Chapter 2: Summarizing data

Yonghyun Kwon

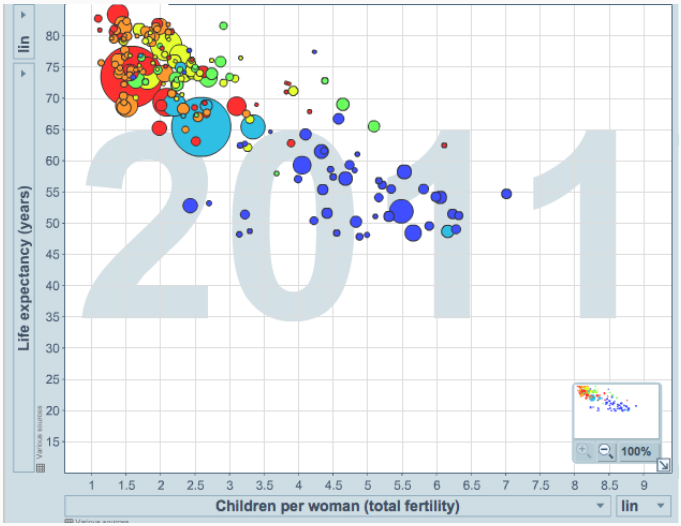
Department of Mathematics, Korea Military Academy

Examining numerical data

Scatterplot

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

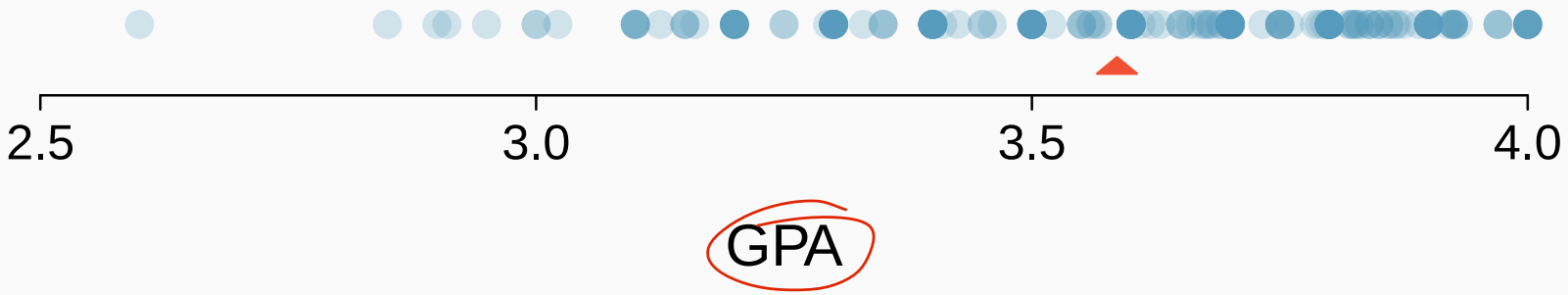
Do life expectancy and total fertility appear to be *associated* or *independent*?



http://www.gapminder.org/world

Mean

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



- 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(cont.)

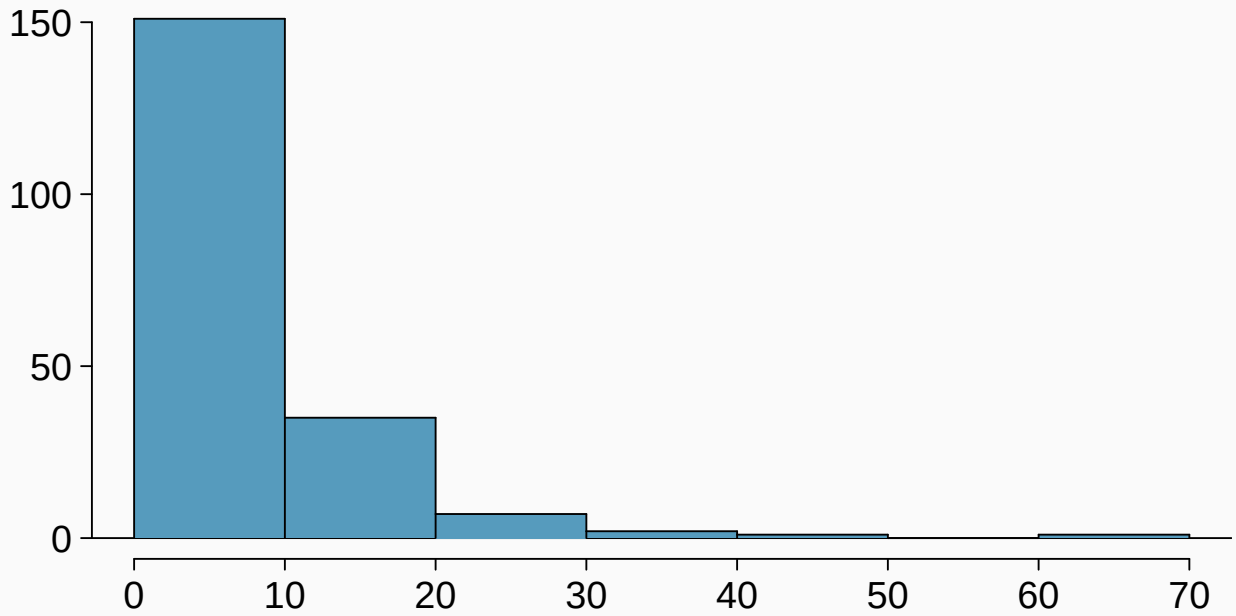
- The *sample mean*, denoted as $\bar{x}$, can be calculated as

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

  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.

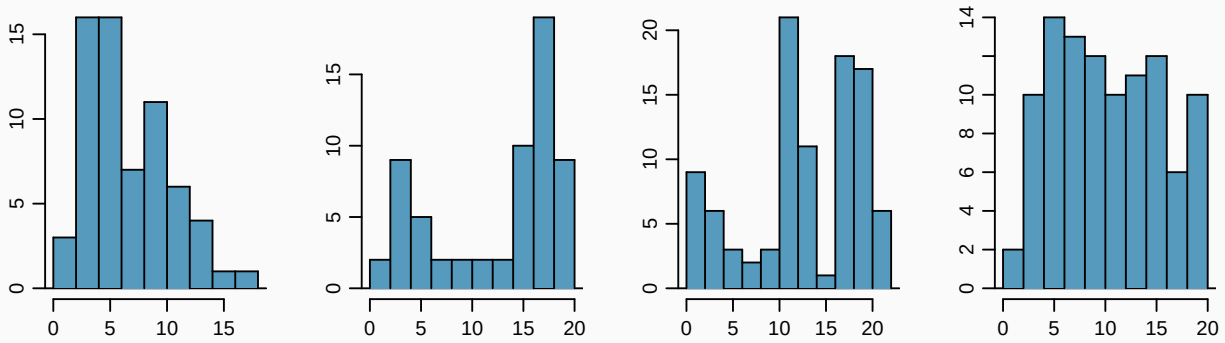
Histograms - Extracurricular hours

- *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.



Shape of a distribution: modality

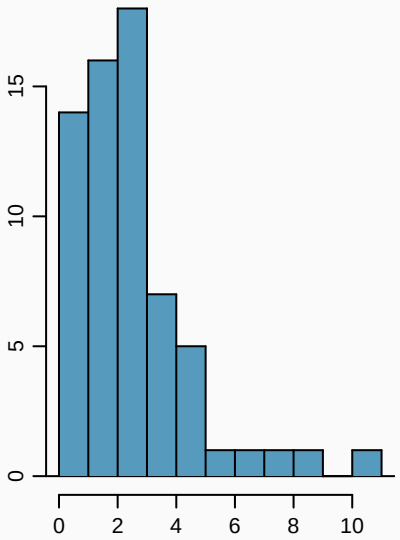
Does the histogram have a single prominent peak (*unimodal*), several prominent peaks (*bimodal/multimodal*), or no apparent peaks (*uniform*)?

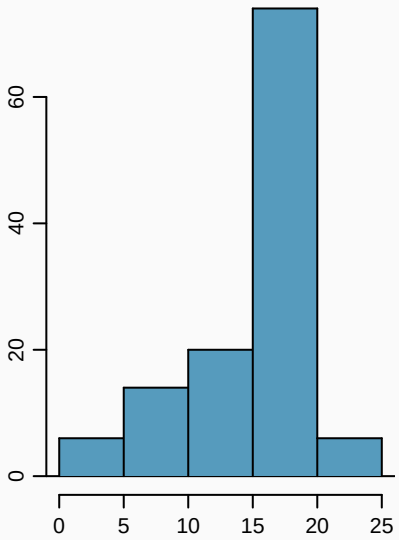


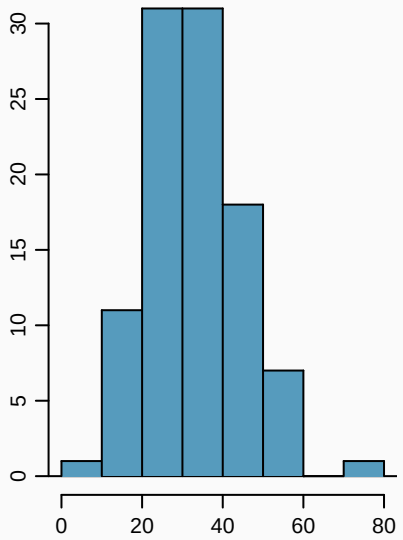
Note: In order to determine modality, step back and imagine a smooth curve over the histogram – imagine that the bars are wooden blocks and you drop a limp spaghetti over them, the shape the spaghetti would take could be viewed as a smooth curve.

Shape of a distribution: skewness

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







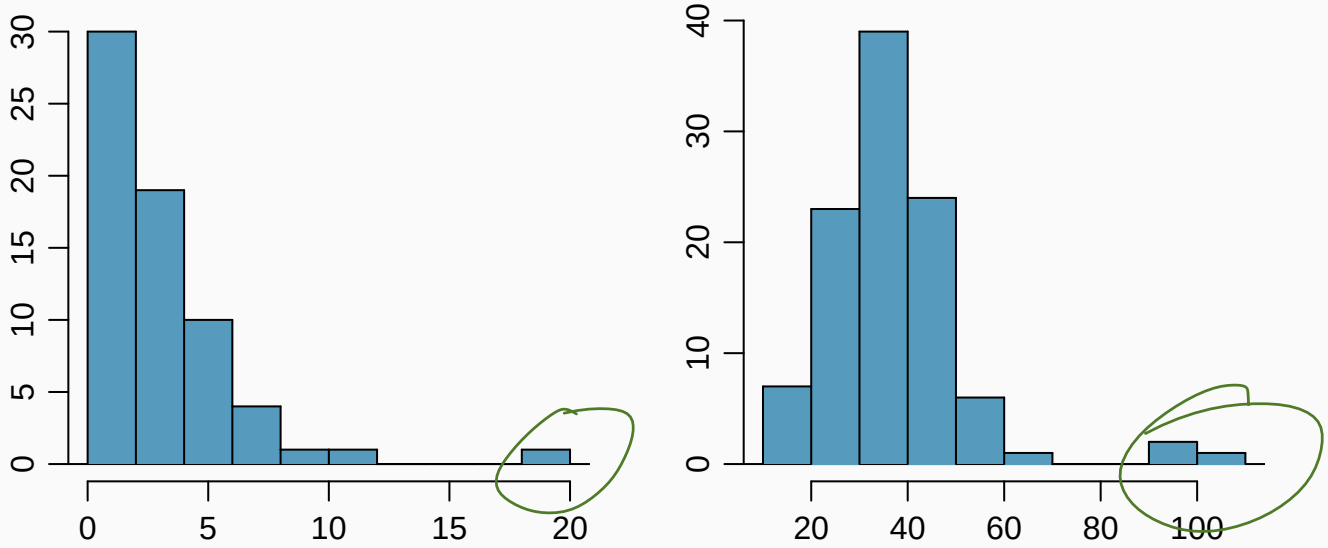
→ right-skewed

→ symmetric.

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

Shape of a distribution: unusual observations

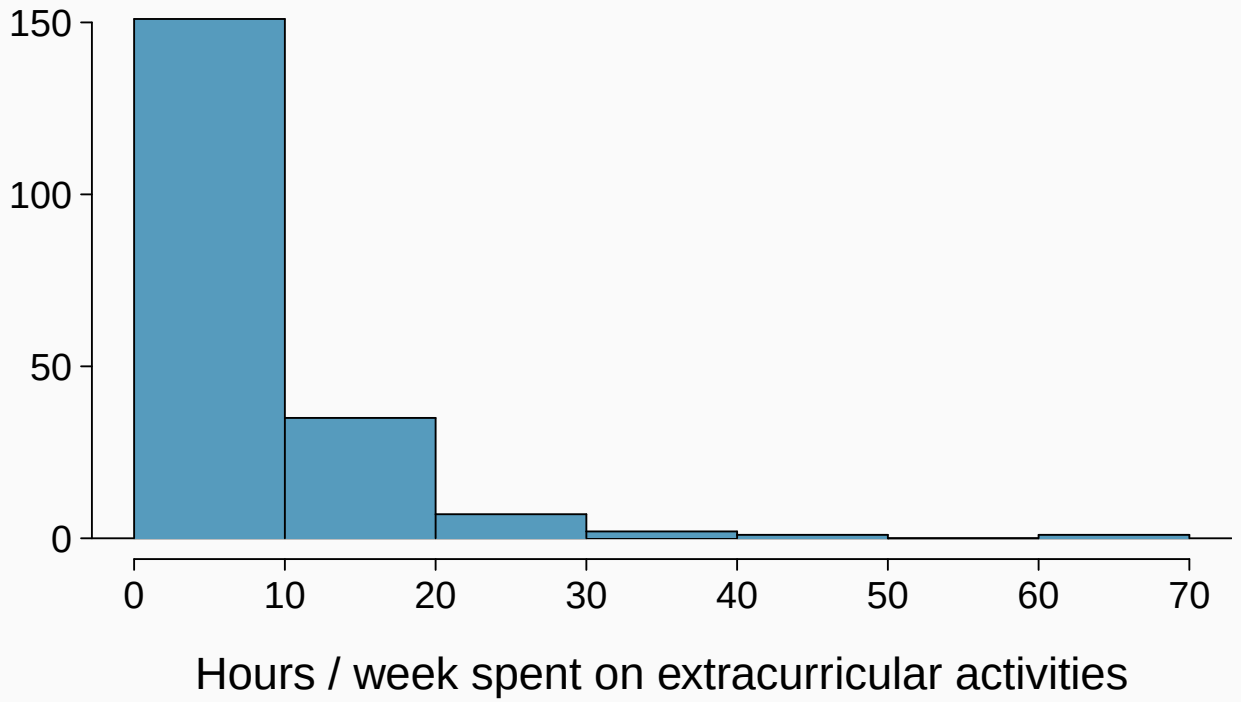
Are there any unusual observations or potential *outliers*?



Extracurricular activities

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

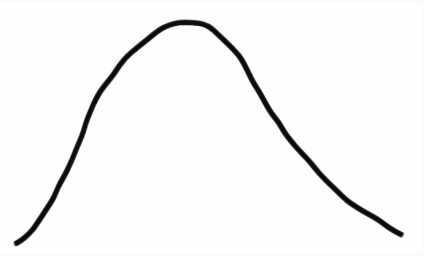
right-skewed



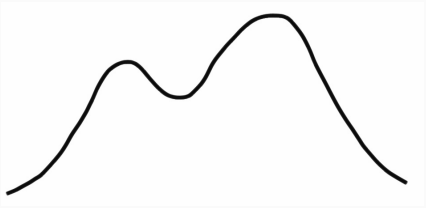
Commonly observed shapes of distributions

- modality

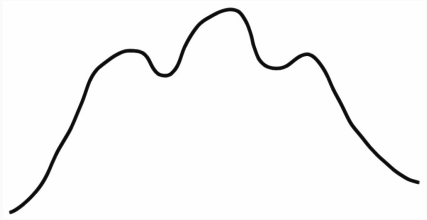
unimodal



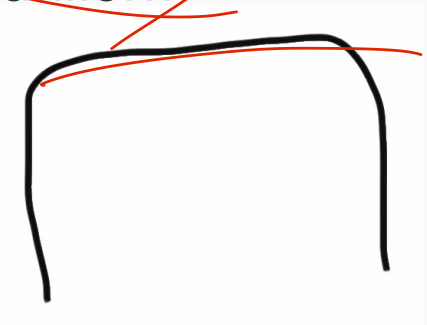
bimodal



multimodal

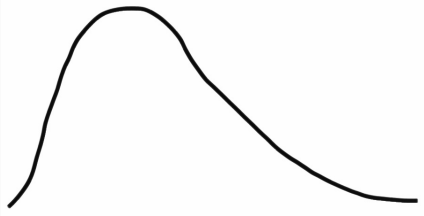


uniform

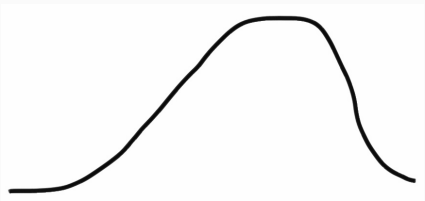


- skewness

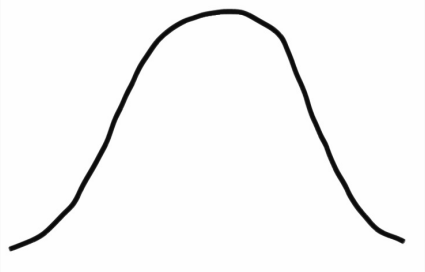
right skew



left skew



symmetric



Practice

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

(a) weights of adult females  ⇒ symmetric

(b) salaries of a random sample of people from North Carolina ⇒ right-skewed

(c) house prices

(d) birthdays of classmates (day of the month) ⇒ uniformly distributed

Variance

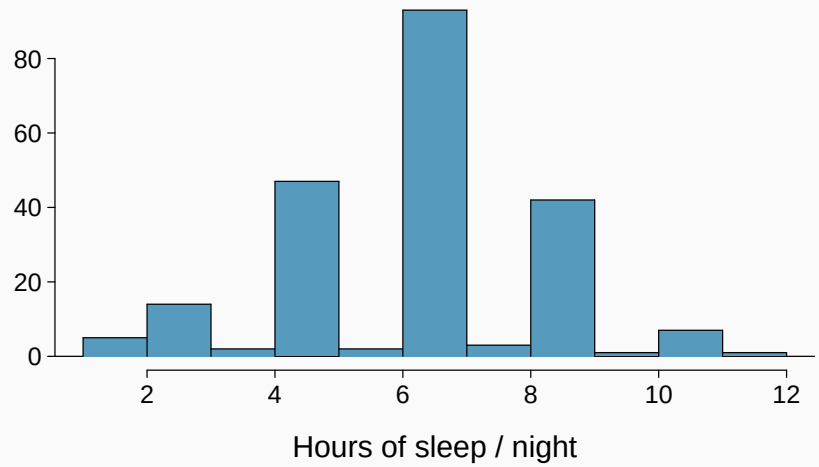
Variance is roughly the average squared deviation from the mean.

Sample variance

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

$x_i - \bar{x}$: deviation

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



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

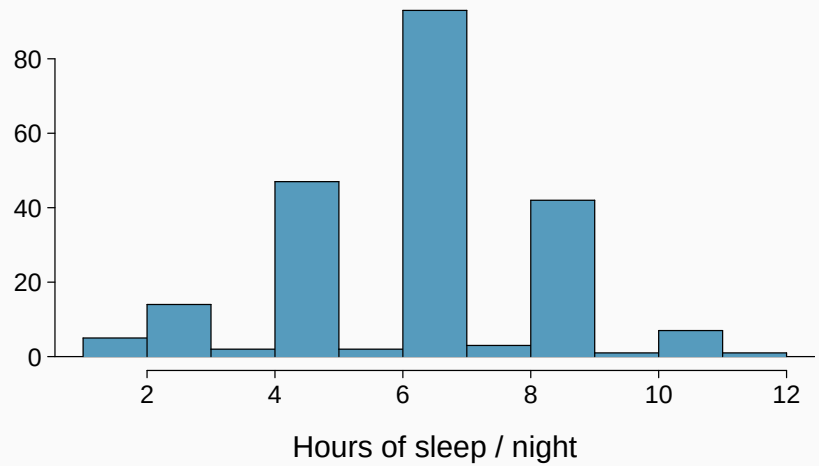
Standard deviation

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:

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



Median

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

$$0, 1, \underline{2}, 3, 5$$

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

$$0, 1, \underline{2, 3}, 4, 6 \rightarrow \frac{2+3}{2} = 2.5$$

- 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

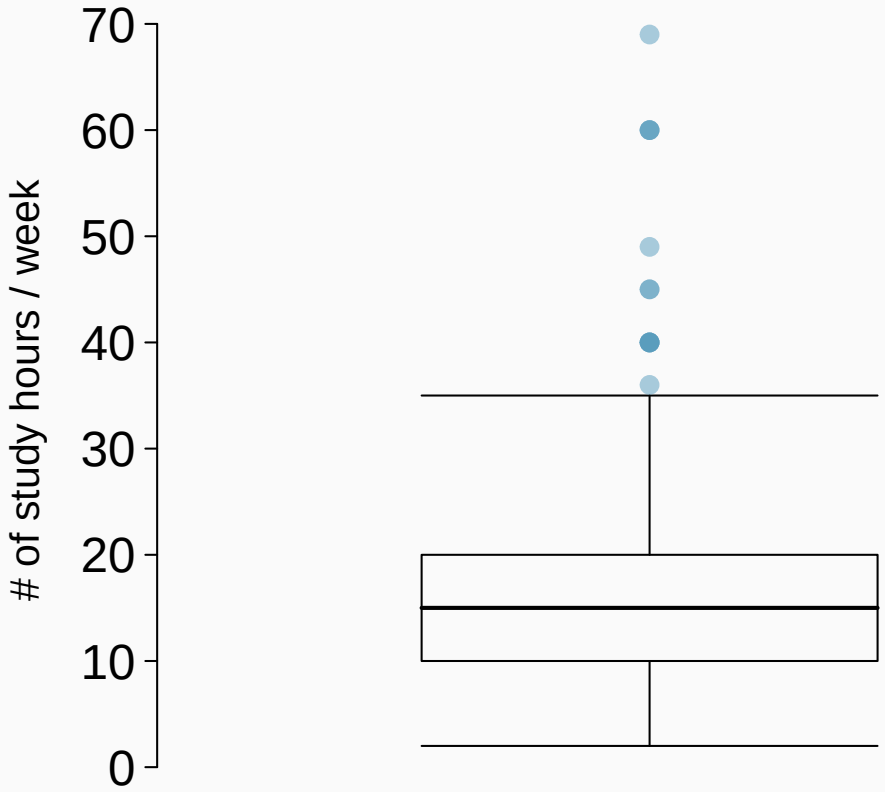
- The 25^{th} percentile is also called the first quartile, *Q1*.
- The 50^{th} percentile is also called the median, *Q2*.
- The 75^{th} percentile is also called the third quartile, *Q3*.
- Between Q1 and Q3 is the middle 50% of the data. The range these data span is called the *interquartile range*, or the *IQR*.

Interquartile range(IQR)

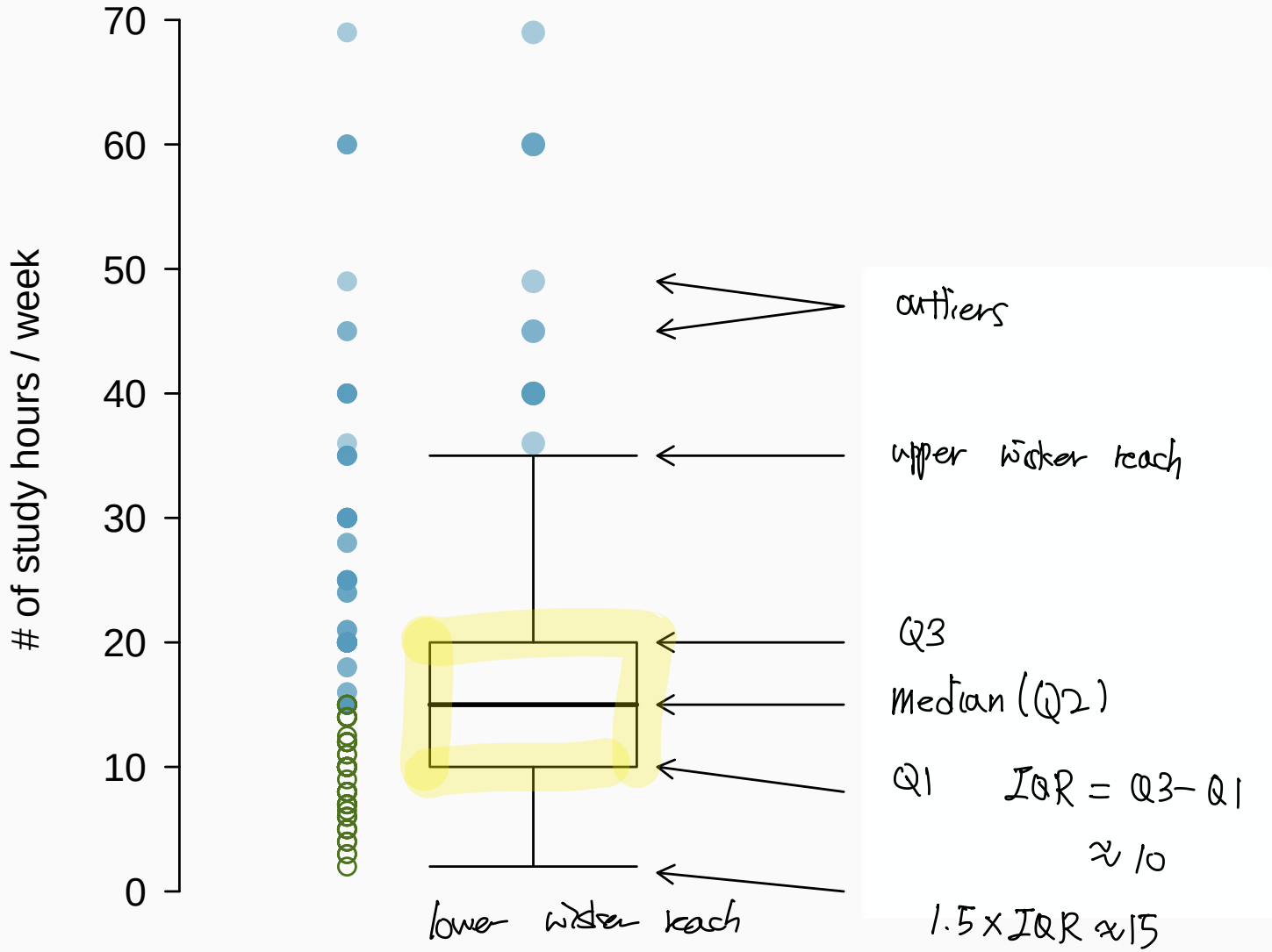
$$IQR = Q3 - Q1$$

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.

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

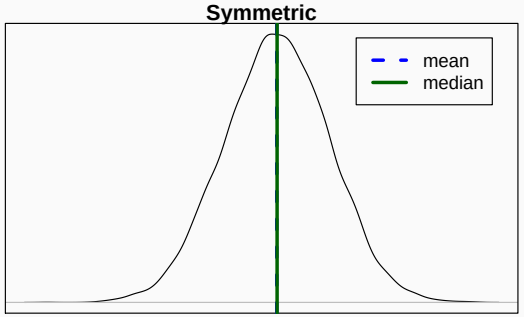
$$Q3 + 1.5 \times IQR = 20 + 1.5 \times 10 = 35$$

$$Q1 - 1.5 \times IQR = 10 - 1.5 \times 10 = -5$$

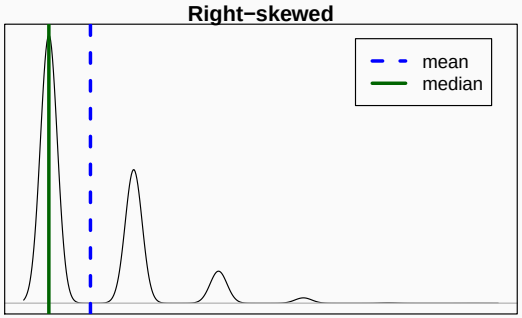
- A potential *outlier* is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

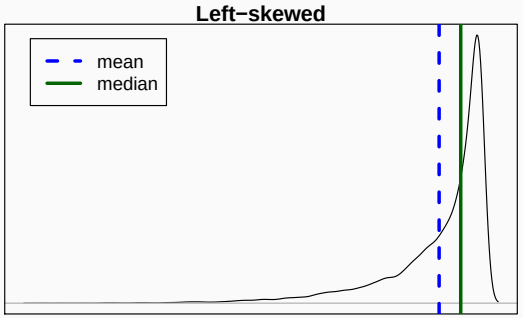
Mean vs. median

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



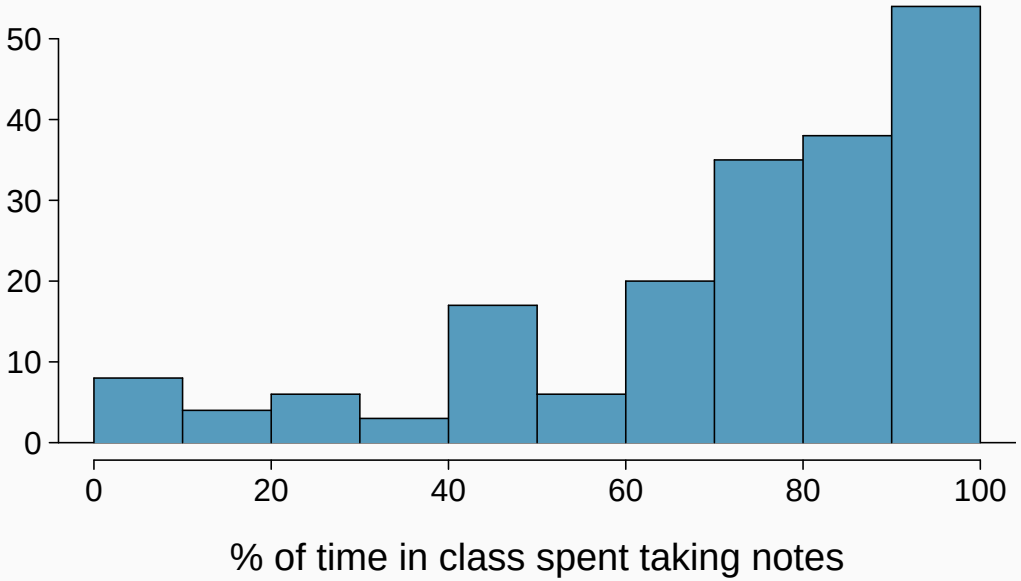
- If the distribution is skewed or has extreme outliers, center is often defined as the median
  - Right-skewed: mean > median
  - Left-skewed: mean < median





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

Considering categorical data

Contingency tables

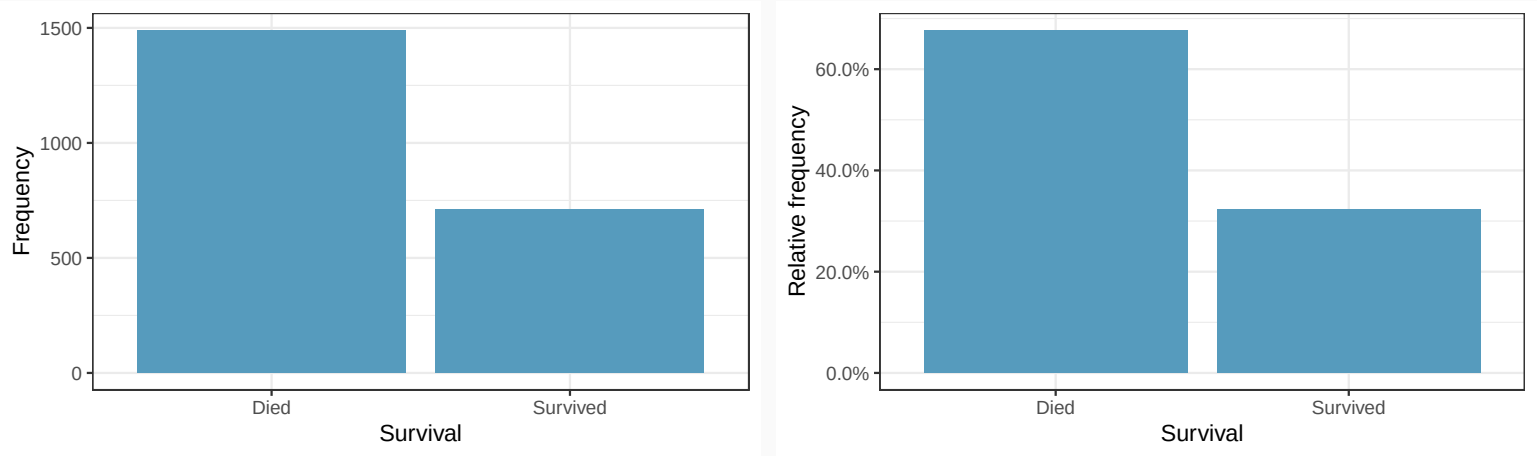
A table that summarizes data for two categorical variables is called a *contingency table*.

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*.



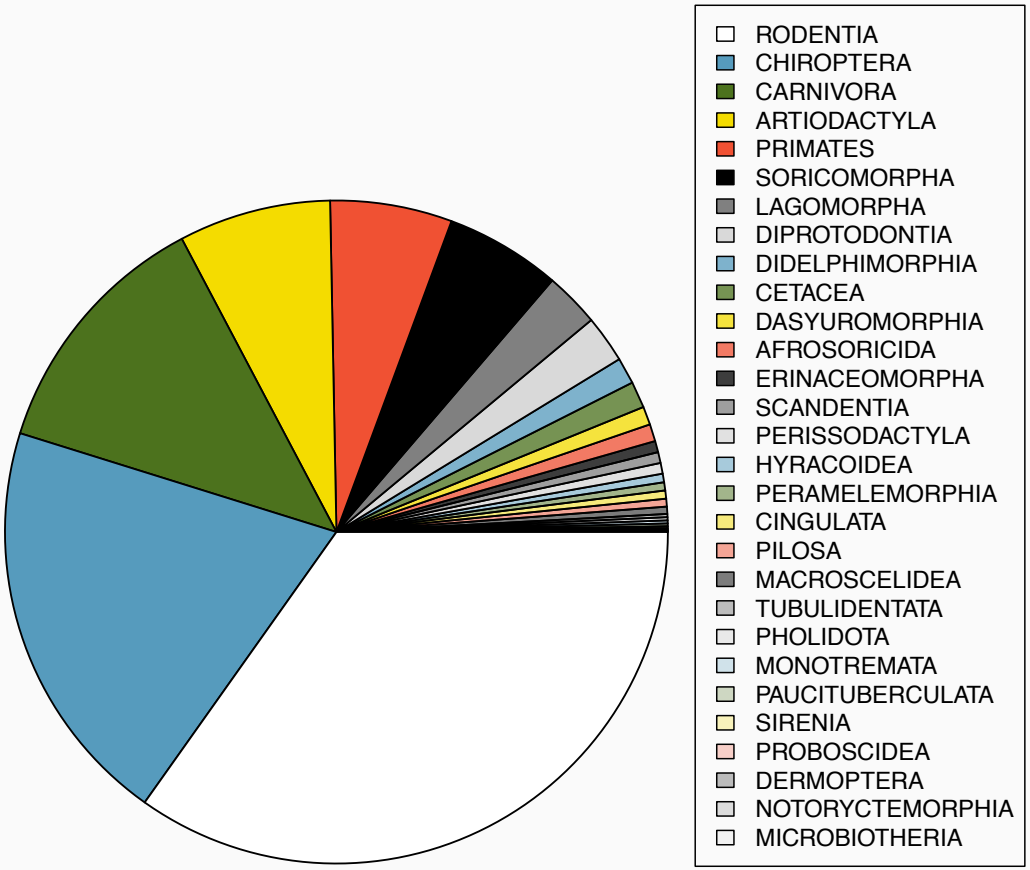
How are bar plots different than histograms?

⇒ categorical

⇒ continuous

Pie charts

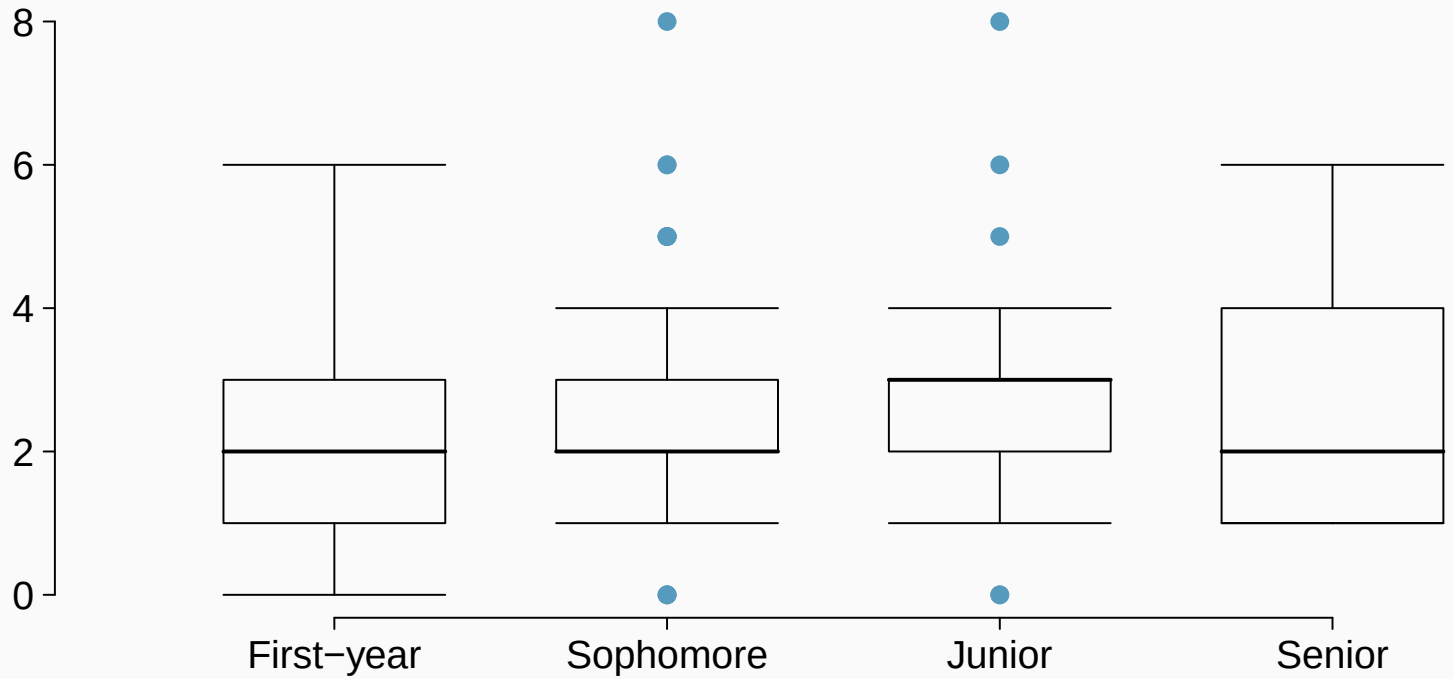
Can you tell which order encompasses the lowest percentage of mammal species?



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?



Exercises

Exercises in OpenIntro Statistics 4th ed.

- 2.9 (c) and (d)
- 2.10
- 2.33