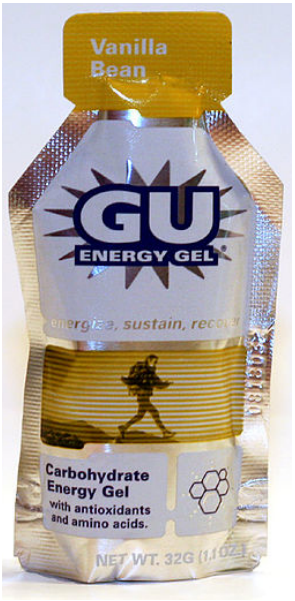
# MATH 1051H-A: Lecture #04

# More on Blocking




We would like to design an experiment to investigate if energy gels makes you run faster:

- **Treatment**: energy gel
- **Control**: no energy gel

It is suspected that energy gels might affect pro and amateur athletes differently, therefore we block for pro status:

- Divide the sample to pro and amateur
- Randomly assign pro athletes to treatment and control groups
- Randomly assign amateur athletes to treatment and control groups
- Pro/amateur status is equally represented in the resulting treatment and control groups

**Why is this important? Can you think of other variables to block for?**

# Practice

A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are equally represented in each group. Which of the below is correct?

1. There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)
2. There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)
3. There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)
4. There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

# Practice

A study is designed to test the effect of light level and noise level on exam performance of students. The researcher also believes that light and noise levels might have different effects on males and females, so wants to make sure both genders are equally represented in each group. Which of the below is correct?}

1. There are 3 explanatory variables (light, noise, gender) and 1 response variable (exam performance)
2. *There are 2 explanatory variables (light and noise), 1 blocking variable (gender), and 1 response variable (exam performance)*
3. There is 1 explanatory variable (gender) and 3 response variables (light, noise, exam performance)
4. There are 2 blocking variables (light and noise), 1 explanatory variable (gender), and 1 response variable (exam performance)

# Difference between Blocking and Explanatory Variables

Factors are conditions we can impose on the experimental units.

Blocking variables are characteristics that the experimental units come with, that we would like to control for.

Blocking is like stratifying, except used in experimental settings when randomly assigning, as opposed to when sampling.

# More Experimental Design Terminology

**Placebo**: fake treatment, often used as the control group for medical studies

**Placebo effect**: experimental units showing improvement simply because they believe they are receiving a special treatment

**Blinding**: when experimental units do not know whether they are in the control or treatment group

**Double-blind**: when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group

# Practice

What is the main difference between observational studies and experiments?

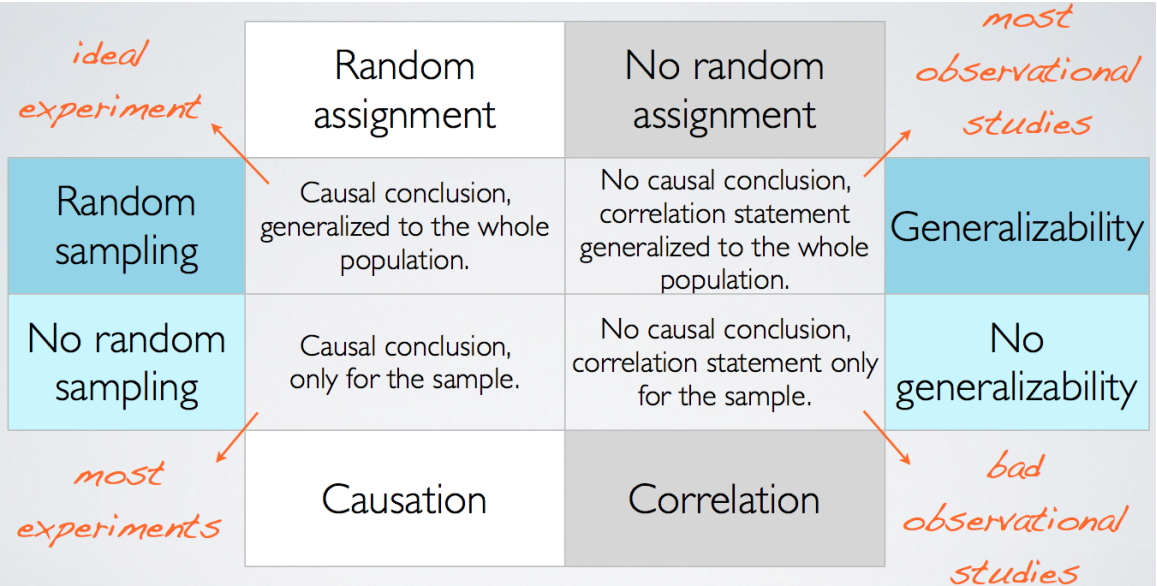
1. Experiments take place in a lab while observational studies do not need to.
2. In an observational study we only look at what happened in the past.
3. Most experiments use random assignment while observational studies do not.
4. Observational studies are completely useless since no causal inference can be made based on their findings.

# Practice

What is the main difference between observational studies and experiments?

1. Experiments take place in a lab while observational studies do not need to.
2. In an observational study we only look at what happened in the past.
3. *Most experiments use random assignment while observational studies do not.*
4. Observational studies are completely useless since no causal inference can be made based on their findings.

# Random Assignment versus Random Sampling

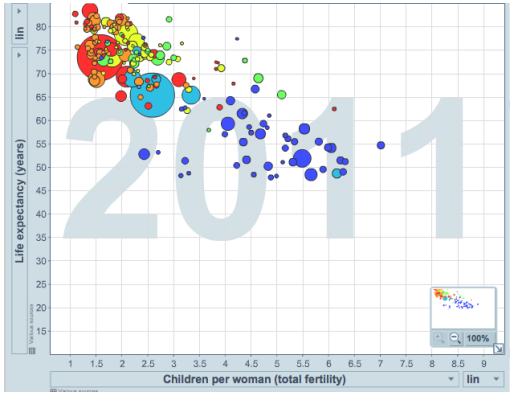


# Examining Numerical Data

# Scatterplots

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

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



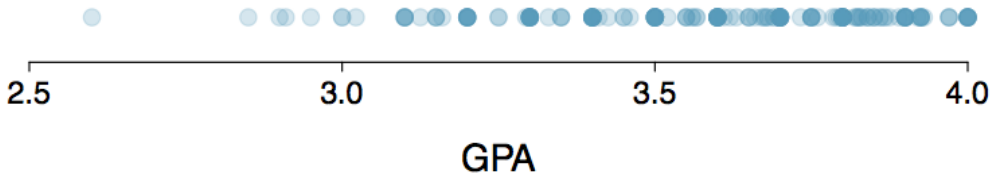

They appear to be linearly and negatively associated: as fertility increases, life expectancy decreases.

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

The relationship changed over the years.

# Dot Plots

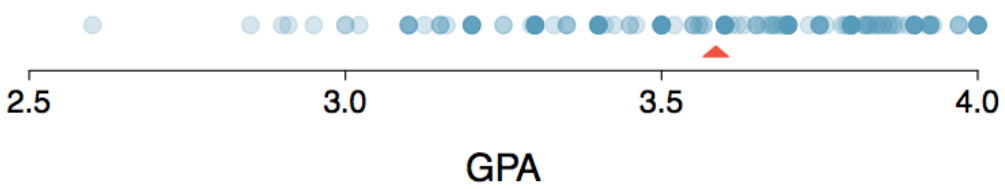
Useful for visualizing one numerical variable. Darker colors represent areas where there are more observations.




**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 and Mean




The **mean**, also called the average (marked with a triangle in the 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

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

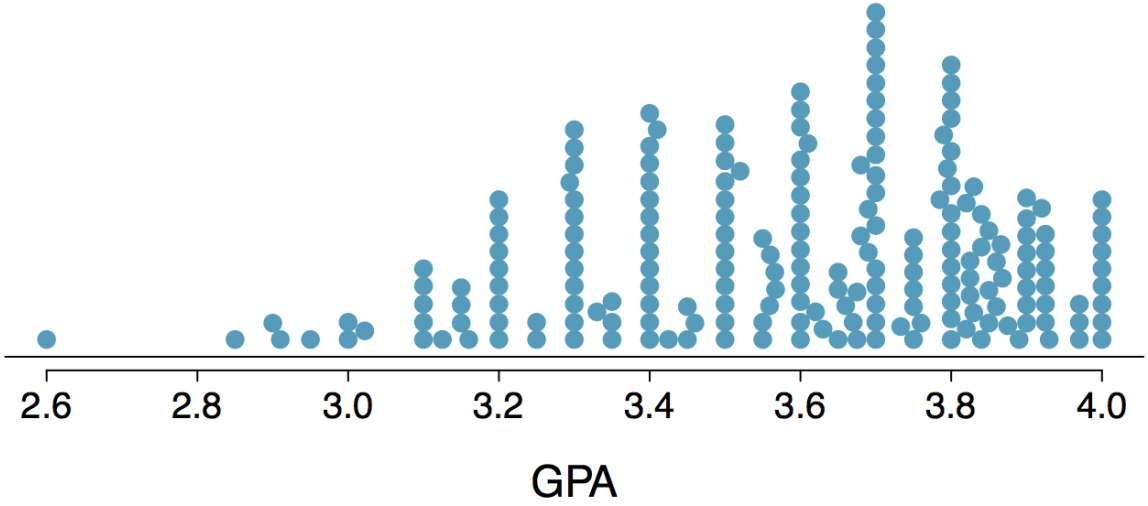
where $x_1, x_2, \cdots, x_n$ represent the $n$ observed values.

The **population mean** is also computed the same way but is denoted as $\mu$. It is often not possible to calculate $\mu$ 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.

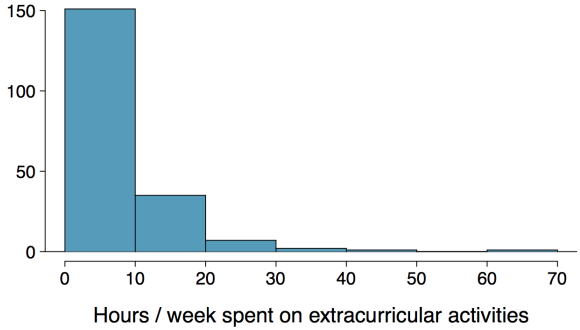
# Stacked Dot Plot

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



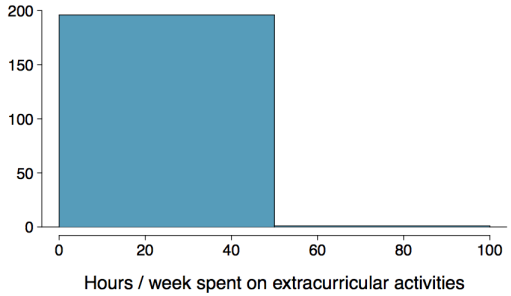
# Histograms — Extracurricular Hours

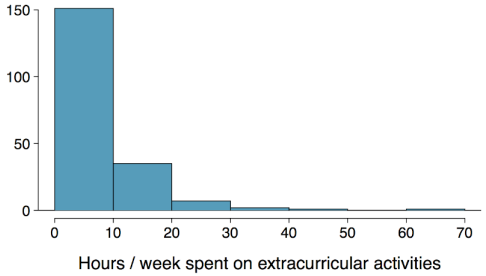
- Histograms provide a view of the **data density**. Higher bars represent where the data are relatively more common.
- Histograms are especially convenient for describing the **shape** of the data distribution.
- The chosen **bin width** can alter the story the histogram is telling.

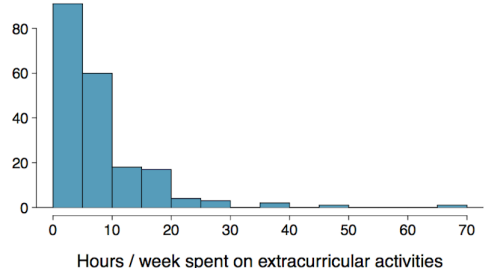


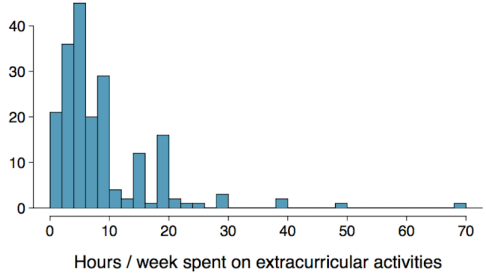
# Bin Width

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



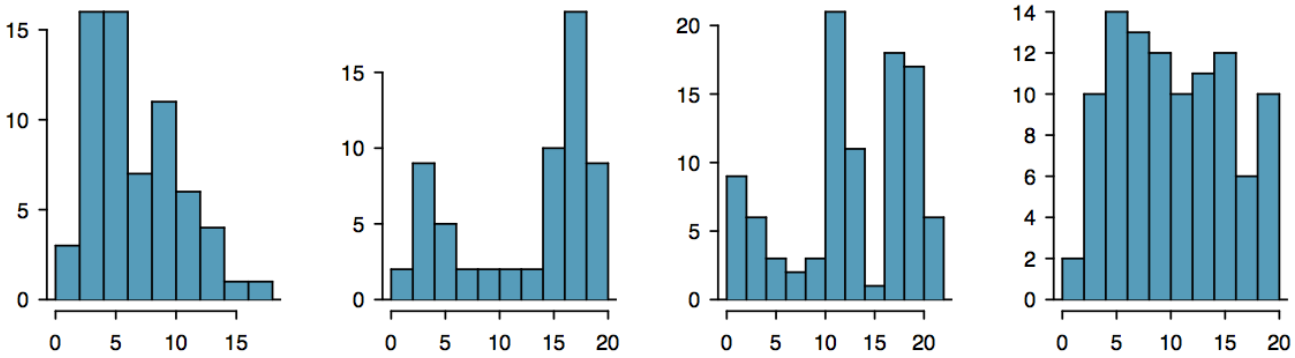



# Shape of a Distribution: Modality

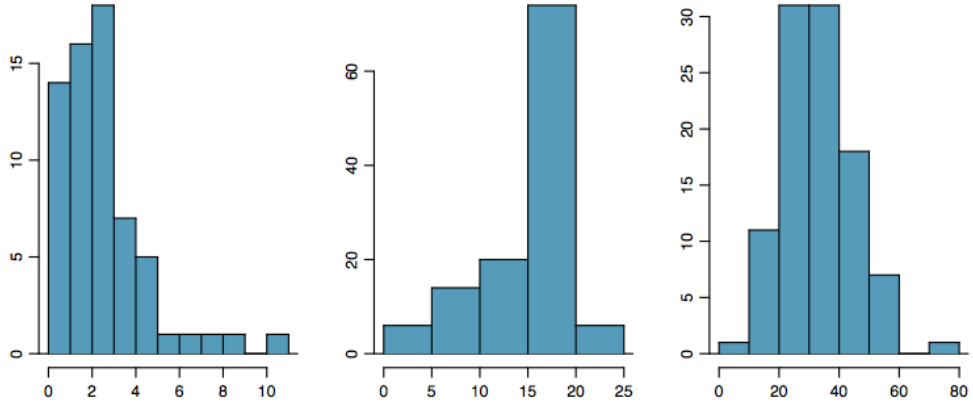
Does the histogram have a single prominent peak (**unimodal**), several prominent peaks (**bimodal/multimodal**), or no apparent peaks (**uniform**)?




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

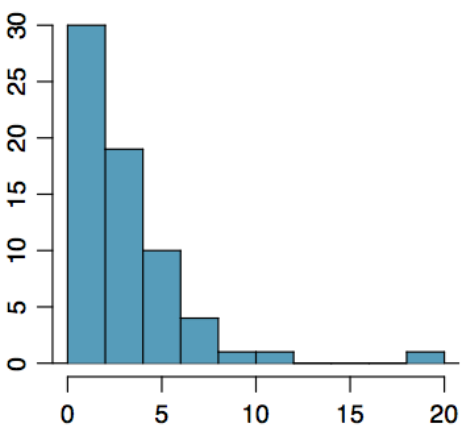
Is the histogram **right skewed**, **left skewed** or **symmetric**?

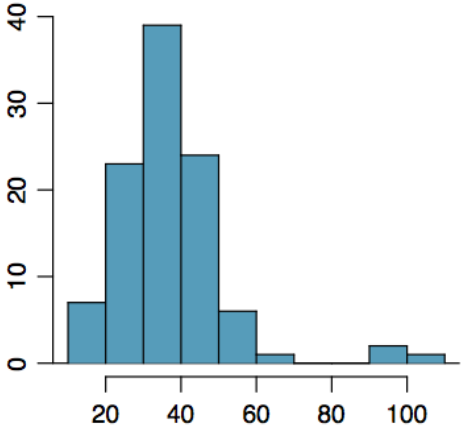



Histograms are said to be skewed to the side of the **long tail**.

# Shape of a Distribution: Unusual Observations

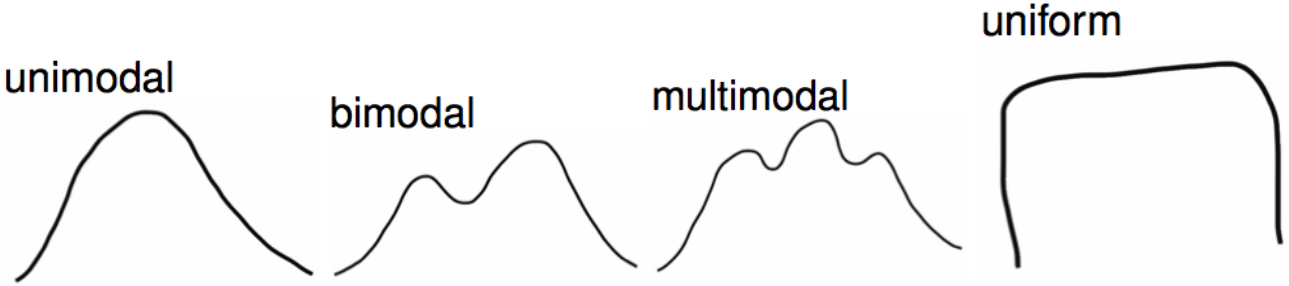
Are there any unusual observations or potential **outliers**?



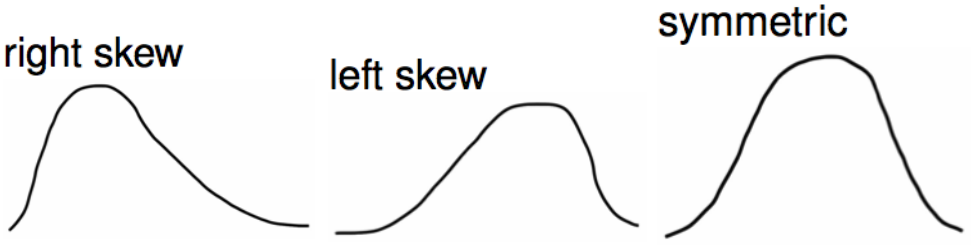



# Commonly observed shapes of distributions

## Modality




## Skewness



# Practice

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

1. weights of adult females
2. salaries of a random sample of people from North Carolina
3. house prices
4. birthdays of classmates (day of the month)

# Practice

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

1. weights of adult females
2. salaries of a random sample of people from North Carolina
3. house prices
4. *birthdays of classmates (day of the month)*

# Application Activity: Shapes of Distributions

Sketch the expected distributions of the following variables:

- number of piercings
- scores on an exam
- IQ scores

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

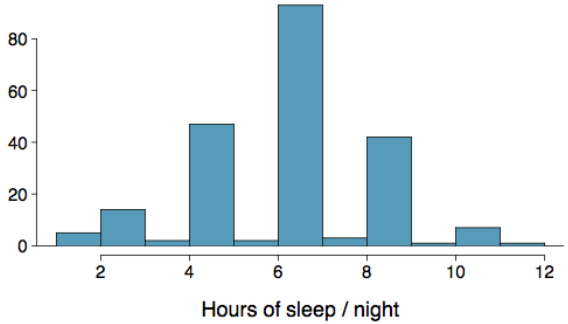
# Are you typical?


are centers
nveying
racteristics
tion?

# Variance

**Variance** is roughly the average squared deviation from the mean.

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




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

$$\begin{aligned} s^2 &= \frac{1}{217-1}\left[(5-6.71)^2 + (9-6.71)^2 + \cdots + (7-6.71)^2\right] \\ &= 4.11 \text{ hours}^2 \end{aligned}$$

# Variance (continued)

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

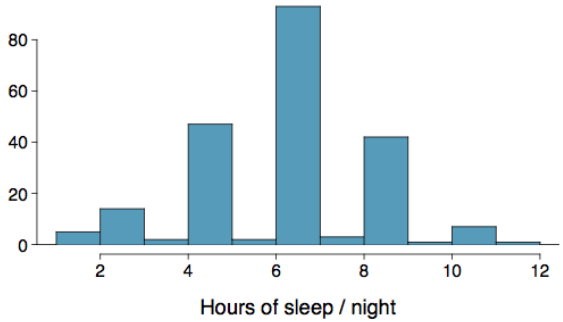
The **standard deviation** is the square root of the variance, and has the same units as the data.

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

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

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

We can see that all of the data are within 3 standard deviations of the mean of $\bar{x} = 6.17$.



# 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 50th **percentile**.

# Percentile

A **percentile** is the the smallest value from an ordered list of numbers which is greater than or equal to that percentage of list elements.

**Example**: The $42^{\text{nd}}$ percentile of the numbers $\{1, 2, 3, \cdots, 99, 100\}$ is 42.

It can become quite complicated when there aren't an even multiple of 100 items!

# 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.

$$\text{IQR} = \text{Q3} - \text{Q1}$$

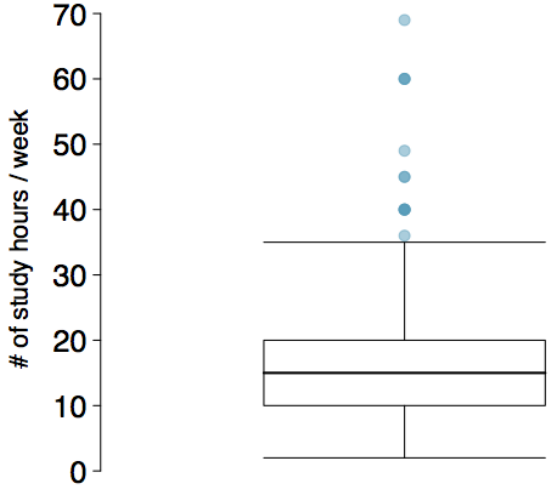
# Example

Now I'd like to switch over to RStudio for a bit, and show you how to **do** some of this!

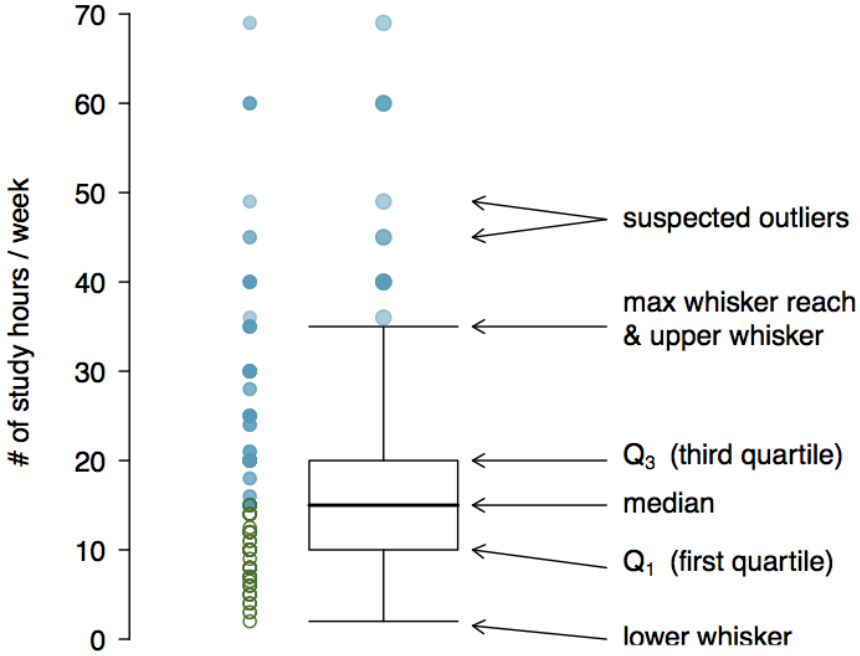
# Plotting as Numerical Summary

# 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.

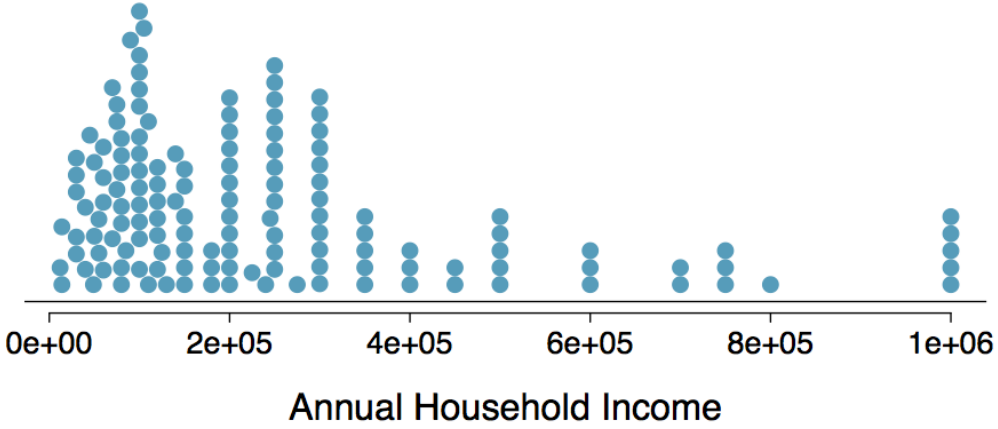
# Outliers (continued)

Why is it important to look for outliers?

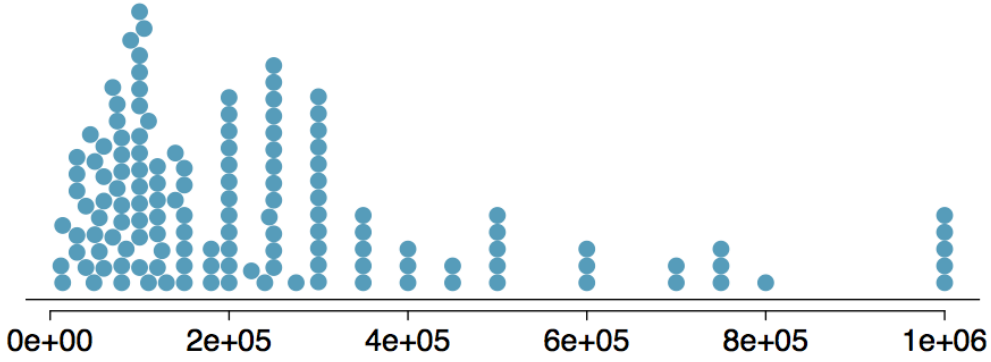
- Identify extreme skew in the distribution.
- Identify data collection and entry errors.
- Provide insight into interesting features of the data.

# 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?

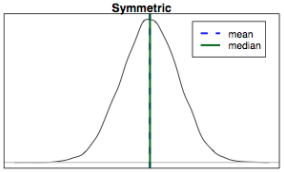
*Median*

# Mean versus Median

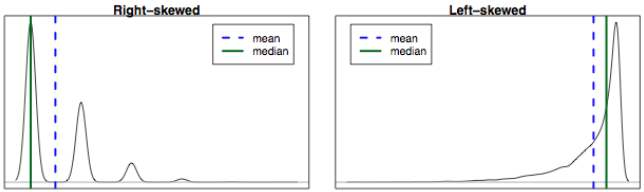
If the distribution is symmetric, center is often defined as the mean:

- mean $\approx$ median

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

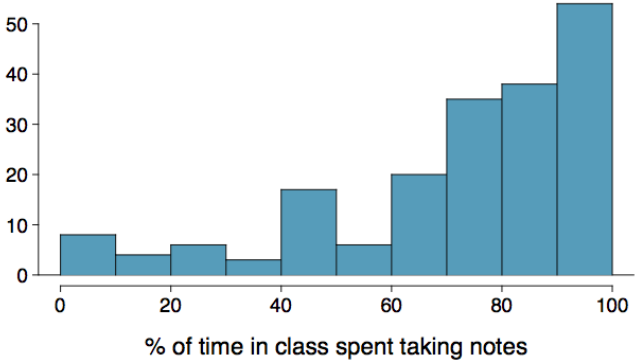



- Right-skewed: mean > median
- Left-skewed: mean < median



# Practice

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



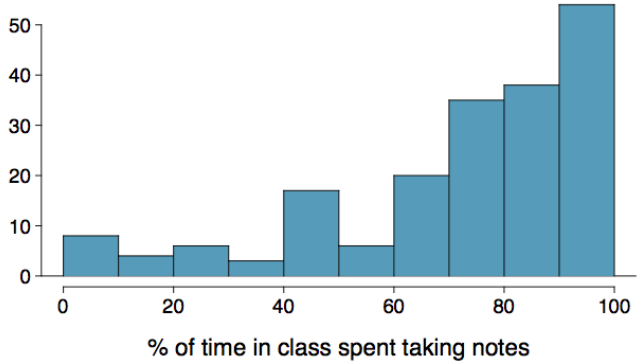

1. mean > median
2. mean < median
3. mean $\approx$ median
4. impossible to tell

# 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 …

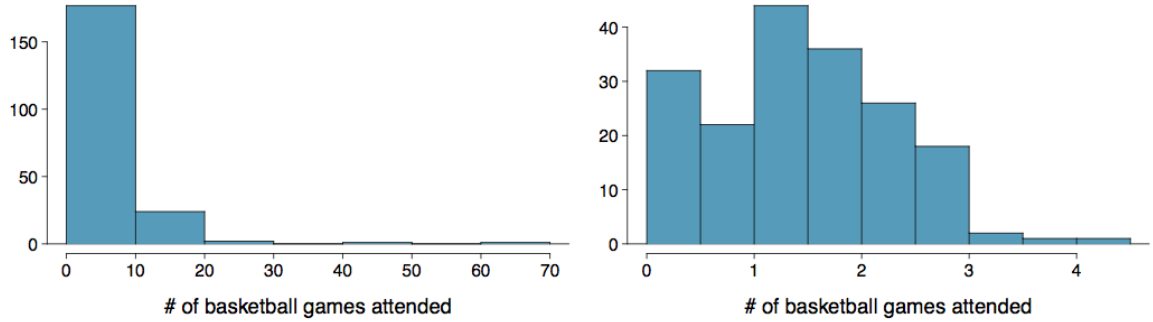
1. mean > median
2. *mean < median*
3. mean ≈ median
4. impossible to tell



# 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, …*