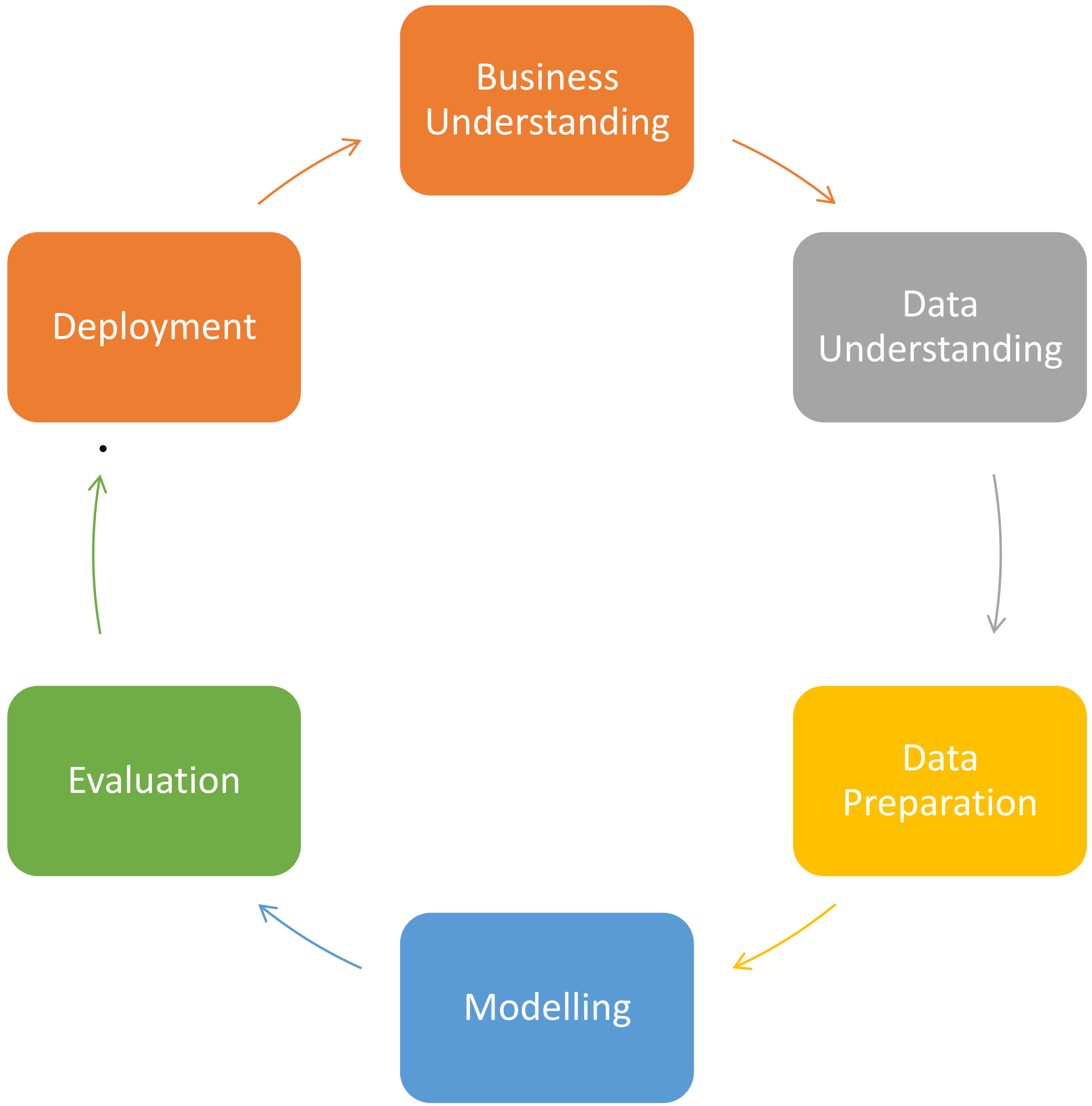
MAKING SENSE OF DATA

Overview:

- Fundamental Statistics
- Basic Data Visualisation
- Distributions of Data
- Transformations

THE PROCESS



CRISP-DM

OBJECTIVES

- Import Numpy and Pandas
- Understand Numpy Arrays
- Create Pandas Data Frames
- Carry out Basic Operations on Data Frames

FUNDAMENTALS OF STATISTICS

Visualization & Exploration

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 students' genders and whether or not they are looking for a spouse while in college.

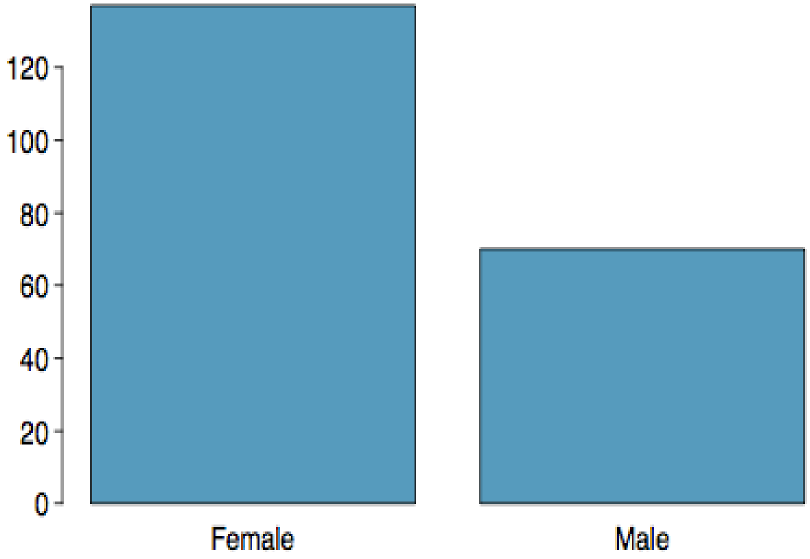
		looking for spouse		
		No	Yes	Total
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

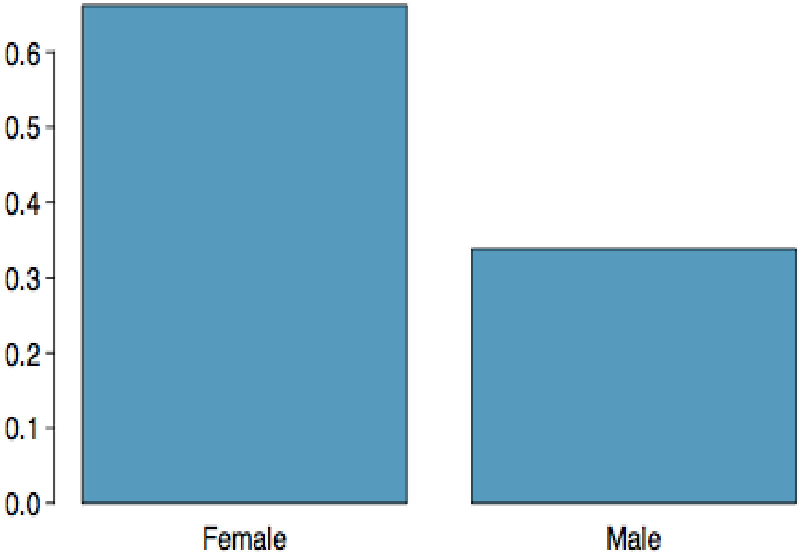
BAR PLOTS

How are bar plots different than histograms?

Bar plots are used for displaying distributions of categorical variables, while 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, while in a bar plot the categories can be listed in any order (though some orderings make more sense than others, especially for ordinal variables.)

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.





CHOOSING THE APPROPRIATE PROPORTION

Does there appear to be a relationship between gender and whether the student is looking for a spouse in college?

To answer this question we examine the row proportions:

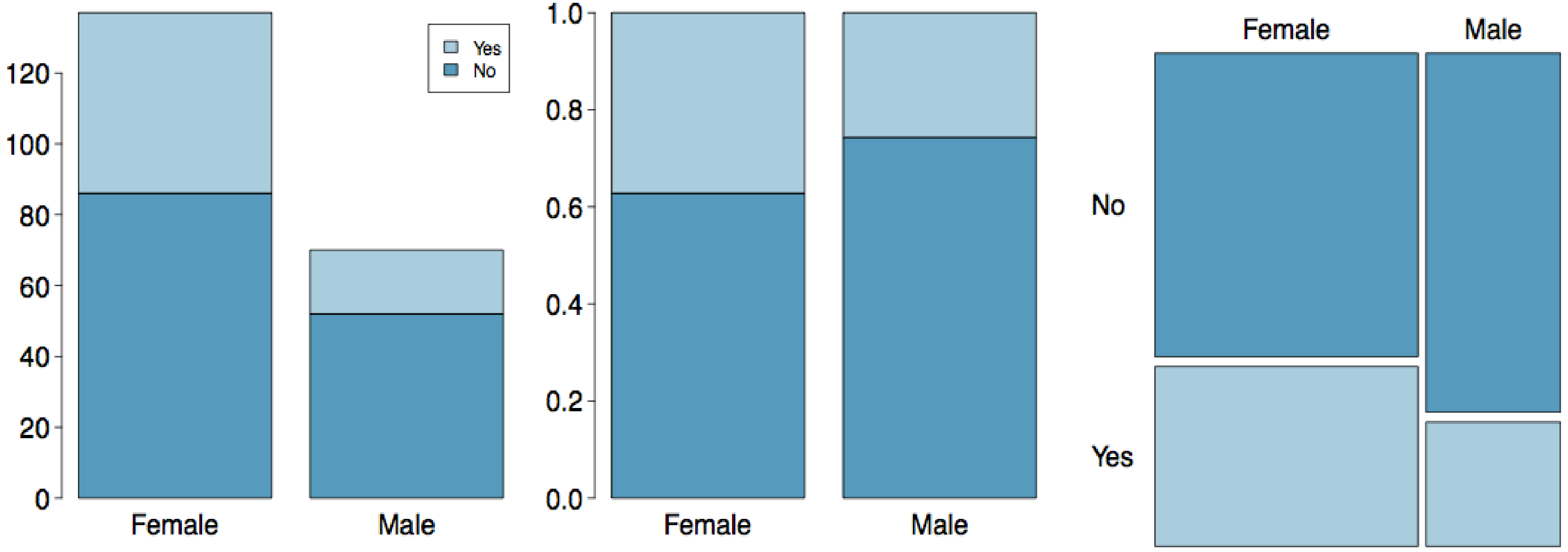
% Females looking for a spouse: 51 / 137 ~ 0.37

% Males looking for a spouse: 18 / 70 ~ 0.26

		looking for spouse		
		No	Yes	Total
gender	Female	86	51	137
	Male	52	18	70
	Total	138	69	207

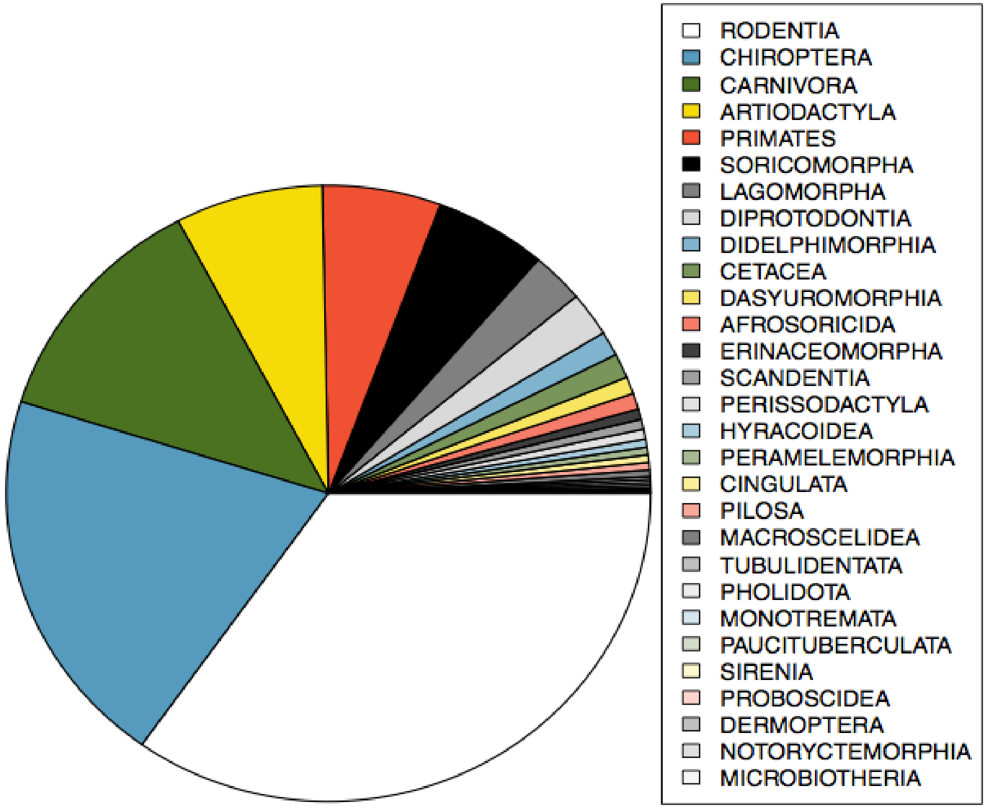
SEGMENTED BAR AND MOSAIC PLOTS

What are the differences between the three visualizations shown below?



PIE CHARTS

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



http://www.bucknell.edu/msw3

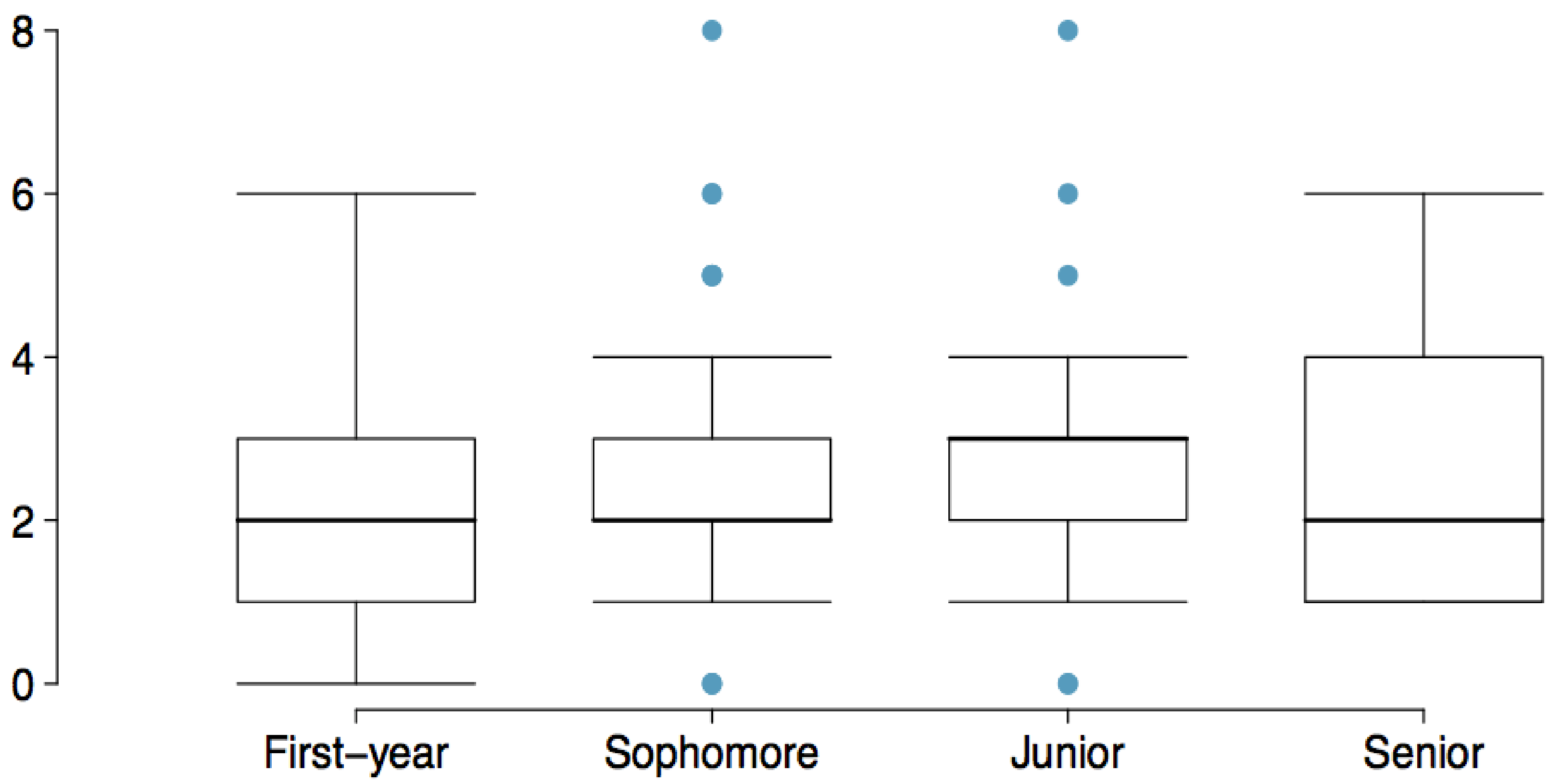
FUNDAMENTALS OF STATISTICS

Visualization & Exploration

Numerical Data

COMPARING NUMERICAL DATA ACROSS GROUPS

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

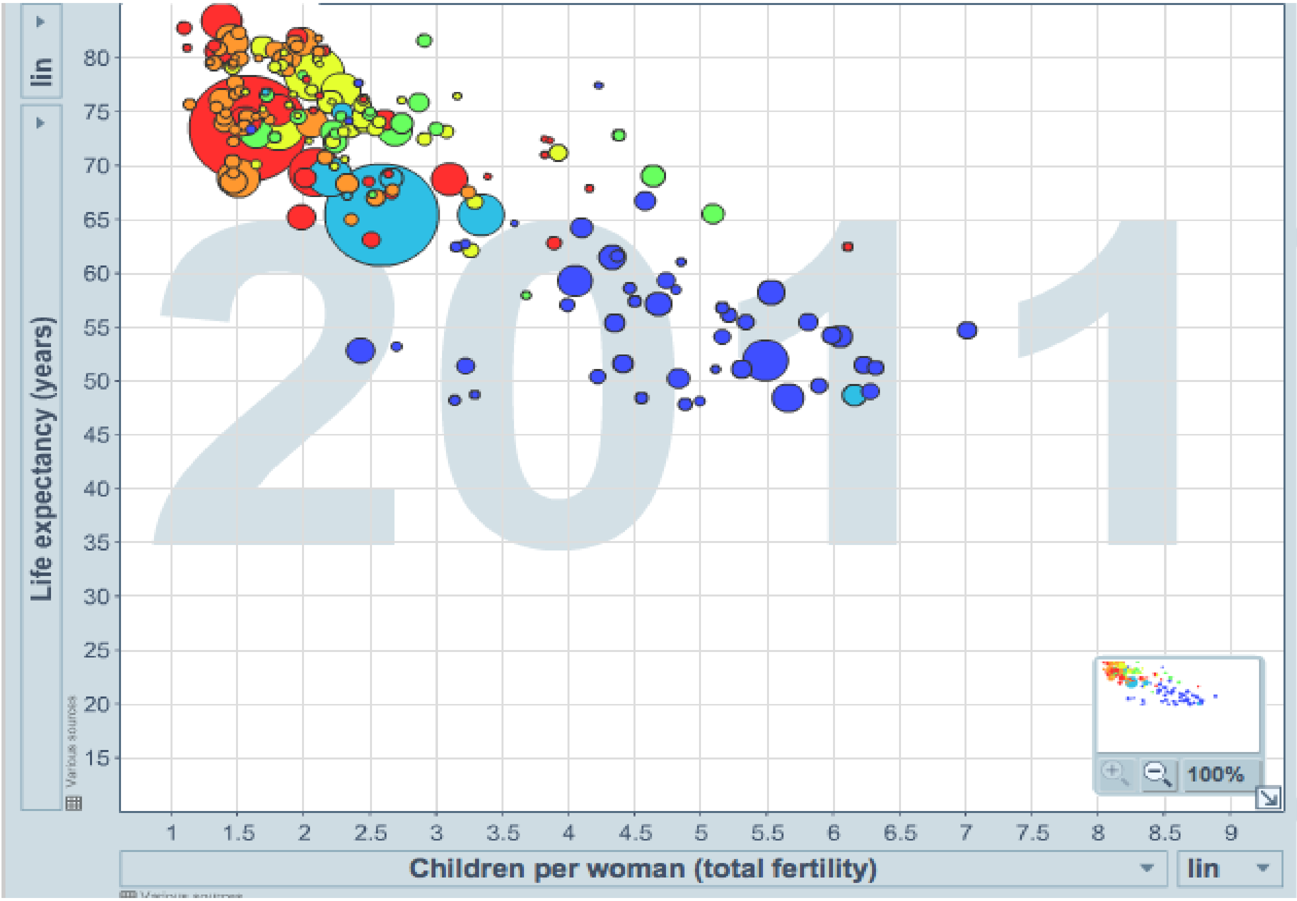


SCATTER PLOT

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

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

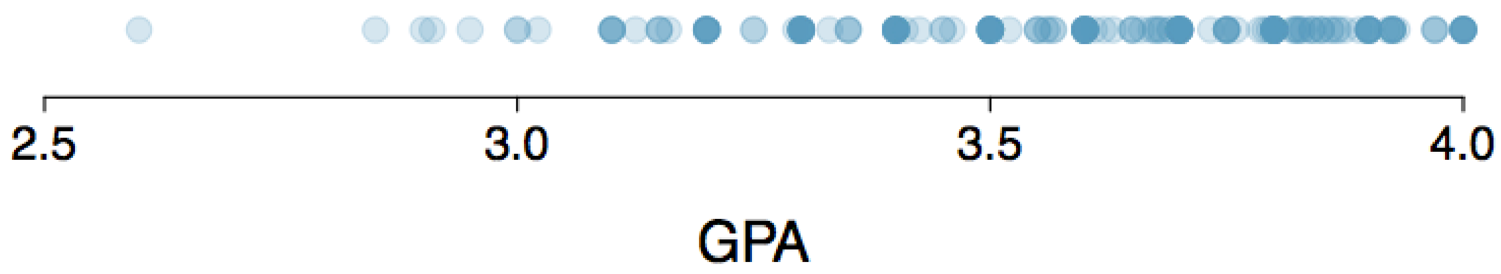
Was the relationship the same throughout the years, or did it change?



DOT PLOTS

Useful for visualizing one numerical variable. Darker colours 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 centre, shape, and spread of the distribution.

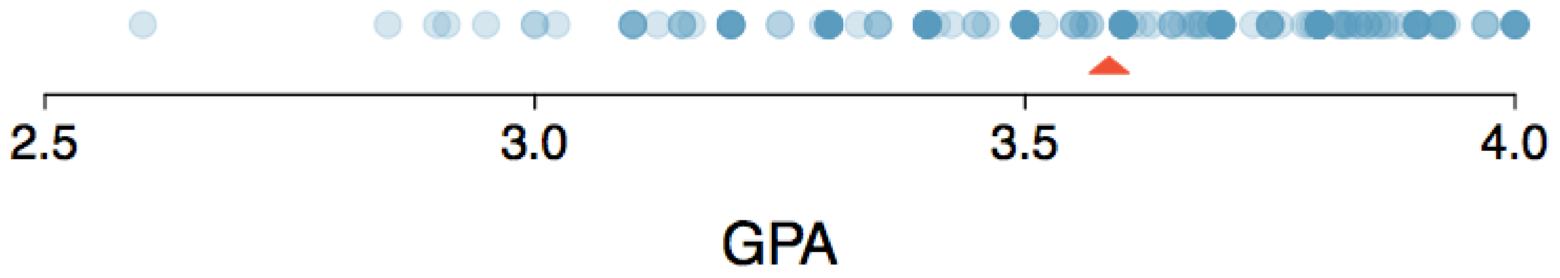


DOT PLOTS AND MEAN

The mean, also called the average (marked with a triangle in the above plot), is one way to measure the centre of a distribution of data.

It is a measure of "central tendency"

The mean GPA is 3.59.



MEAN

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

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

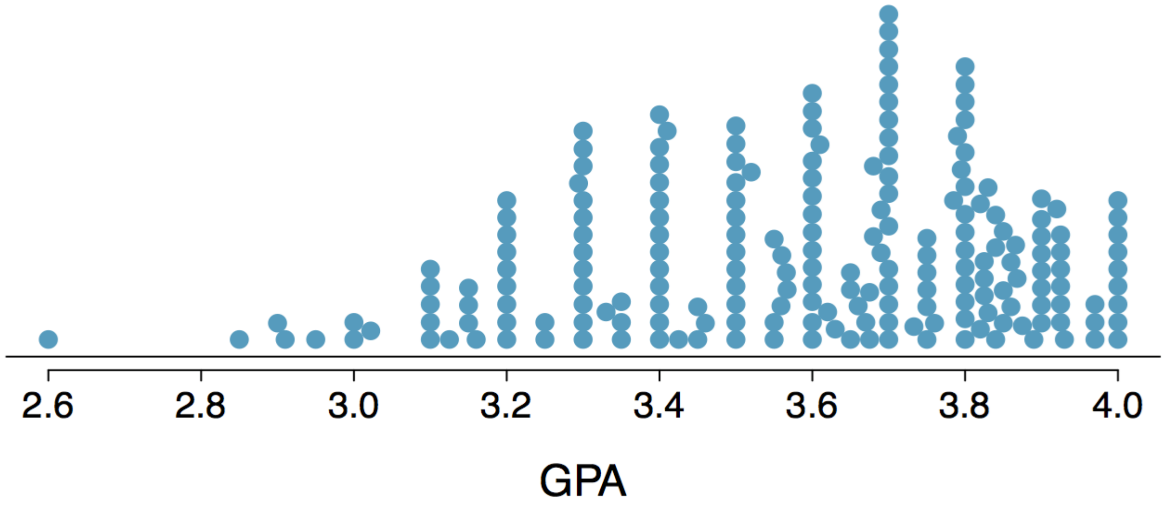
where $x_1, x_2, \ldots, x_n$ represent the n observed values.

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

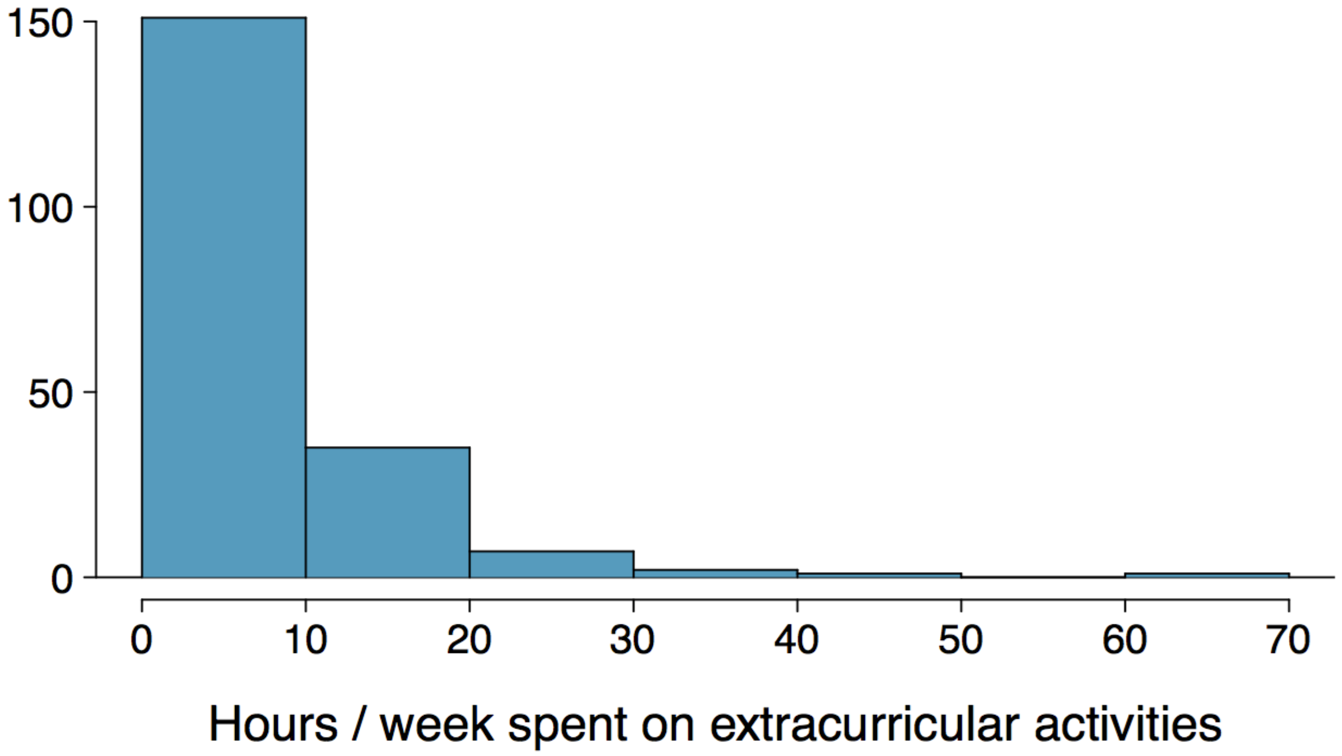


HISTOGRAMS

Histograms provide a view of the data density. Higher bars represent where the data are relatively more common.

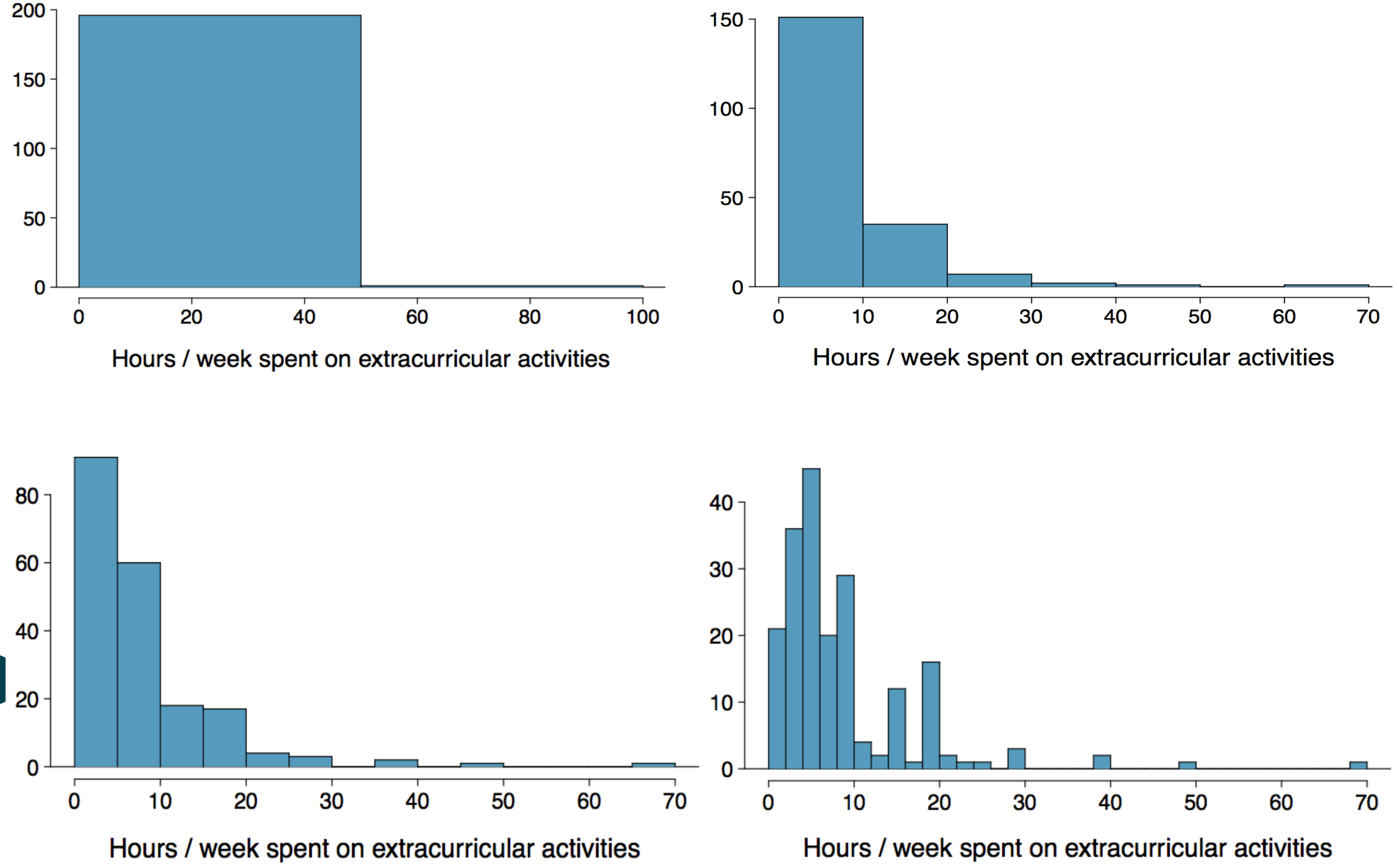
Histograms are especially convenient for describing the shape of the data distribution.

The chosen bin width can alter the story the histogram is telling.



BAR WIDTH (BINS)

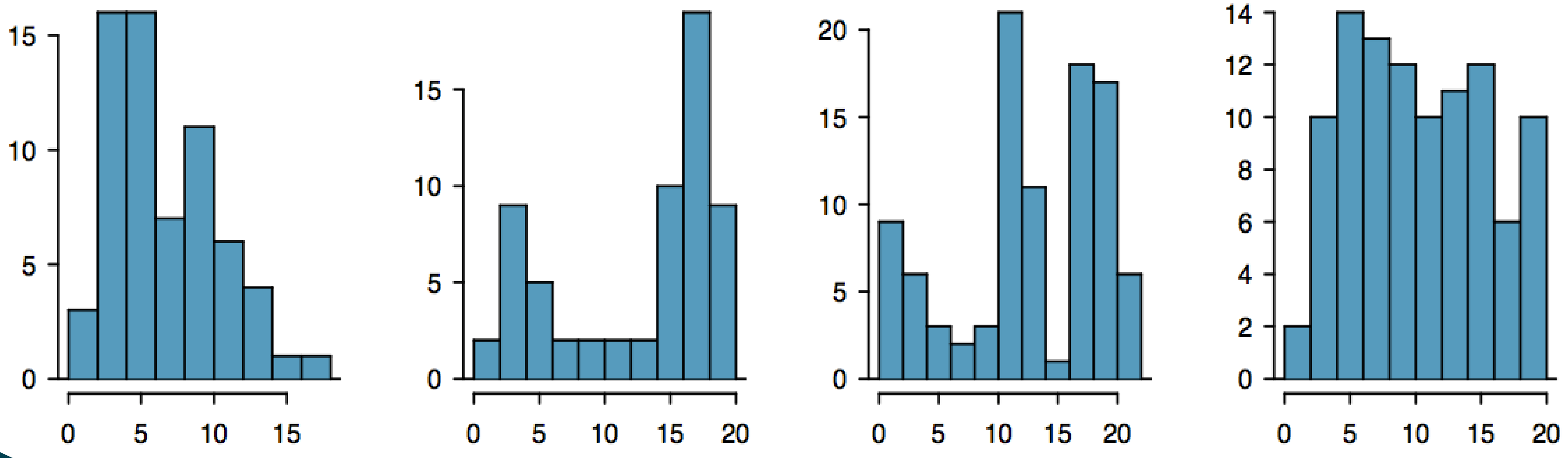
Which one(s) of these histograms are useful? Which reveal too much about the data? Which hide too much?



SHAPE OF A DISTRIBUTION: MODALITY

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

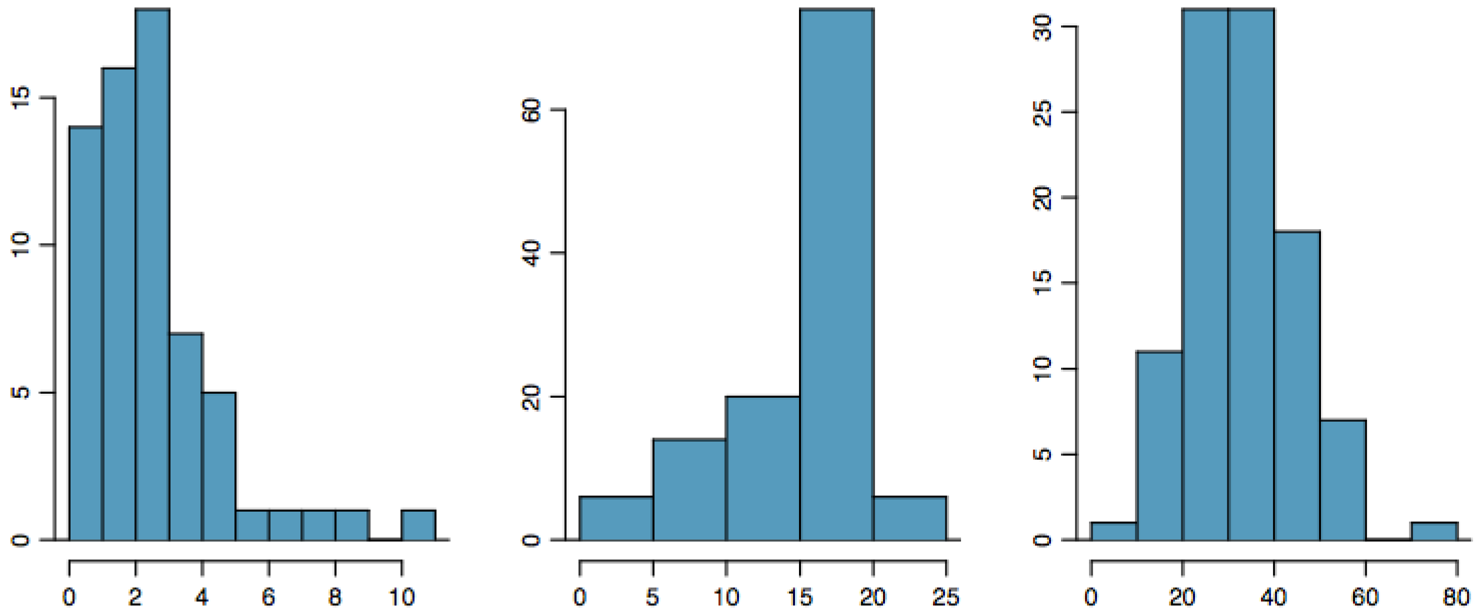
Note: In order to determine modality, step back and imagine a smooth curve over the histogram -- imagine that the bars are wooden blocks and you drop a limp spaghetti over them, the shape the spaghetti would take could be viewed as a smooth curve.



SHAPE OF A DISTRIBUTION: SKEWNESS

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

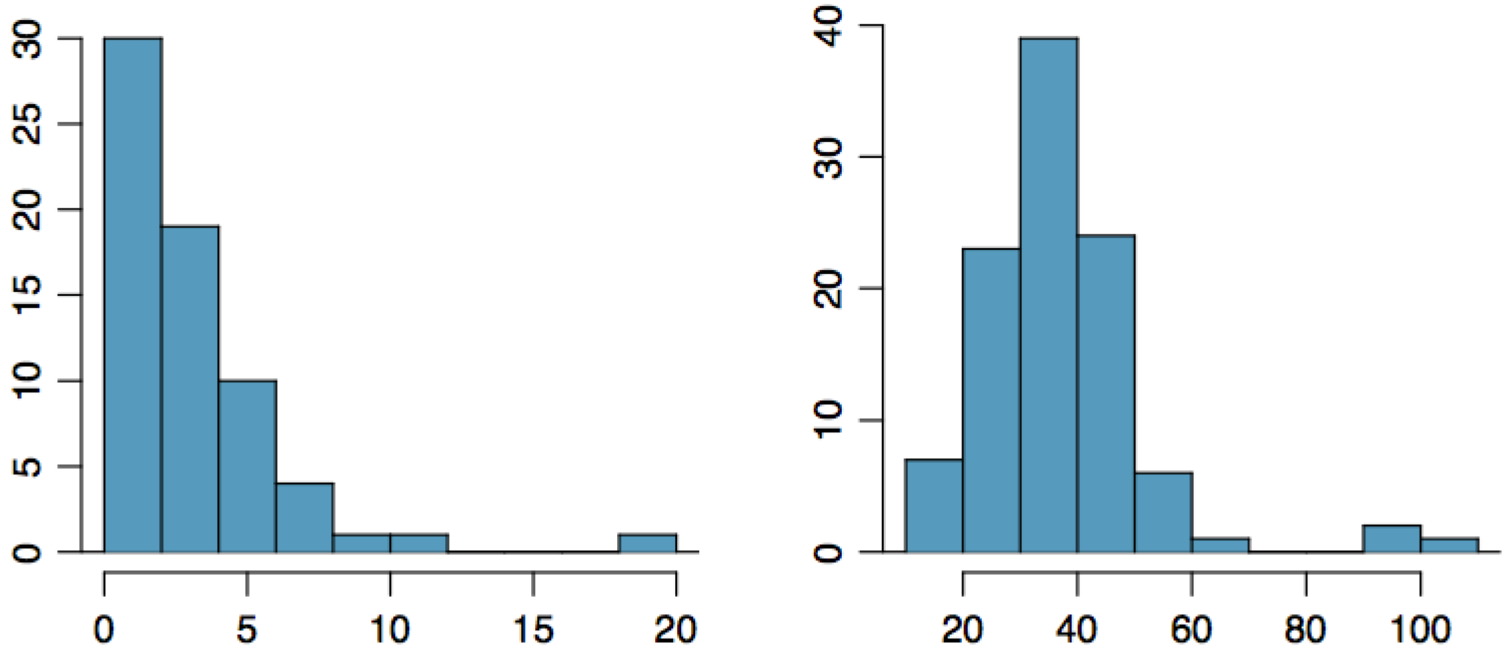
Histograms are said to be skewed to the side of the long tail.



SHAPE OF A DISTRIBUTION:

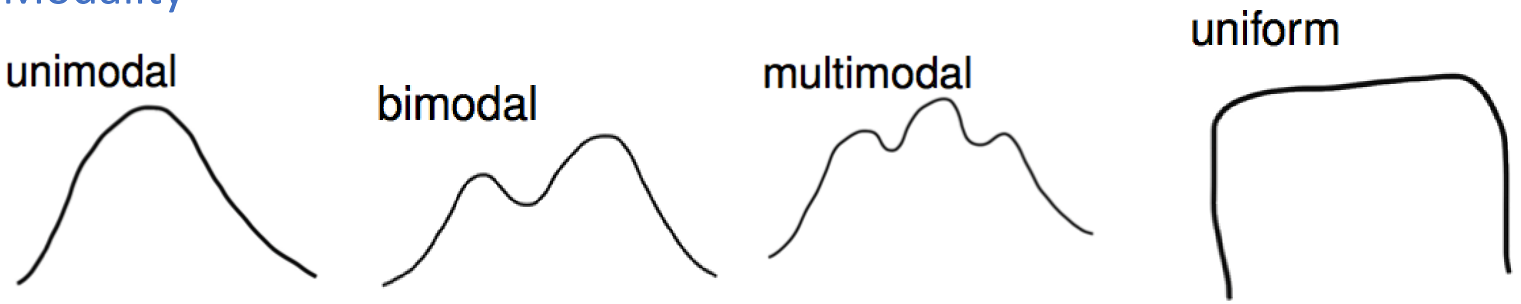
UNUSUAL OBSERVATIONS

Are there any unusual observations or potential outliers?

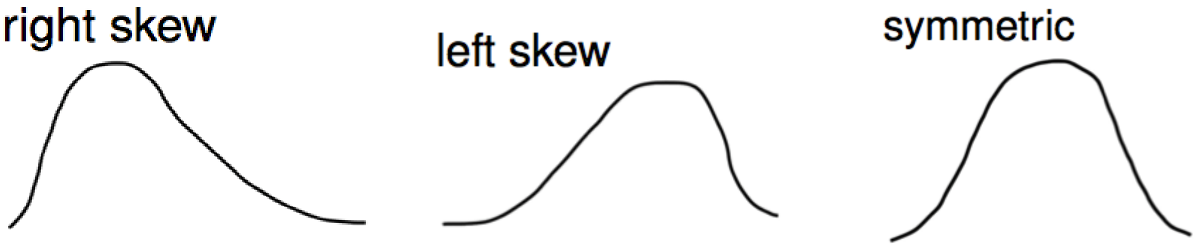


COMMONLY OBSERVED SHAPES OF DISTRIBUTIONS

Modality



Skewness



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)

APPLICATION ACTIVITY:

SHAPES OF DISTRIBUTIONS

Sketch the expected distributions of the following variables:

- People visiting a restaurant per hour
- Distribution of Wealth in the World
- Retirement age
- IQ Score

- Come up with a concise way (1-2 sentences) to teach someone how to determine the expected distribution of any variable.

ARE YOU TYPICAL?

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

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

VARIANCE

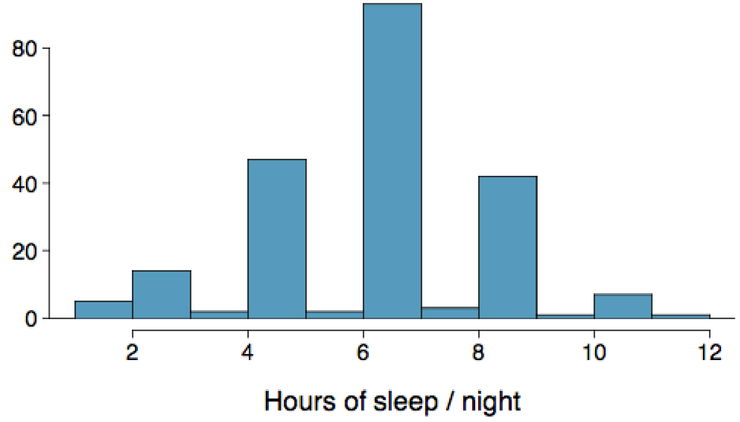
Variance is roughly the average squared deviation from the mean.

The variance of amount of sleep students get per night can be calculated as:

The sample mean is
and the sample size is n = 217.

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

$$\bar{x} = 6.71,$$



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

VARIATION

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

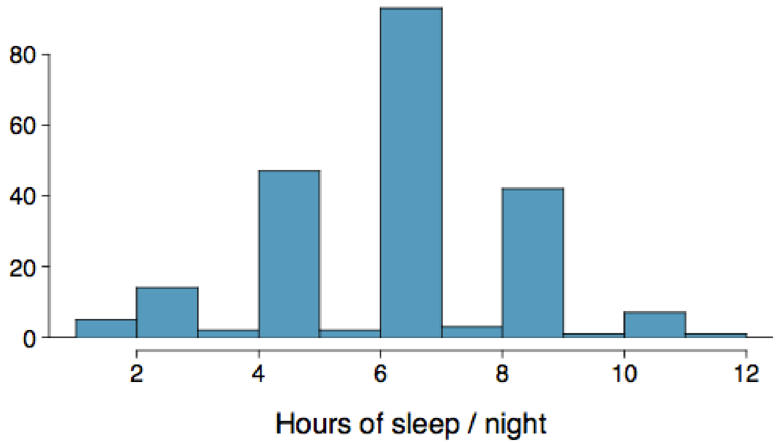
The standard deviation is the square root of the variance, and has the same units as the data.

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

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

We can see that all of the data are within 3 standard deviations of the mean.

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



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.

$$0, 1, \underline{2, 3}, 4, 5 \to \frac{2+3}{2} = 2.5$$

Since the median is the midpoint of the data, 50% of the values are below it. Hence, it is also the 50th percentile.

Q1, Q3, AND IQR

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

The 50th percentile is also called the median.

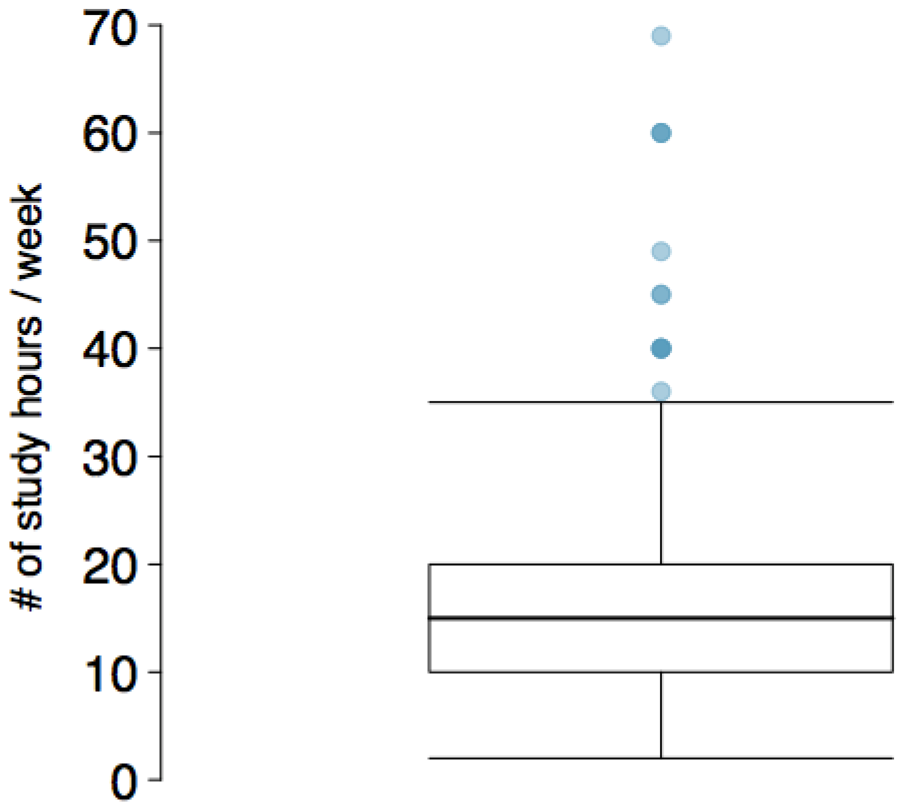
The 75th percentile is also called the third quartile, Q3.

Between Q1 and Q3 is the middle 50% of the data. The range these data span is called the interquartile range, or the IQR.

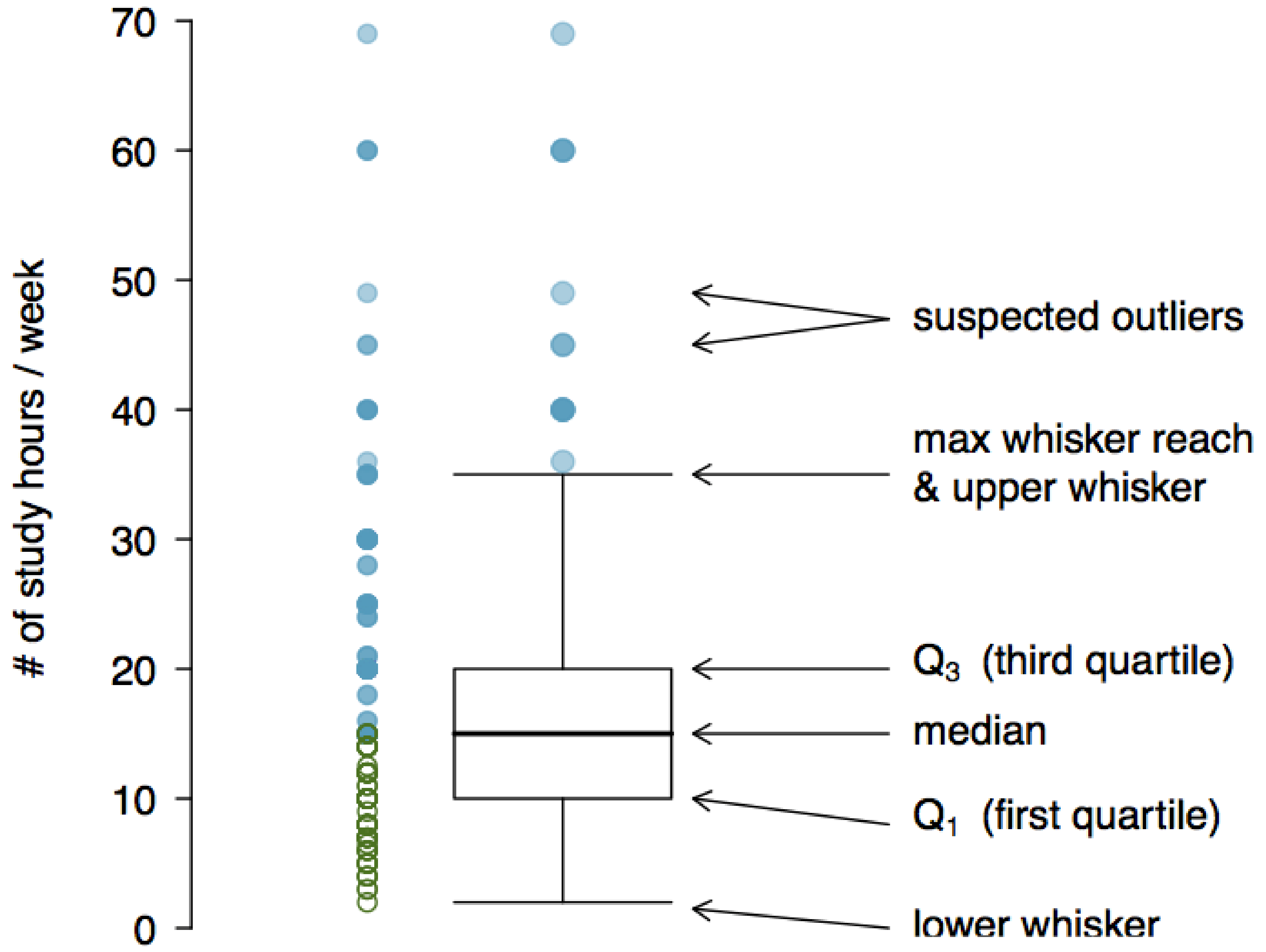
$$IQR = Q3 - Q1$$

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

IQR: 20 - 10 = 10

max upper whisker reach = 20 + 1.5 x 10 = 35

max lower whisker reach = 10 - 1.5 x 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.

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

OUTLIERS

Why is it important to look for outliers?

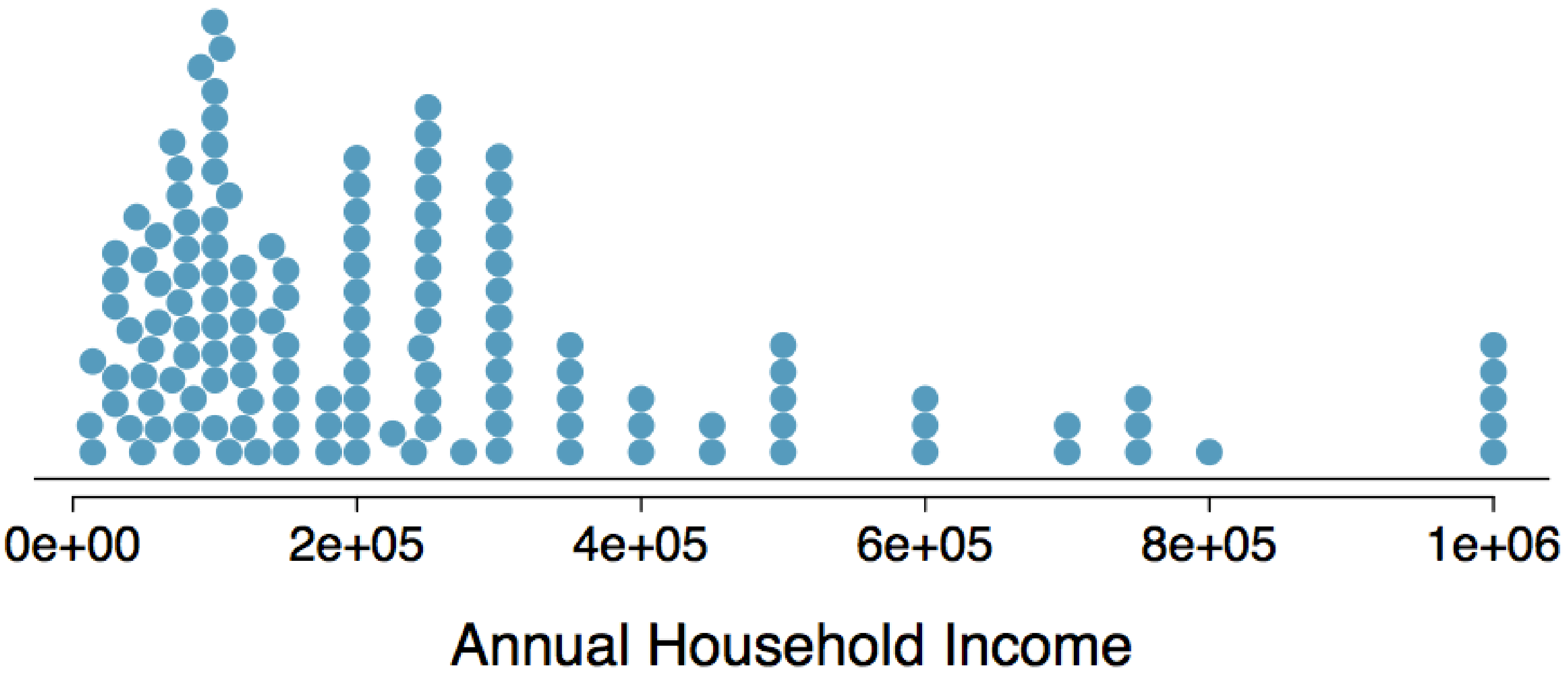
Identify extreme skew in the distribution.

Identify data collection and entry errors.

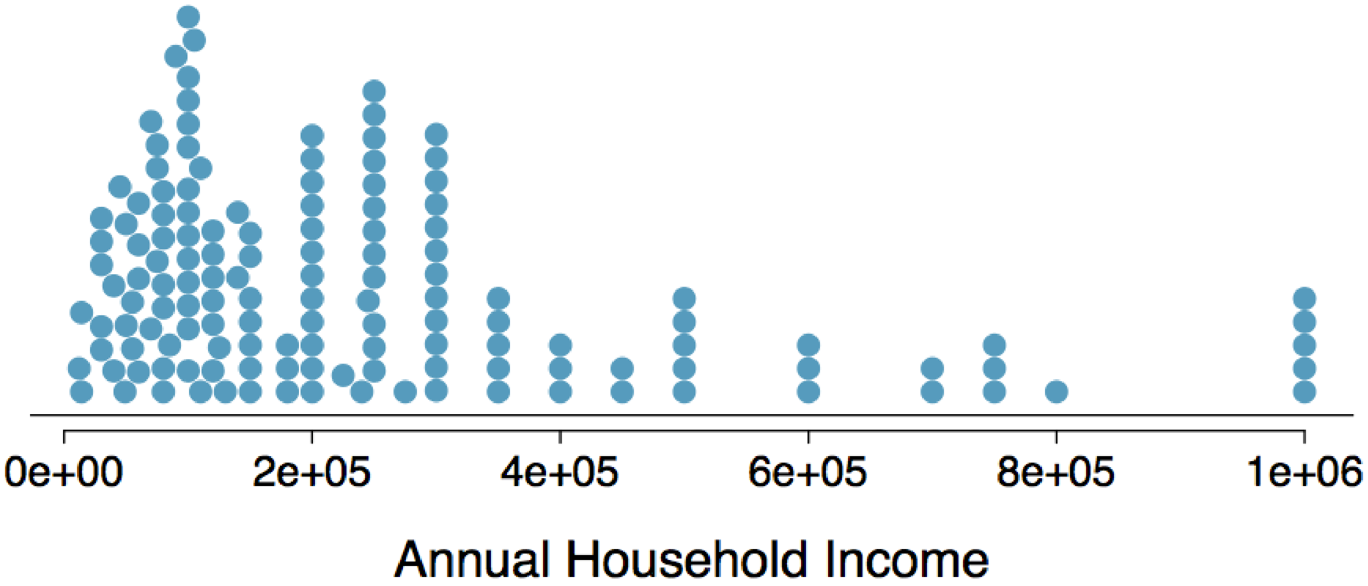
Provide insight into interesting features of the data.

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



	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?

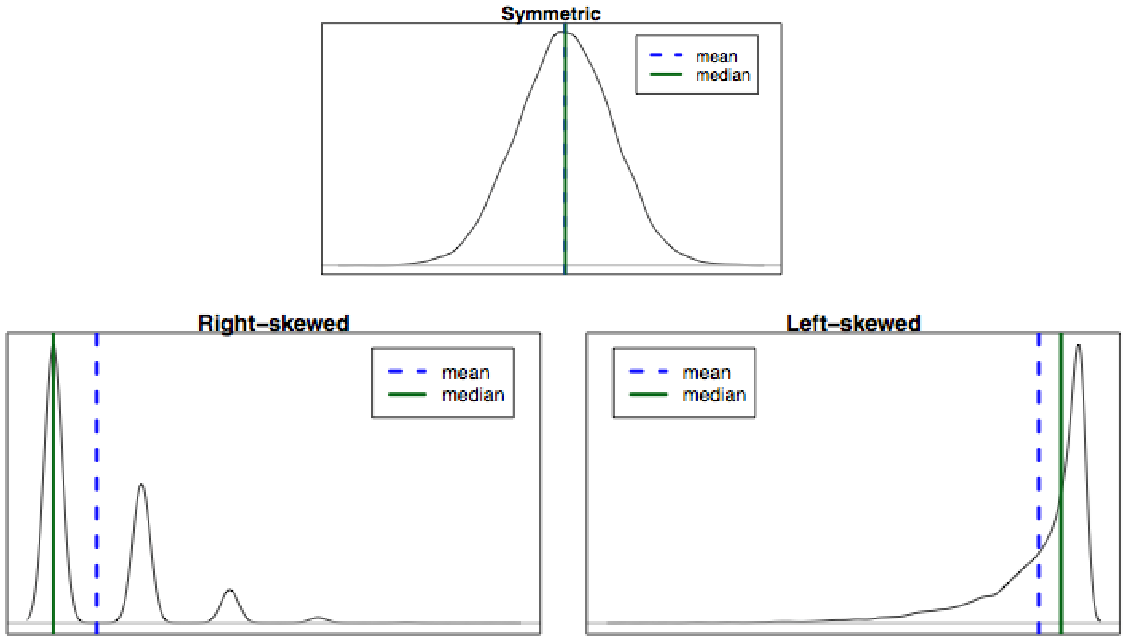
MEAN VS MEDIAN

If the distribution is symmetric, centre is often defined as the mean: mean ~ median

If the distribution is skewed or has extreme outliers, centre is often defined as the median

Right-skewed: mean > median

Left-skewed: mean < median



PRACTICE

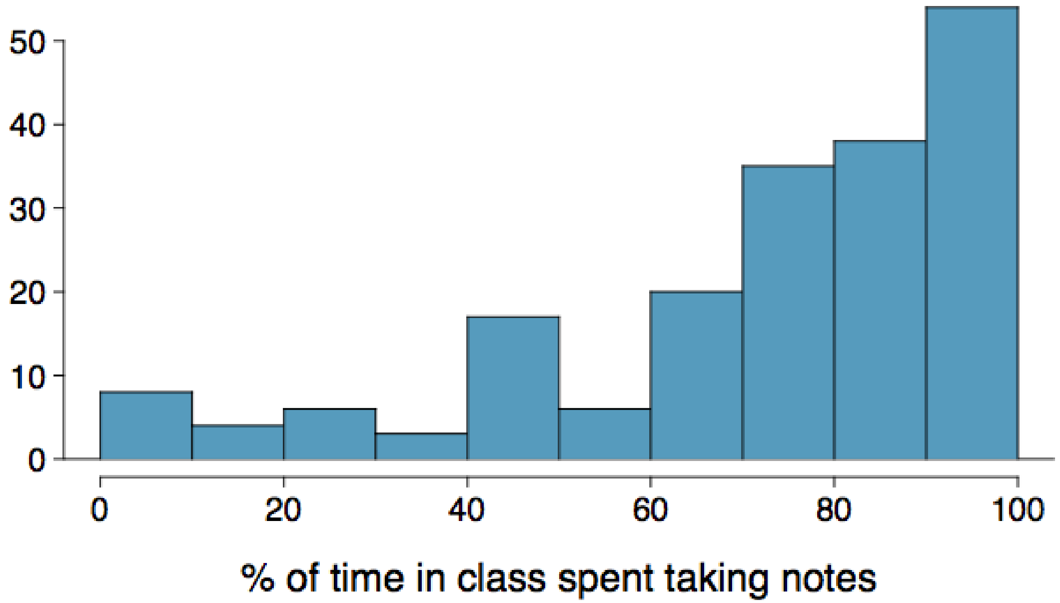
Which is most likely true for the distribution of percentage of time actually spent taking notes in class versus on Facebook, Twitter, etc.?

(a) mean > median

(b) mean ~ median

(c) mean < median

(d) impossible to tell

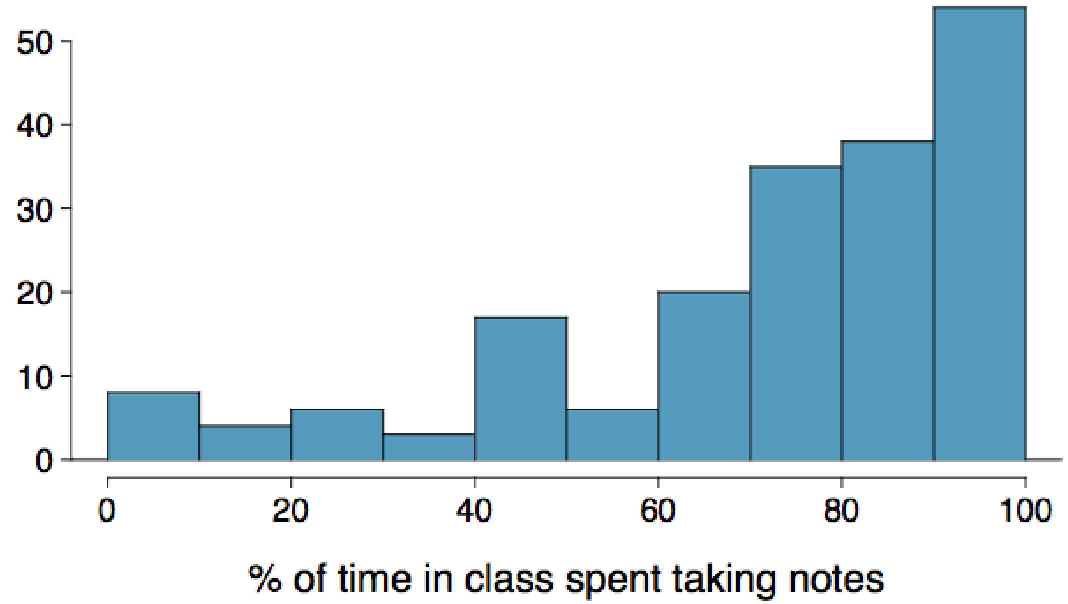


PRACTICE

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

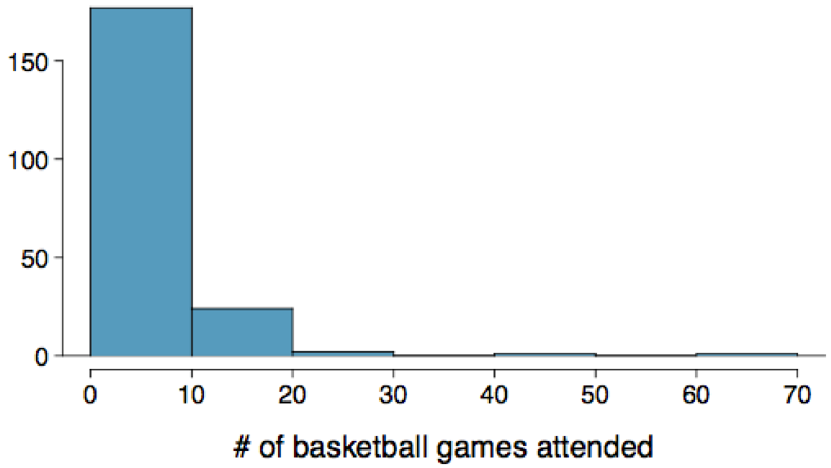
(c) mean < median

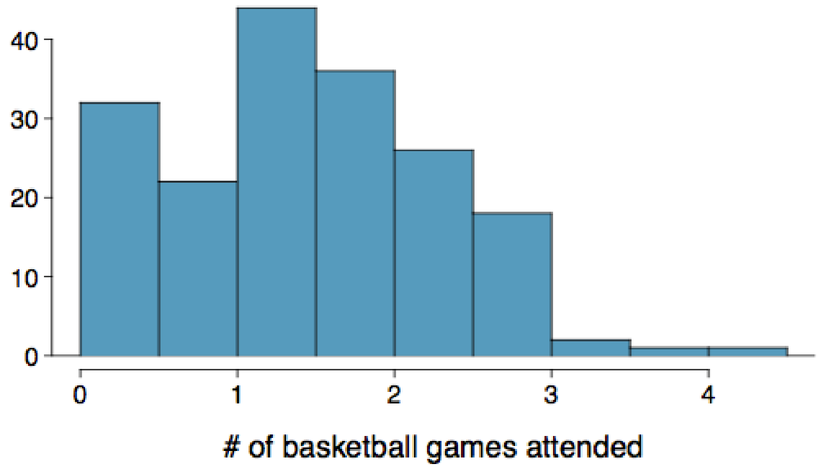
(d) impossible to tell

EXTREMELY SKEWED DATA

When the data is extremely skewed, transforming it 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	...
# of games	4.25	3.91	3.22	...

However, results of an analysis might be difficult to interpret because the log of a measured variable is usually meaningless.

What other variables would you expect to be extremely skewed?

Salary, housing prices, etc.