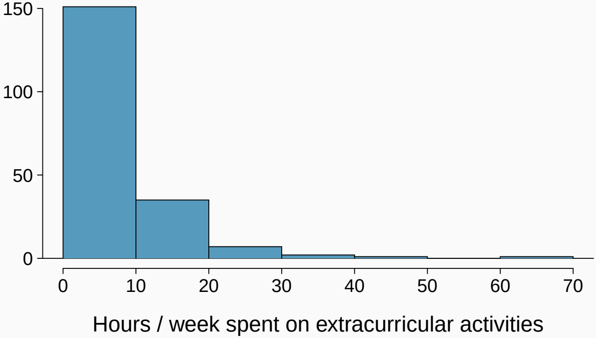
Stacked dot plot

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



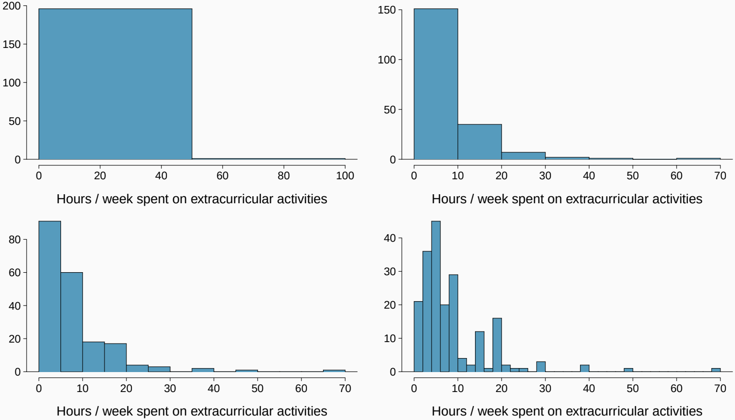
Histograms - Extracurricular hours

- 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.



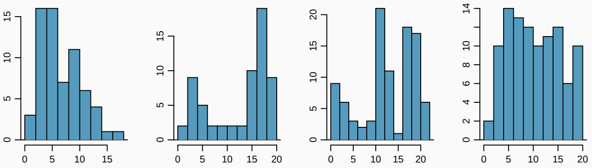
Bin width

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



Shape of a distribution: modality

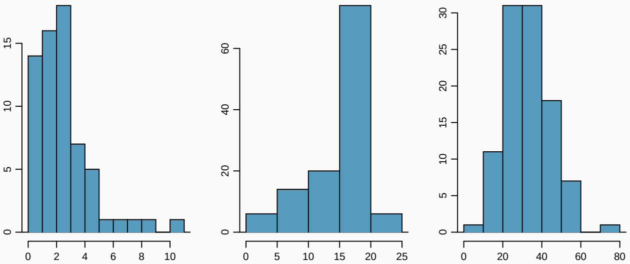
Does the histogram have a single prominent peak (*unimodal*), several prominent peaks (*bimodal/multimodal*), or no apparent peaks (*uniform*)?



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

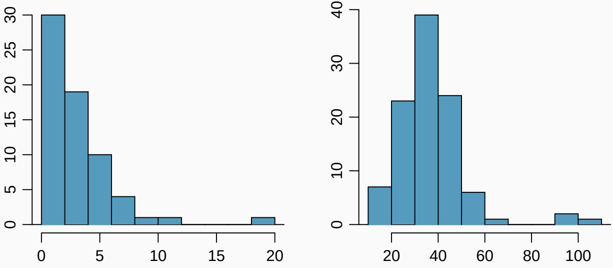
Is the histogram *right skewed*, *left skewed*, or *symmetric*?



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

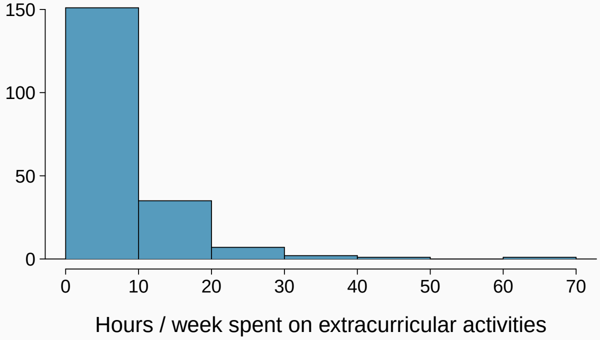
Shape of a distribution: unusual observations

Are there any unusual observations or potential *outliers*?



Extracurricular activities

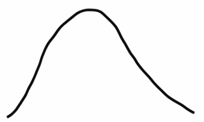
How would you describe the shape of the distribution of hours per week students spend on extracurricular activities?



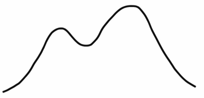
Commonly observed shapes of distributions

- modality

unimodal

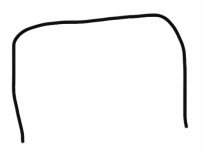


bimodal



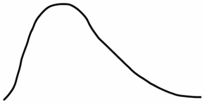
multimodal

uniform

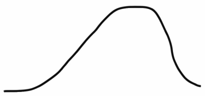


- skewness

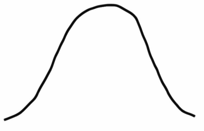
right skew



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)

Are you typical?

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

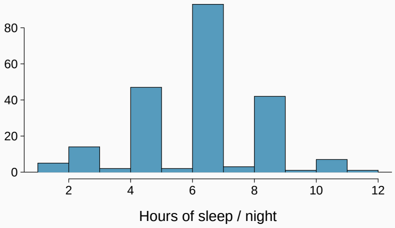
How useful are centers alone for conveying the true characteristics of a distribution?

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:



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

Variance (cont.)

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

Standard deviation

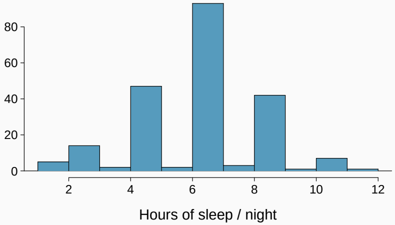
The *standard deviation* is the square root of the variance, and has the same units as the data.s

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

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

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

- 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*.

$$IQR = Q3 - Q1$$

How to calculate the sample quartiles?

1. Arrange the data set in order of magnitude from smallest to largest.
2. Calculate the quartile positions:
   - Position of Q_1: $.25(n + 1)$
   - Position of Q_3: $.75(n + 1)$
3. If the positions are integers, then Q_1 and Q_3 are the values in the ordered data set found in those positions.
4. If the positions in step 2 are not integers, find the two measurements in positions just above and just below the calculated position. Calculate the quartile by finding a value either one-fourth, one-half, or three-fourths of the way between these two measurements.

Example

Let's compute the quartiles of the following data set representing the heights (in cm) of 12 plants:

$$12.3, 4.1, 16.2, 8.9, 14.7, 2.5, 10.6, 7.8, 9.1, 15.4, 17.6, 6.3$$

Step 1: Order the Data and Compute Positions

First, we order the data from smallest to largest:

$$2.5, 4.1, 6.3, 7.8, 8.9, 9.1, 10.6, 12.3, 14.7, 15.4, 16.2, 17.6$$

Then, we calculate the positions for Q1 and Q3 as $0.25(12+1) = 3.25$ and $0.75(12+1) = 9.75$ respectively.

Step 2: Compute the Quartiles

1:	2:	3:	4:	5:	6:	7:	8:	9:	10:	11:	12:
2.5	4.1	6.3	7.8	8.9	9.1	10.6	12.3	14.7	15.4	16.2	17.6

We interpolate to find the values at these positions:

- For Q1, the positions 3 and 4 are the closest integers to 3.25. The values at these positions are 6.3 and 7.8. So, Q1 = $6.3 + 0.25 * (7.8 – 6.3) = 6.675$.
- For Q3, the positions 9 and 10 are the closest integers to 9.75. The values at these positions are 14.7 and 15.4. So, Q3 = $14.7 + 0.75 * (15.4 – 14.7) = 15.225$.

The median Q2 is the average of the values at positions $12/2 = 6$ and $(12/2 + 1 = 7$, which are 9.1 and 10.6 in our case. So, Q2 = $(9.1 + 10.6)/2 = 9.85$.

Result

So, the quartiles of our data set are:

- Q1 = 6.675
- Q2 = 9.85
- Q3 = 15.225

Data Set

Here's the data set we are working with:

$$12.3, 4.1, 16.2, 8.9, 14.7, 2.5, 10.6, 7.8, 9.1, 15.4, 17.6, 6.3$$

```
# Store the data in a vector
data <- c(12.3, 4.1, 16.2, 8.9, 14.7, 2.5, 10.6, 7.8,
9.1, 15.4, 17.6, 6.3)
```

Computing Mean and Median in R

Here's how you can compute the mean and median for the data set using R:

```
# Compute the mean
mean_value <- mean(data)
mean_value

# Compute the median
median_value <- median(data)
median_value
```

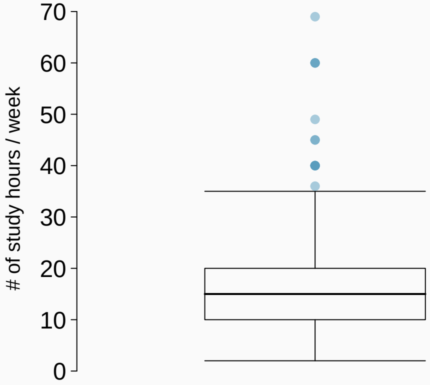
Computing Quartiles in R (Type 6)

Here's how you can compute the quartiles for the data set using R with ‘type = 6‘:

```
# Compute the quartiles using type 6
quartiles <- quantile(data, type = 6)
quartiles
```

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.

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

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

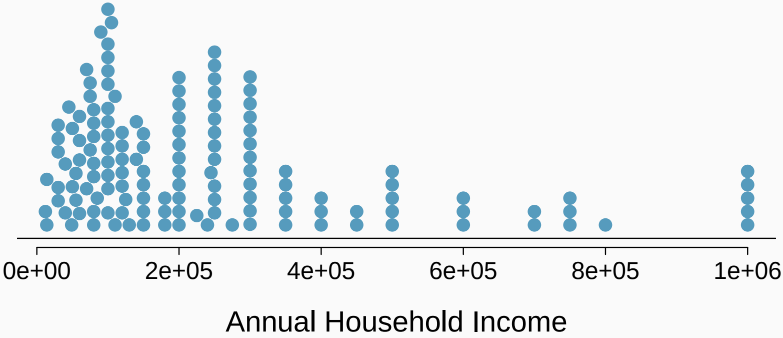
- A potential *outlier* is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

Outliers (cont.)

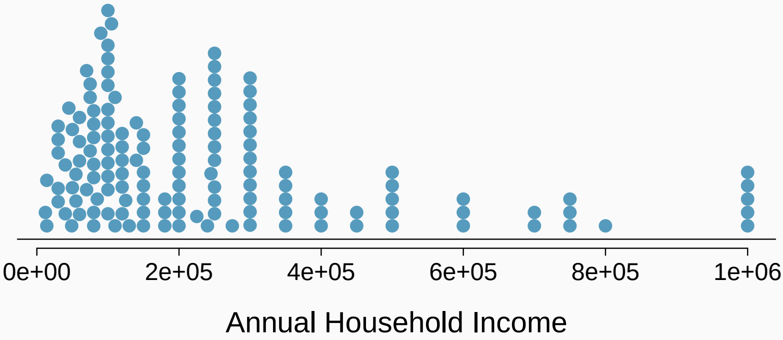
Why is it important to look for outliers?

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

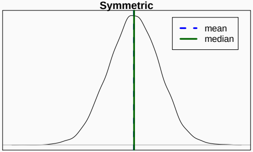
Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

- 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

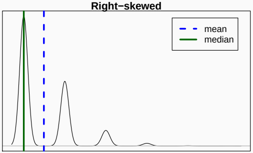
If you would like to estimate the typical household income for a student, would you be more interested in the mean or median income?

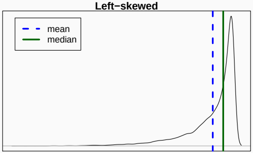
Mean vs. median

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



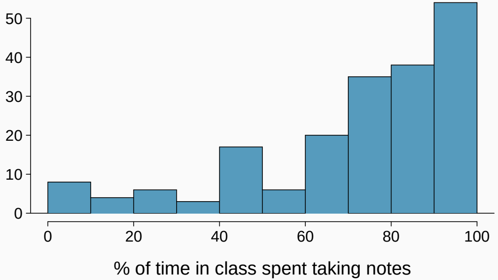
- If the distribution is skewed or has extreme outliers, center is often defined as the median
  - Right-skewed: mean $>$ median
  - Left-skewed: mean $<$ median





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

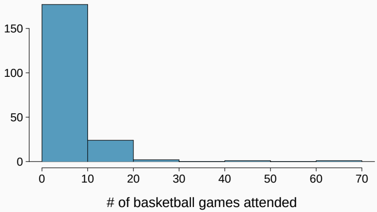
(c) mean ≈ median

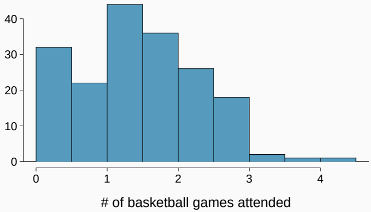
(d) impossible to tell

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.

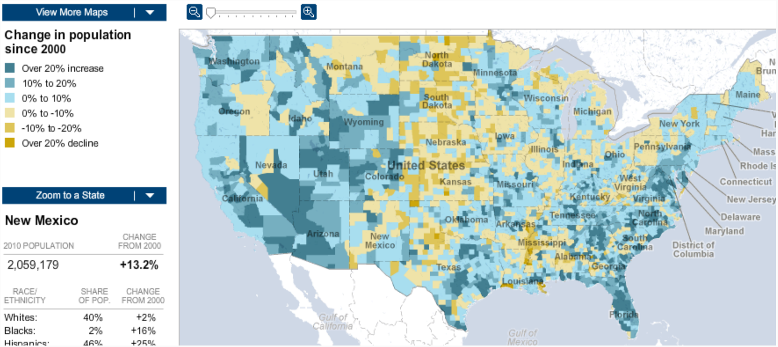
# of games	70	50	25	$\cdots$
log(# of games)	4.25	3.91	3.22	$\cdots$

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

What other variables would you expect to be extremely skewed?

Intensity maps

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



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

Considering categorical data

Contingency tables

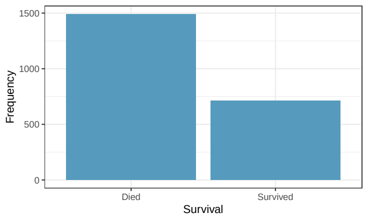
A table that summarizes data for two categorical variables is called a *contingency table*.

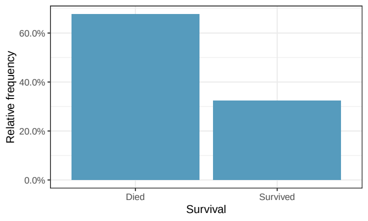
The contingency table below shows the distribution of survival and ages of passengers on the Titanic.

		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*.





How are bar plots different than histograms?

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

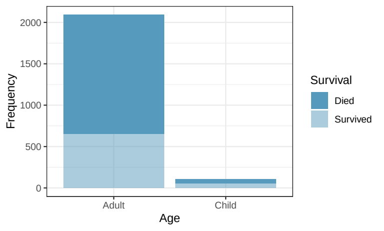
To answer this question we examine the row proportions:

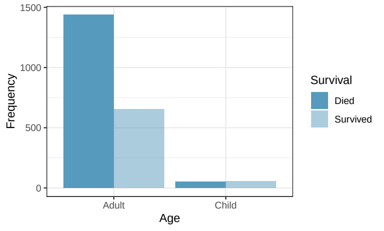
- % Adults who survived: 654 / 2092 $\approx$ 0.31
- % Children who survived: 57 / 109 $\approx$ 0.52

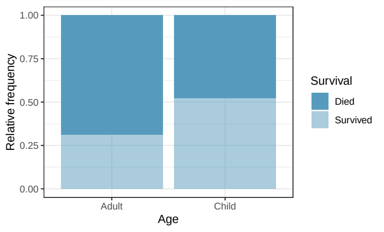
Bar plots with two variables

- *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?

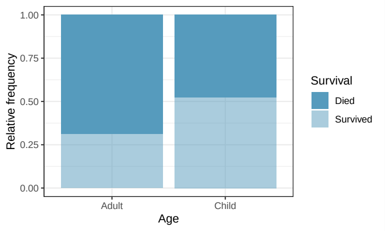


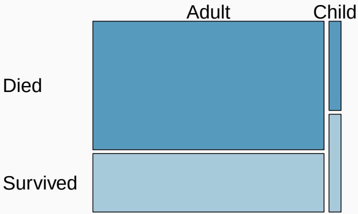




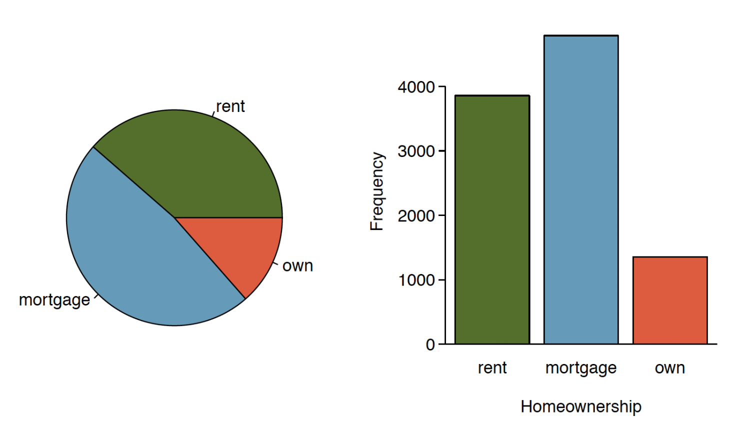
Mosaic plots

What is the difference between the two visualizations shown below?



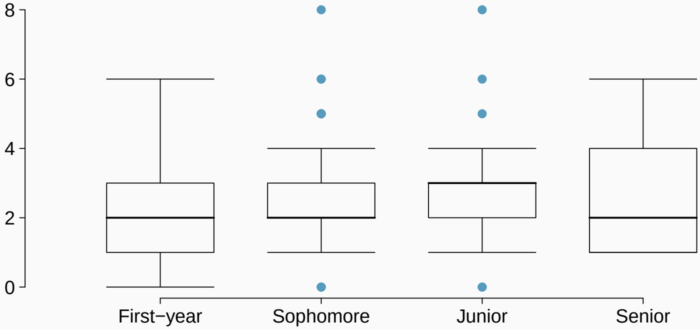


Pie chart vs bar plot

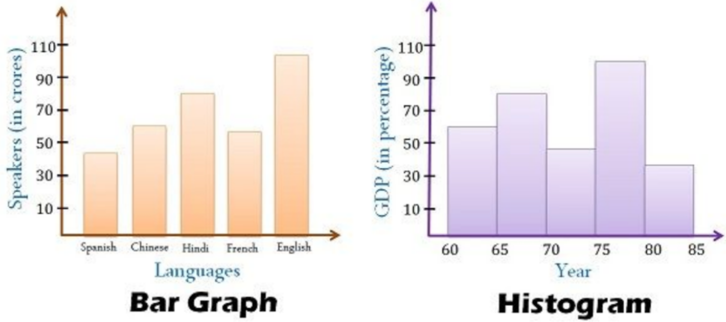


Side-by-side box plots

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



Bar plot (bar graph) versus histogram



The fundamental difference between histogram and bar graph will help you to identify the two easily is that there are gaps between bars in a bar graph but in the histogram, the bars are adjacent to each other.

Bar plot (bar graph) versus histogram

BASIS FOR COMPARISON	HISTOGRAM	BAR GRAPH
Meaning	Histogram refers to a graphical representation, that displays data by way of bars to show the frequency of numerical data.	Bar graph is a pictorial representation of data that uses bars to compare different categories of data.
Presents	Quantitative data	Categorical data
Spaces	Bars touch each other, hence there are no spaces between bars	Bars do not touch each other, hence there are spaces between bars.
Elements	Elements are grouped together, so that they are considered as ranges.	Elements are taken as individual entities.
Can bars be reordered?	No	Yes
Width of bars	Need not to be same	Same