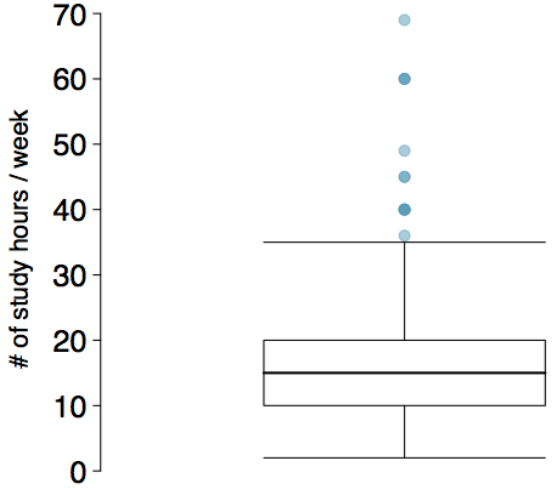
MATH 1051H-A: Lecture #05

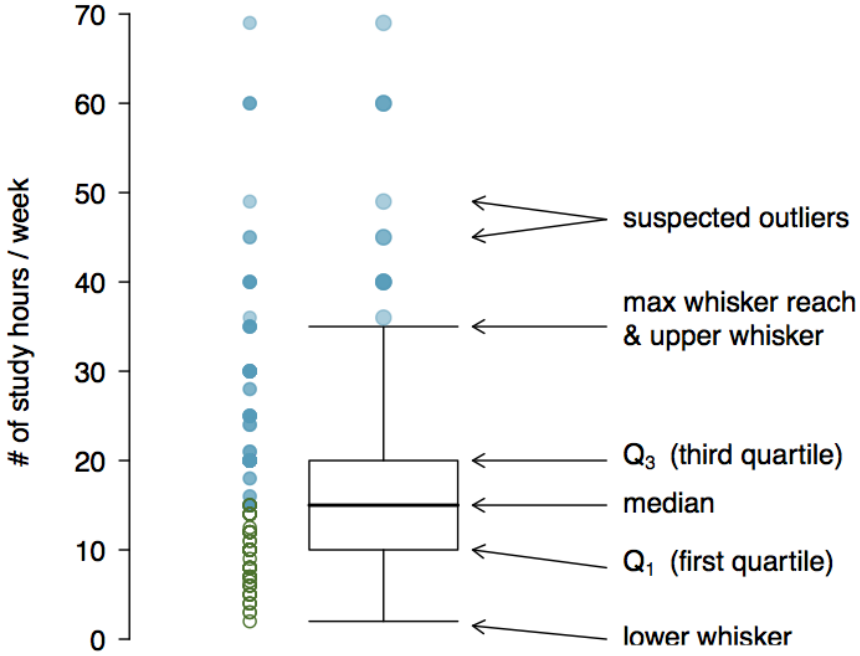
Plotting as Numerical Summary

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

The **whiskers** of a box plot can extend up to $1.5 \times \text{IQR}$ away from the quartiles.

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

Example: IQR: 20 - 10 = 10

- max upper whisker reach = $20 + 1.5 \times 10 = 35$
- 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.

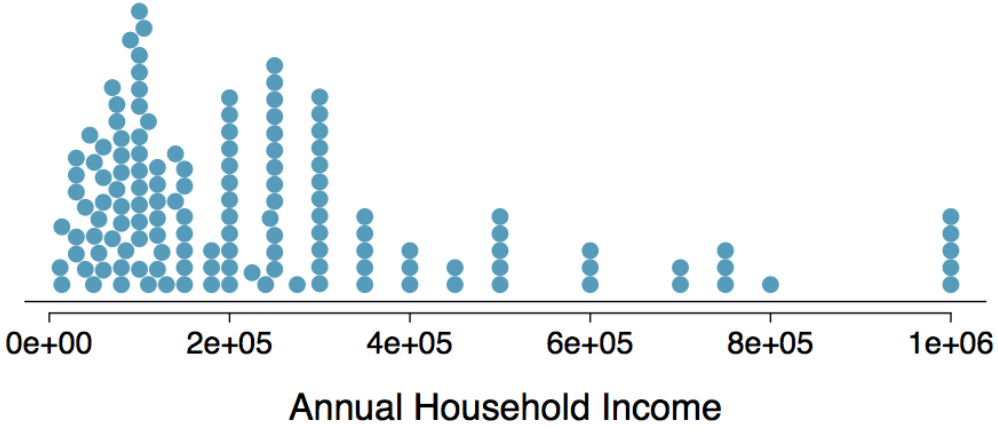
Outliers (continued)

Why is it important to look for outliers?

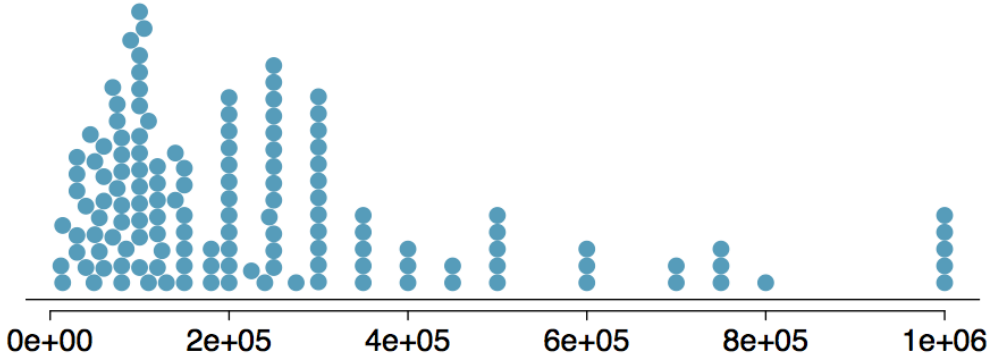
- Identify extreme skew in the distribution.
- Identify data collection and entry errors.
- Provide insight into interesting features of the data.

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?

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

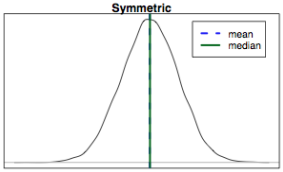
If you would like to estimate the typical household income for a student, would you be more interested in the mean or median income?

Median

Mean versus Median

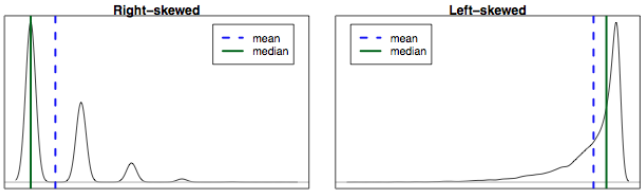
If the distribution is symmetric, center is often defined as the mean:

- mean $\approx$ median



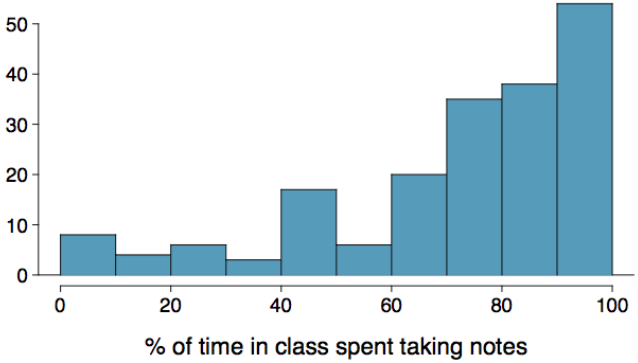
If the distribution is skewed or has extreme outliers, center is often defined as the median

- 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.?



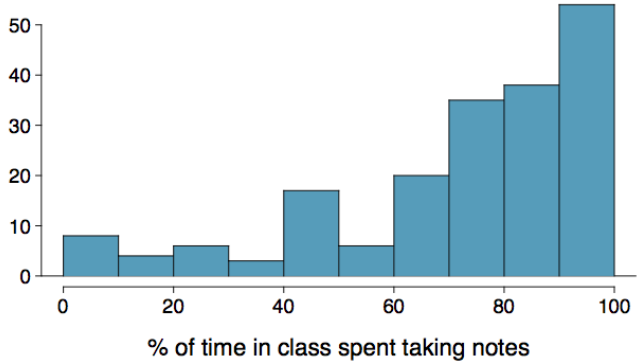
1. mean > median
2. mean < median
3. mean ≈ median
4. impossible to tell

Practice

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

If we compute, the mean = 80% and the median = 76%. So …

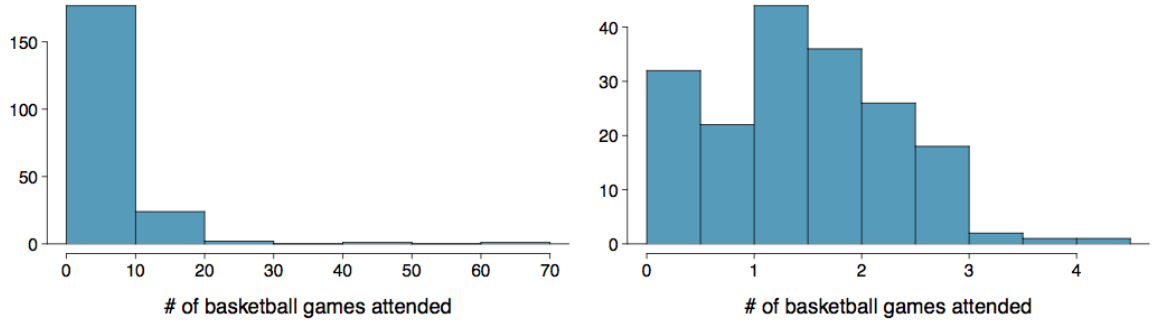
1. mean > median
2. *mean < median*
3. mean ≈ median
4. impossible to tell



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	. . .
# 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?

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, ability to throw a football, …

If Time Allows …

Back to RStudio!