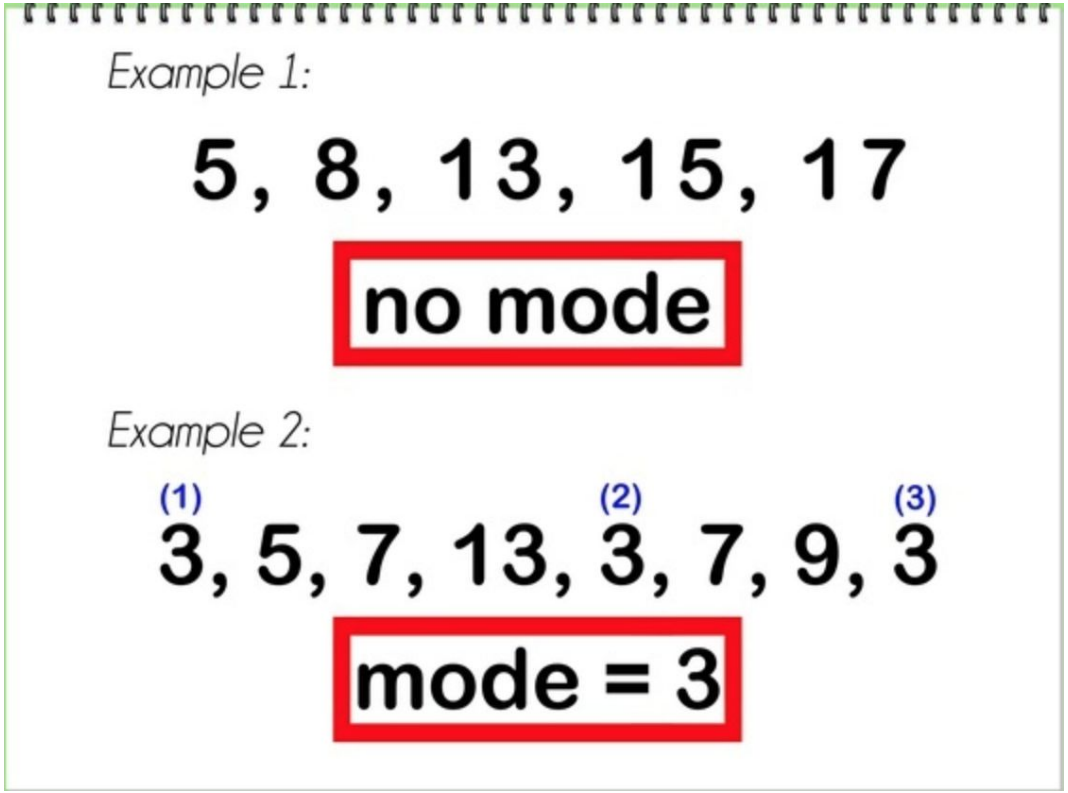
Intro to Statistics

Measures of Central Tendency

Measures of Central Tendency

- The **MODE** of a data set is the most frequently occurring element.
  a. For example, in a list like `[1, 1, 2]`, 1 would be the mode.
- The **MEDIAN** of a data set is the middle element
  a. To find the median, we first sort the data, and select the middle element. In `[1, 2, 3]`, 2 is the median.
  b. For even-length data sets, we have *two* elements in the middle of the list.
     i. We generally return the average of the two elements as the median of such a list.
- Explain that the **MEAN** of a data set is what is commonly called the *average* of a data set.
  a. To calculate the mean, we sum all of the numbers in the data set, and divide by the length of the data set.

Examples - Mode



Examples - Median

1, 3, 3, **6**, 7, 8, 9

Median = 6

1, 2, 3, **4**, **5**, 6, 8, 9

Median = (4 + 5) ÷ 2

= 4.5

Examples - Mean

The marks of seven students in a mathematics test with a maximum possible mark of 20 are given below:

15 13 18 16 14 17 12

Find the mean of this set of data values.

Solution:

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

$$= \frac{15+13+18+16+14+17+12}{7}$$

$$= \frac{105}{7}$$

$$= 15$$

So, the mean mark is 15.

Measures of Central Tendency - The Why?

- The purpose of each of these numbers is to allow us to describe an entire data set with a single number.
- While each of these numbers allows us to describe a data set, **they are not always equally descriptive**….

$$\text{Mean} = \frac{\text{sum of all values}}{\text{total number of values}}$$

$$\text{Median} = \text{middle value (when the data are arranged in order)}$$

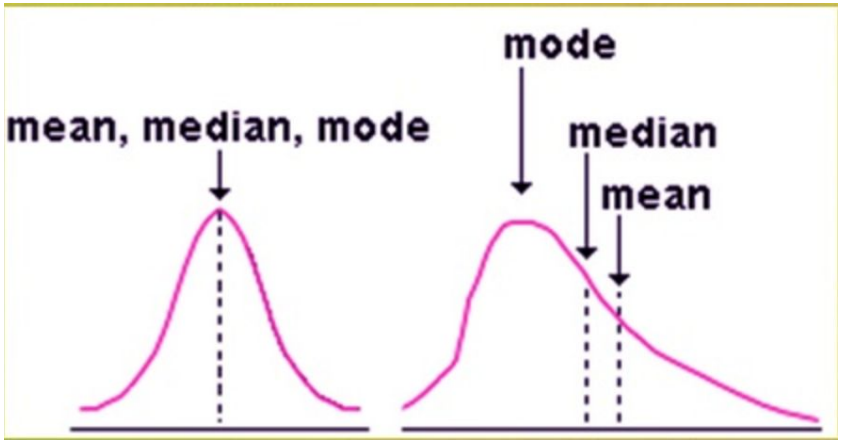
$$\text{Mode} = \text{most common value}$$

A Case for the Median - Skewed Data

- The median is a good choice for describing *just about any data set*.
- **Example**:
  a. These are prices from a local department store: [30, 31, 31, 32, 32, 40, 41, 41, 1000].
- The **median**, 40, is a reasonable description of the "average" price in the data set.
- The mean, which works out to 154.75, does not describe the data very well — in fact, the mean is a different order of magnitude than *any* of the prices in the data set!
- This illustrates that the median tends to give relatively faithful descriptions even in the face of outliers, such as 1000.

A Case for the Median - Skewed Data

- Advantages:
  a. In general, the median is more "resistant" to extreme fluctuations in data than the mean. (outliers)
  b. This property often makes the median a better choice than the mean, in spite of the fact that the mean is more common. It's almost always safe to use the median to describe data.

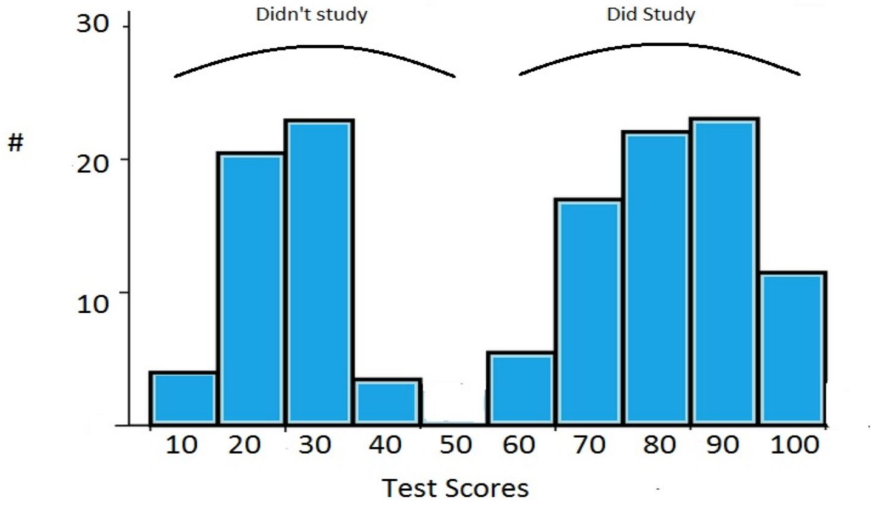


A Case for the Mode - Clusters

- Sometimes, we might have data where values cluster in *several* places.
- Consider this longer list of prices: [30, 31, 31, 32, 32, 40, 41, 41, 1000, 1210, 1210, 1567]
  a. In this case, the mean, 438.75, still differs from the typical price in either the low or high clusters by an order of magnitude.
  b. The median, 40.5, also doesn't adequately describe the data.
  c. *Most* prices are close to 40.5, but a large number of them — 30% — are much higher.
- Data like this, which has two or more *clusters* of data that are spread apart from one another, is often best described by the **mode**.
- The modes are 31, 32, 41, and 1210, numbers which *do* represent the spread of the data quite nicely.
- If this data set had included only one instance of 1210, this "trick" would not work.
  a. If this is not the case, we can describe the data using the modes alongside the median.

A Case for the Mode - Clusters

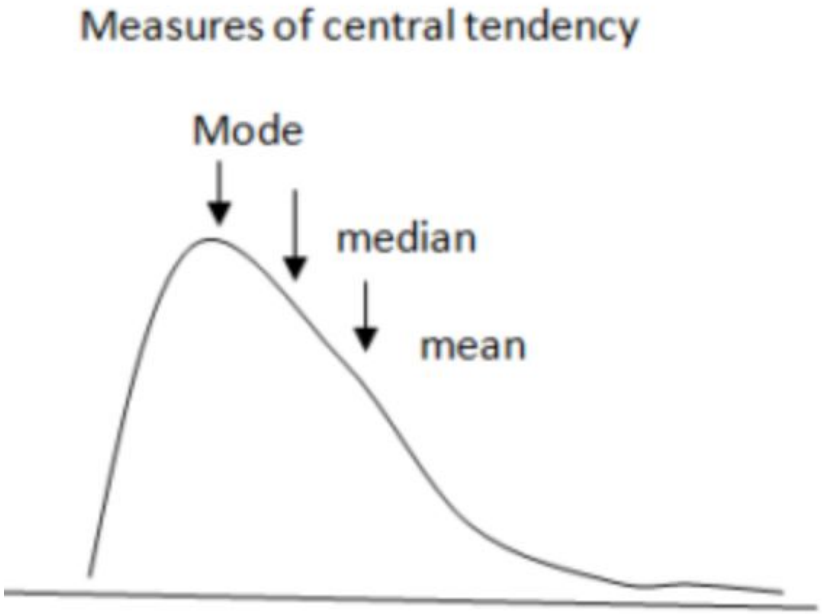
- Data like this, which has two or more *clusters* of data that are spread apart from one another, is often best described by the **mode**.
- This is because a data set can only have one median or mean, but multiple modes; and that the list of modes is likely to contain numbers from each cluster.



A Case for the Mean

- The **mean** is most useful for describing data that are *close together.*
- **Example**:
  a. Consider the list of only low prices: [30, 31, 31, 32, 32, 40, 41, 41].
- In this case, the **mean**, 34.75, *is* an accurate summary of the price data.
- Real data sets are often more "spread out" than this, and that it is difficult to guarantee that a data set does not contain extreme values that skew the mean.
- The median remains a good default statistic, even in these cases, as it will be fairly close to the mean — for this data set, the median is 32, which is just as "accurate" a summary as 34.75.
- Advantage:
  a. That one important potential advantage of the mean over the median is that it factors in *every value of the data set*, which the median and mode do *not*.

Central Tendency Example



Variance, Standard Deviation and Z-score

How about the Spread of the Data?

- Consider these two lists: [3, 4, 5, 6, 7] and [-1525, -200, 5, 745, 1000].
  a. For both of these data sets, the mean and median are 5 — but they are *obviously* different!
- In addition to knowing the the mean, median, and mode of a dataset, we clearly need to know something more about the data, namely, we need to know something about the *spread of the data.*
- Important statistics that measure spread are:
  a. Variance
  b. Standard deviation
  c. Z-score

Variance

- The **variance** of a data set is a single number that describes how "far apart" its values are.
- Measured in units *squared*

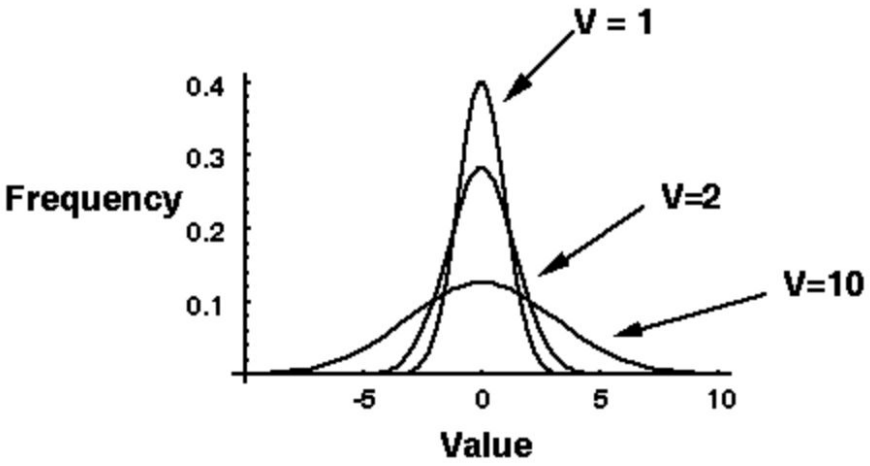
The variance is computed as the average squared deviation of each number from its mean. For example, for the numbers 1, 2, and 3, the mean is 2 and the variance is:

$$\sigma^2 = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = 0.667 \cdot$$

http://davidmlane.com/hyperstat/A16252.html

Variance

- The more "spread out" a data set is, the higher its variance.
  a. **Example**:
     i. for the data set [3, 4, 5, 6, 7], the variance is 2.
     ii. for the data set [-1525, -200, 5, 745, 1000], the variance is 784,110.



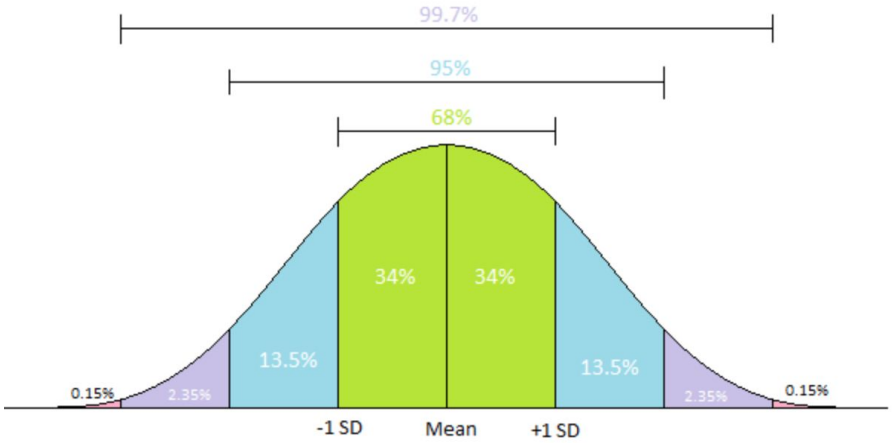
Standard Deviation

- The **standard deviation** is simply the square root of the variance.
  a. For example, the variance of a data set containing the heights of average Australians would be in units of inches (or centimeters) *squared*.
- It is not natural to talk about how much a data set measured in inches spreads in terms of *square* inches — it makes more sense to talk about how much data measured in inches spreads in terms of *inches*.
- One of the motivators for using the standard deviation instead of the variance: It is easier to interpret.
  a. The standard deviation for the data set [3, 4, 5, 6, 7] is approximately 1.414.
  b. The standard deviation for the data set [-1525, -200, 5, 745, 1000] is approximately 885.5.

Standard Deviation

- The way we often use the standard deviation as a unit to describe how far individual numbers in a data set are away from the mean.
  a. Intuitively, the standard deviation tells us how far away from average any number in the data set is. For example, consider the price data we had before: [30, 31, 31, 32, 32, 40, 41, 41, 1000].
  b. If you run the numbers, 1000 would 2.83 standard deviations away from avera
  far above the mean

For comparison, 41 is only -0.3 standard deviations away from average, meaning it is *just a little below* the mean.



Z-Score

- Variance and standard deviation are *single* numbers that describe the *whole* data set
- The **z-score** is a statistic that describes how far away from the mean any *single* number in the data set is.
- The z-score for a number in a data set tells us how many standard deviations away from the mean that number is.
- Recall, the z-score for 1000 in our price data set is about 2.83, whereas for 41, it is -0.3.
- No need to memorize! The libraries we will work with — namely SciPy — have them built-in.

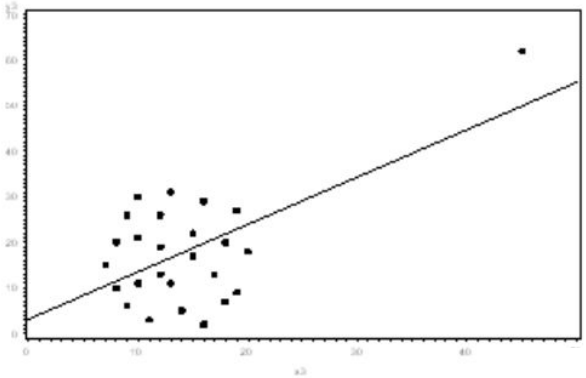
$$z\ score = \frac{(x - \mu)}{\sigma}$$

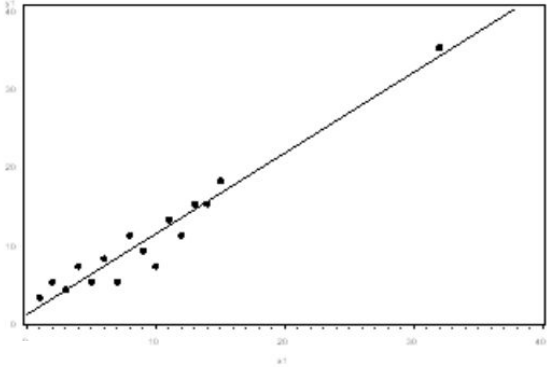
Outliers

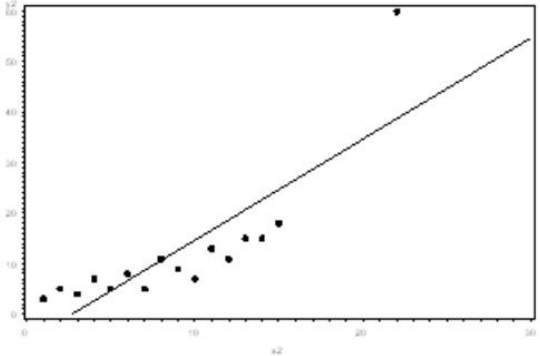
Outliers

- Extreme values often do not describe the data…
- It is okay to remove outliers if any of the following are true:
  - The data is due to bad measurements. (If your data includes reaction time measurements, and one of the values is 1ms, you should probably remove it — human reaction times are *at least* 100ms, so 1ms is clearly bad data.)
  - If the outliers *create* trends that wouldn't exist without them, you *should* drop them.

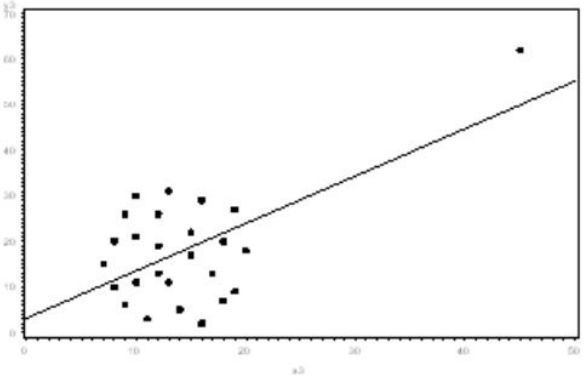
Outliers - To remove or not to remove?



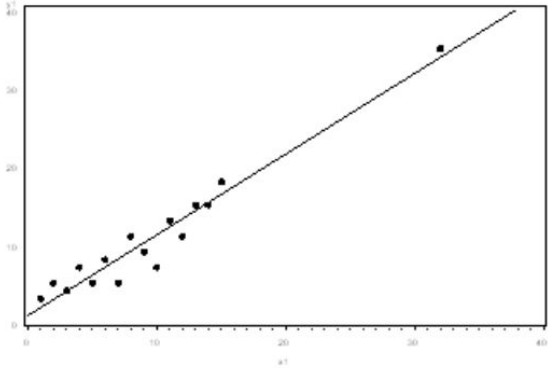




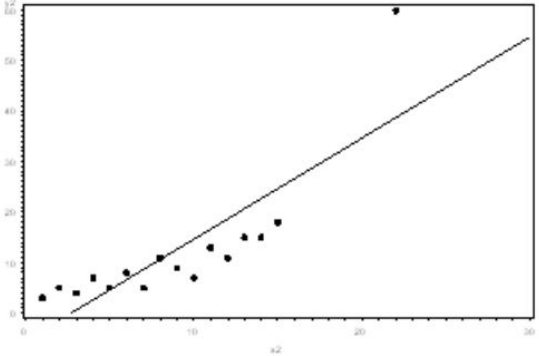
Outliers- To remove or not to remove?



The outliers do *not* change your results. In this case, it is okay to drop them, but it is best to make a note of having done so.



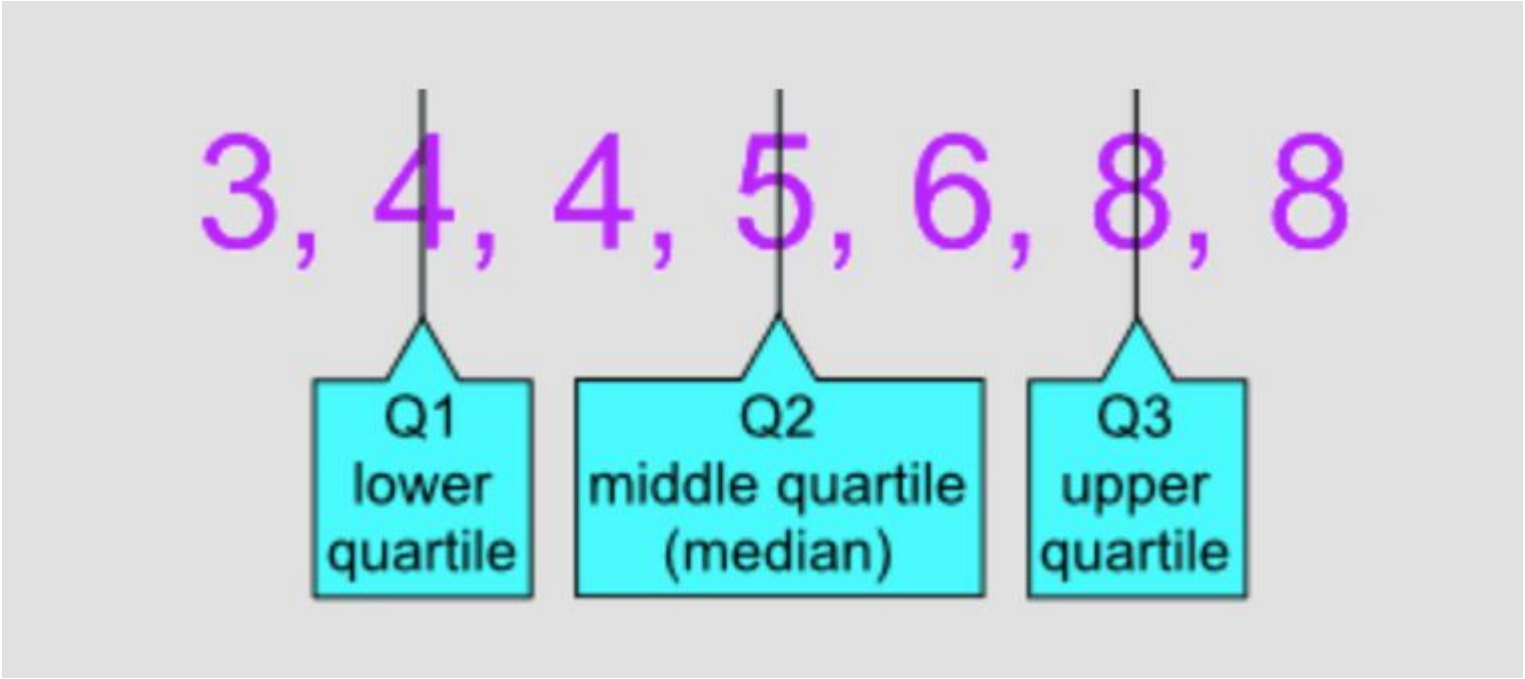
If the outliers *create* trends that wouldn't exist without them, you *should* drop them



If your outlier *does* change your results, you *should not* drop it

http://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/

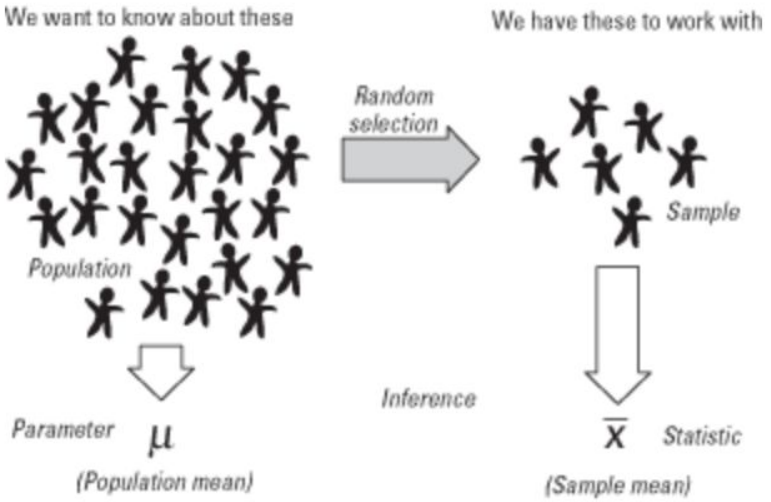
Outliers- Seem ‘fuzzy’?



Integrity of Data

Do we trust our data? SEM and Error Bars

We will focus on the notions of standard error and how well measurements of a section of a population (e.g., voting habits of Americans in cities) represent that population as a whole (e.g., all Americans).



This process can lead to imperfect predictions….why?

This process can lead to imperfect predictions.why?

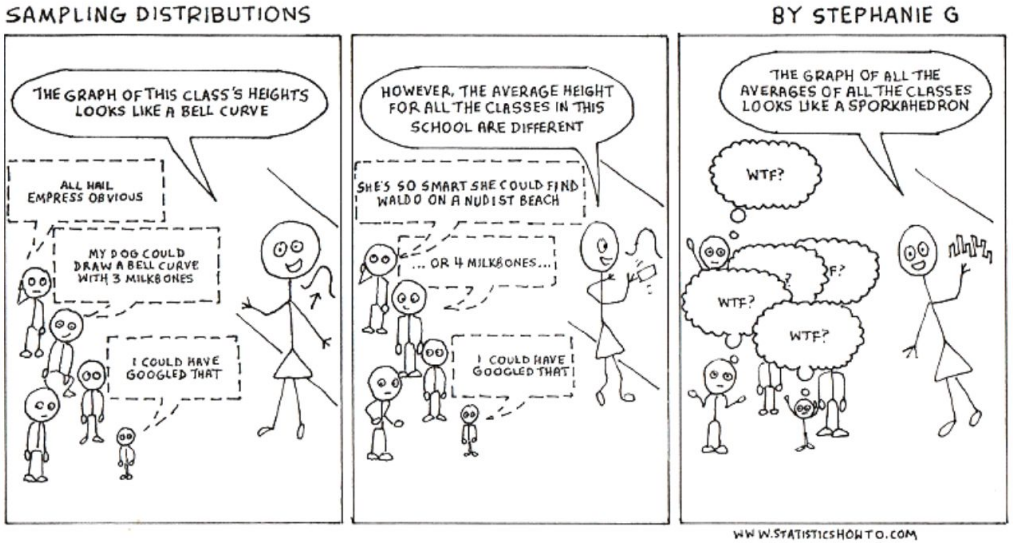
- A small sample might not realistically represent a large population, and that even a large sample, if chosen poorly, might not describe the population at large.
- For example, if we poll only college students and people who live in cities, we will almost certainly make bad predictions, because this sample does not represent the entire population.

Do we trust our data? SEM and Error Bars

- Let's say we read election predictions from two different sources.
  a. One source made predictions from a sample of 1,000 people;
  b. The other made predictions from a sample of 10,000.
  c. Which we would trust more, and why?

Do we trust our data? SEM and Error Bars

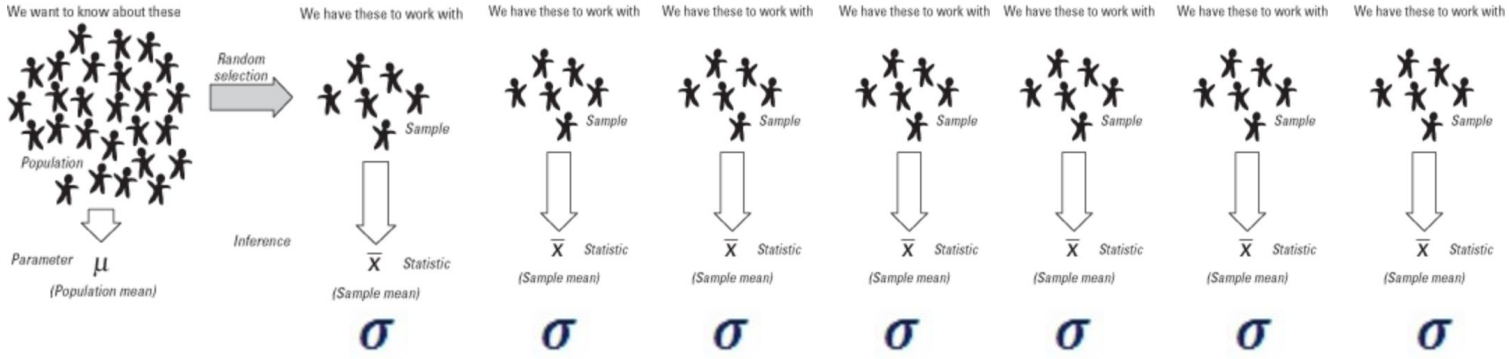
- Since a sample is just a subset of the whole population, we can poll *multiple* samples to get more accurate estimates of how the population behaves.



http://www.statisticshowto.com/sampling-distribution/

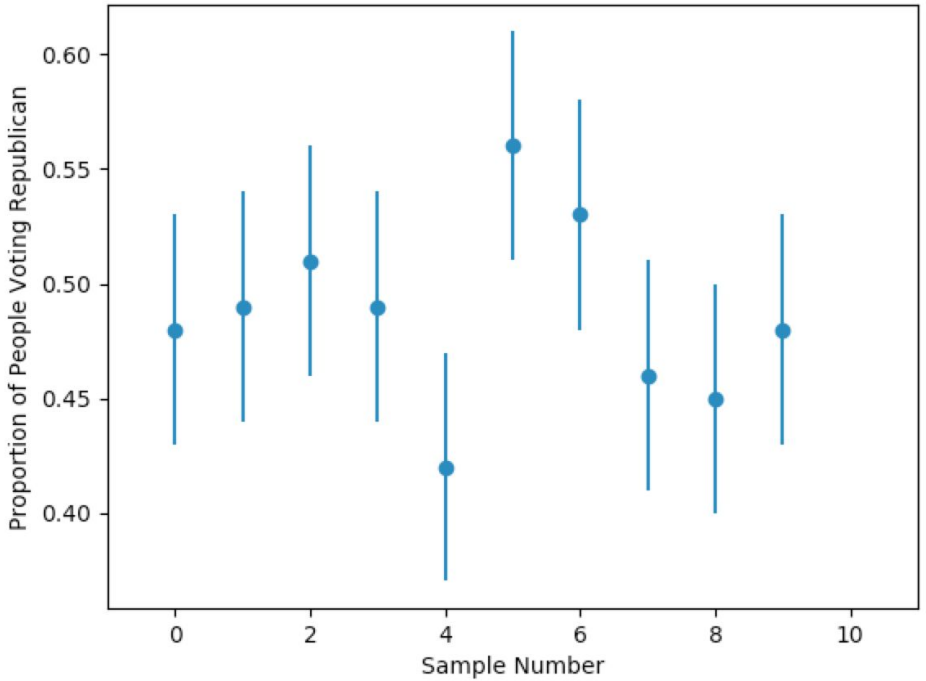
Do we trust our data? SEM and Error Bars

- If we take multiple samples, the collection of samples *themselves* is a data set.
- If we have 100 samples, we create a list of the samples' standard deviations.
- We can use these numbers to calculate something called standard error, which is an estimate how well the samples represent the population.
- Each sample's *standard error* describes how far its mean is from the **population's** "true" mean.
- The formula is unimportant — there is a [function in SciPy] that does this for us.



SEM and Error Bars - Example

- Most of the error bars overlap
- **Summary**: errorbars to provide a visual indicator as to how confident we were in the proximity of our sample means to the "true" population mean.



Student’s t-test

Student's t-test

- If we plotted the errorbars of our data and found that most of them did not overlap, it would raise questions about the data.
- Suppose, for instance, that about half of the means had error bars that overlapped one another, and the other half had error bars that overlapped one another, but neither cluster's error bars overlapped the other's.
- In this case, we might expect that the two clusters *were not* randomly selected from the same population

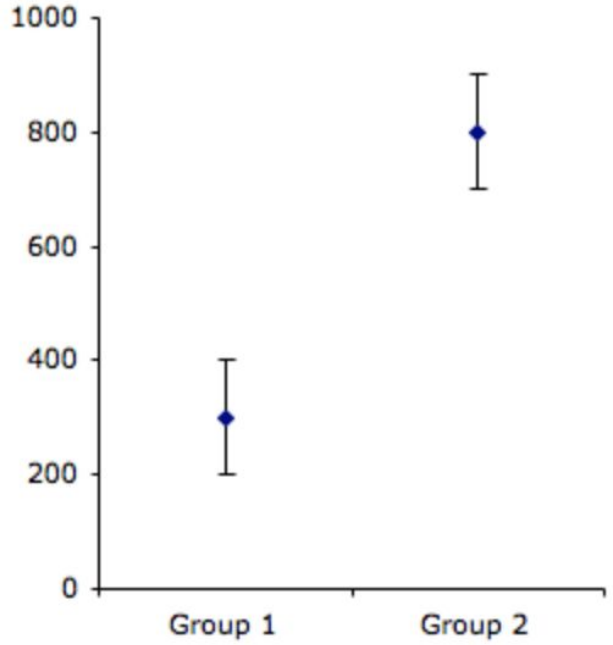
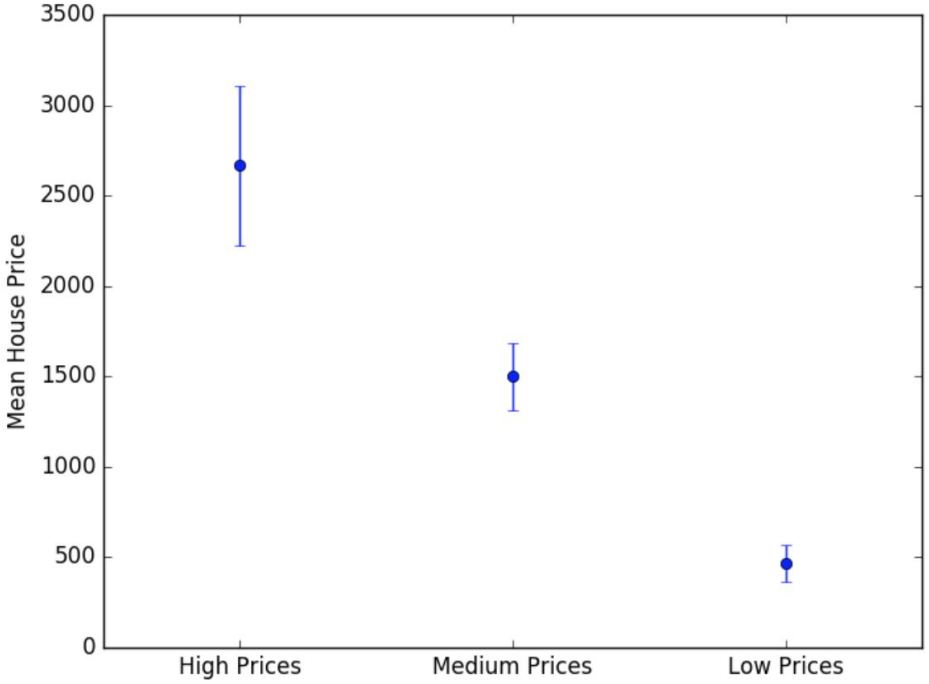


Figure 2: Mean reaction time (ms) and standard error for Group 1 (n=36) and Group 2 (n=34).

Student's t-test

- One way we might end up with an effect like this would be to generate one sample with prices from a rich neighborhood; another with prices from a poor neighborhood; and one with prices from a middle-class neighborhood.
- In this case, we would end up with three wildly different means, with error bars that (probably) would not overlap.



Fits & Regression

Regression

- Regression analysis allows us to take a data set and "reverse engineer" an equation describing it.
- Measures like the median, variance, and IQR *describe* data sets, but do not allow us to make *predictions* with it.
- It is tools like regression that allow us to predict where data points we *did not* measure might end up if we *had* collected more data.