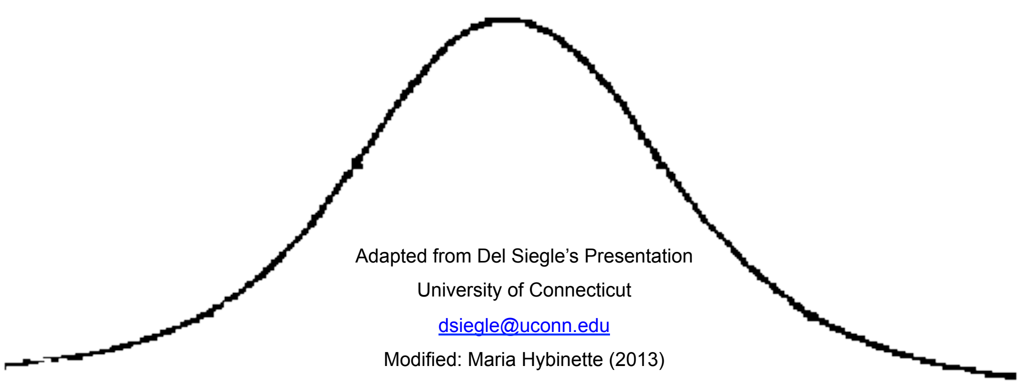
Today

- Today: Review statistics
  - In Excel
  - Goal: Understand the Normal Distribution (adapted from Siegle's Cornell Lecture)
- Don't forget: Summary Due Friday
  - Last Chapter 6: Names.
- Next week will continue with: Super Freakonomics
  - Chapter 1 Not due until Monday b/c Fall Break
- Attendance Today.

Understanding the Normal Curve



Adapted from Del Siegle's Presentation

University of Connecticut

dsiegle@uconn.edu

Modified: Maria Hybinette (2013)

Motivating Data Set: Foot Sizes

Our Simplified Study:
- 30 Professors
- Measured their foot size

Popular Press:
- **2009 Study of Foot Sizes** (differences between right and left foot)

Popular Press:
- **USA Today: American Shoe Sizes**
- **Daily Mail 2009**

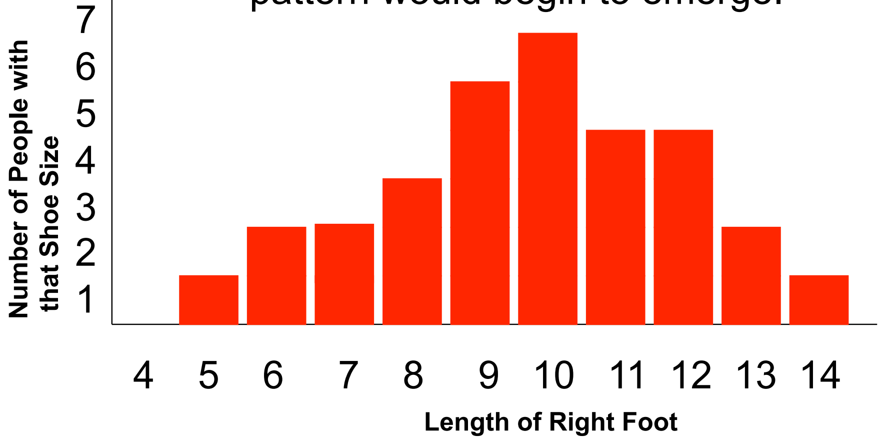
http://en.wikipedia.org/wiki/Shoe_size

Suppose we measured the **right** foot length of 30 UGA female professors and graphed the results.

Assume the first person had a 10 inch foot. We could create a bar graph and plot that person on the graph.

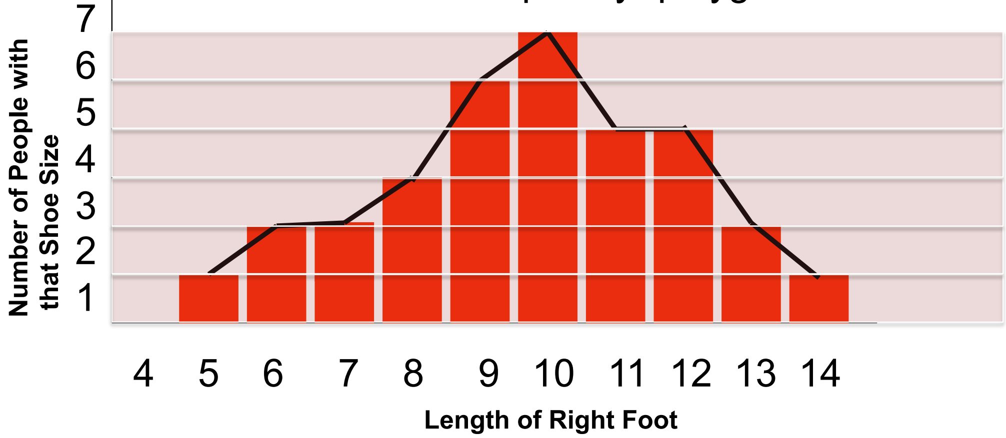
If our second subject had a 9 inch foot, we would add her to the graph.

As we continued to plot foot lengths, a pattern would begin to emerge.

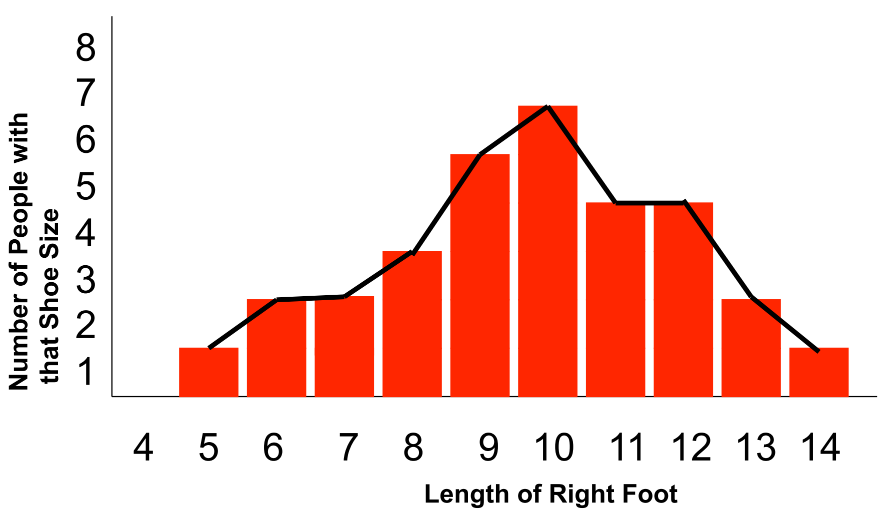


Notice how there are more people (n=6) with a 10 inch right foot than any other length. Notice also how as the length becomes larger or smaller, there are fewer and fewer people with that measurement. This is a characteristics of many variables that we measure. There is a tendency to have most measurements in the middle, and fewer as we approach the high and low extremes.

If we were to connect the top of each bar, we would create a frequency "polygon."

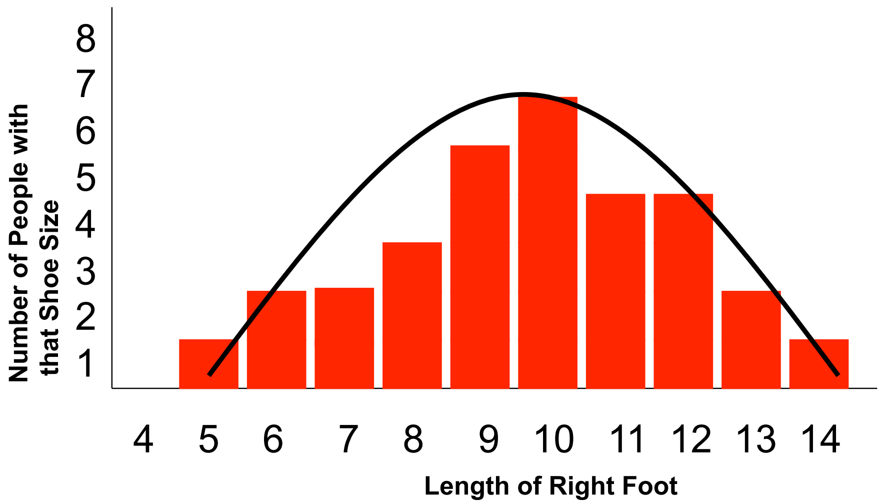


You will notice that if we smooth the lines, our data almost creates a bell shaped curve.



You will notice that if we smooth the lines, our data almost creates a bell shaped curve.

This bell shaped curve is known as the "Bell Curve" or the "Normal Curve."



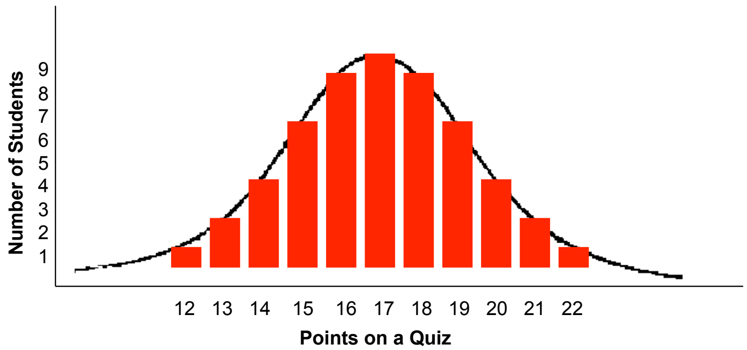
Revisiting: Real Life Foot Sizes

US Foot Size are getting Bigger

- 20th Century Average women wore size 3.5
- 40s-60s: size 5.5 (8 13/16 inches)
- 70s: size 7.5 (9 ½ inches)
- 2000s: size 8.5-9 (9 13/16 - 10 inches)
- **Why**: People are getting taller and heavier, feet are getting larger in proportion of their bodies (Podiatric Historian William Rossi).

http://newsfeed.time.com/2012/10/16/size-8-is-the-new-7-why-our-feet-are-getting-bigger/

Whenever you see a normal curve, you should imagine the bar graph within it.



Now lets look at quiz scores for 51 students.

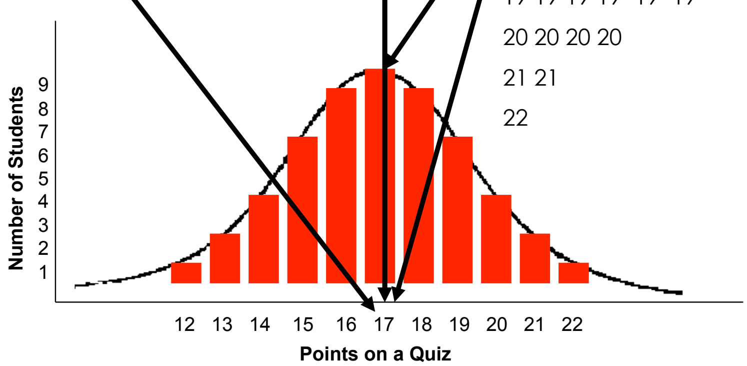
The mean, mode, and median will all fall on the same value in a normal distribution.

12+13+13+14+14+14+14+15+15+15+15+15+15+16+16+16+16+16+16+16+
17+17+17+17+17+17+17+17+17+18+18+18+18+18+18+18+18+19+19+19+19+
19+ 19+20+20+20+20+ 21+21+22 = 867

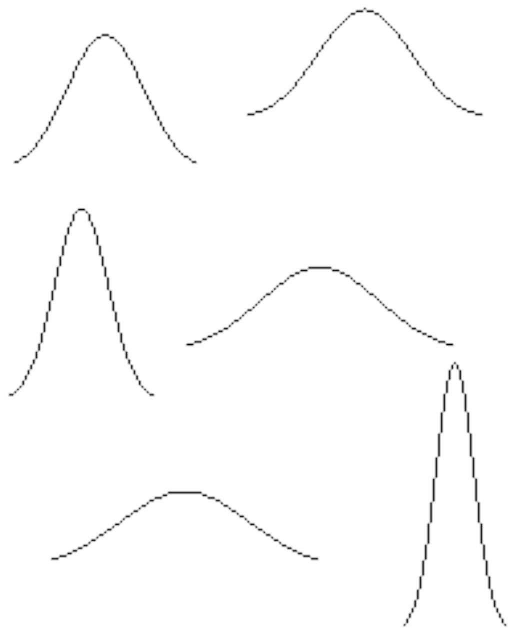
12, 13, 13, 14, 14, 14, 14, 15, 15, 15, 15, 15, 15, 16, 16, 16, 16, 16, 16, 16, 17, 17, 17, 17, 17, 17, 17, 17, 17, 18, 18, 18, 18, 18, 18, 18, 18, 19, 19, 19, 19, 19, 19, 20, 20, 20, 20, 21, 21, 22

867 / 51 = 17

12
13 13
14 14 14 14
15 15 15 15 15 15
16 16 16 16 16 16 16 16
17 17 17 17 17 17 17 17 17
18 18 18 18 18 18 18 18
19 19 19 19 19 19
20 20 20 20
21 21
22



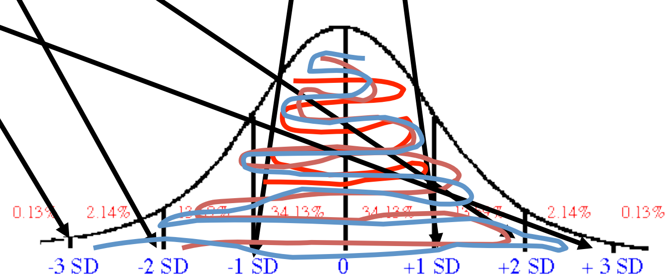
Normal distributions (bell shaped) are a family of distributions that have the same general shape. They are **symmetric** (the left side is an exact mirror of the right side) with scores more concentrated in the middle than in the tails. Examples of normal distributions are shown to the right. Notice that they differ in how **spread** out they are. The area under each curve is the same.



If your data fits a normal distribution, approximately 68% of your subjects will fall within one standard deviation of the mean.

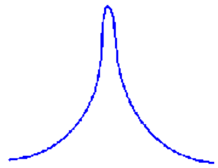
Approximately 95% of your subjects will fall within two standard deviations of the mean.

Over 99% of your subjects will fall within three standard deviations of the mean.



Normal Distribution

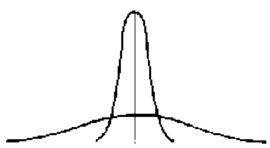
The mean and standard deviation are useful ways to describe a set of scores. If the scores are grouped closely together, they will have a smaller standard deviation than if they are spread farther apart.



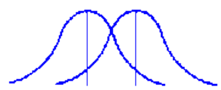
Small Standard Deviation

Large Standard Deviation

Variety of pairs of normal distributions are plotted below.



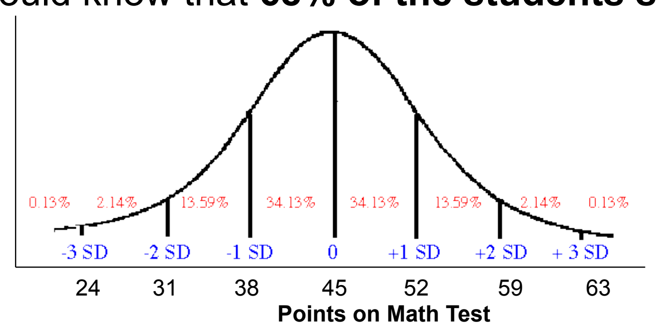
Same Means
Different Standard Deviations



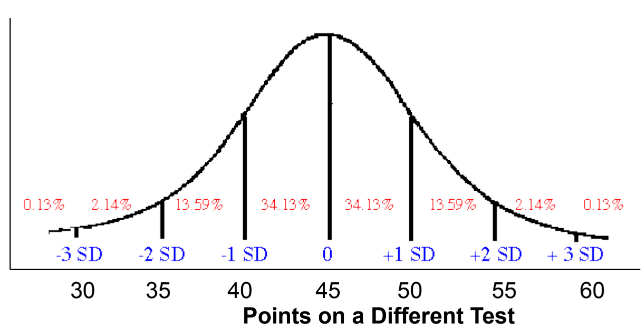
Different Means
Same Standard Deviations

Different Means
Different Standard Deviations

The number of points that one standard deviations equals varies from distribution to distribution. On one math test, a **standard deviation may be 7 points**. If the **mean were 45,** then we would know that **68% of the students scored from 38 to 52.**

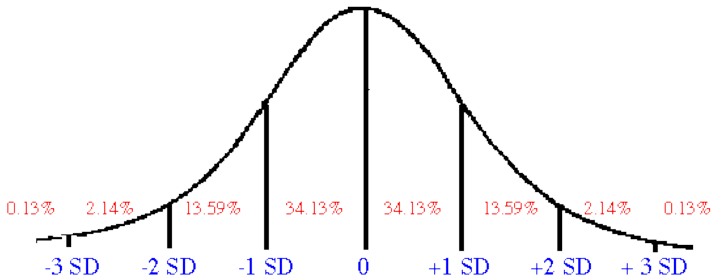


On another test, a **standard deviation may equal 5 points**. If the **mean were 45**, then 68% of the students would score from **40 to 50 points**.



Standardized Scores Motivation

When you have a subject's raw score, you can use the mean and standard deviation to calculate his or **her standardized score** if the distribution of scores is normal. Standardized scores are useful when **comparing a student's performance across different tests**, or **when comparing students with each other**. Your next assignment will involve calculating and using standardized scores.



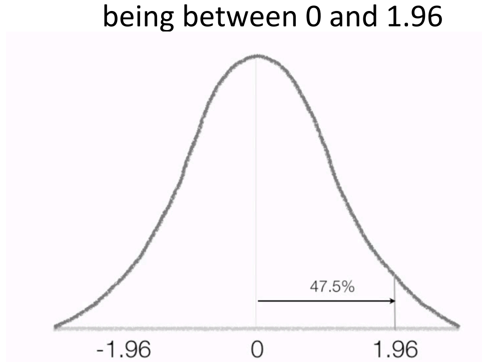
z-score	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**
T-score	20	30	40	50	60	70	80
IQ-score	65	70	85	100	115	130	145
SAT-score	200	300	400	500	600	700	800

Standardizing Scores

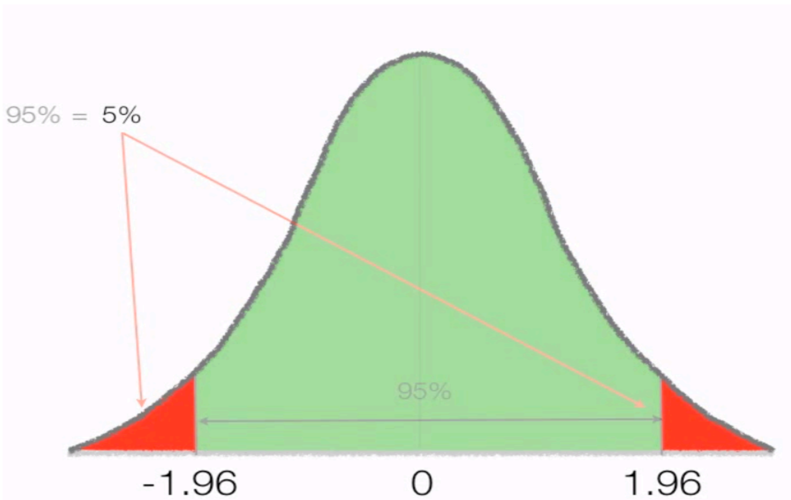
Z-Scores

- Z-score is the distance between the mean and an observation
  - Distance is measured in standard deviations (see previous slide)
- Z-score is 0 at the mean

- Z-score of 1.96
  - Probability of an observation being between 0 and 1.96



  - Probability an observation between -1.96 to +1.96 is 95%



Z-Score Calculations

- Z = (observation - mean)/ (standard deviation)

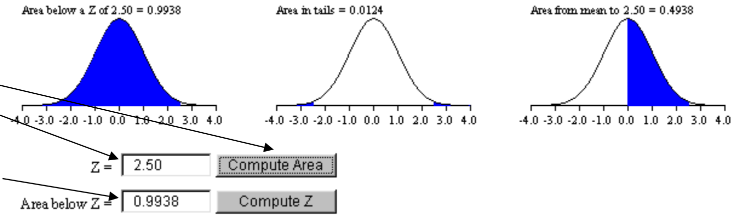
$$z = \frac{x - \mu}{\sigma}$$

- Z scores tells you how far you are form the mean (previous slide) and is measured in standard deviations
- **Example**: Link Below

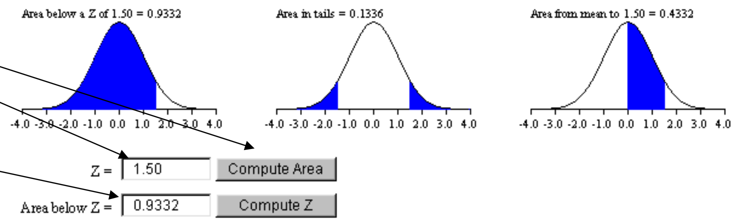
http://www.measuringusability.com/zcalc.htm

Assuming that we have a normal distribution, it is easy to calculate what percentage of students have z-scores between 1.5 and 2.5. To do this, use the Area Under the Normal Curve Calculator at http://davidmlane.com/hyperstat/z_table.html.

Enter 2.5 in the top box and click on *Compute Area*. The system displays the area below a z-score of 2.5 in the lower box (in this case .9938)



Next, enter 1.5 in the top box and click on *Compute Area*. The system displays the area below a z-score of 1.5 in the lower box (in this case .9332)

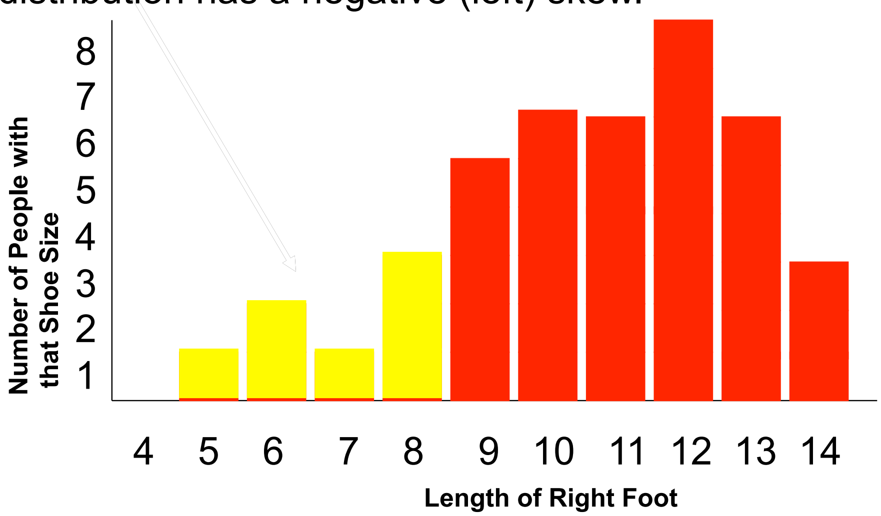


If .9938 is below z = 2.5 and .9332 is below z = 1.5, then the area between 1.5 and 2.5 must be .9939 - .9332, which is .0606 or 6.06%. Therefore, 6% of our subjects would have z-scores between 1.5 and 2.5.

Data do not always form a normal distribution. When most of the scores are high, the distributions is not normal, but negatively (left) skewed.

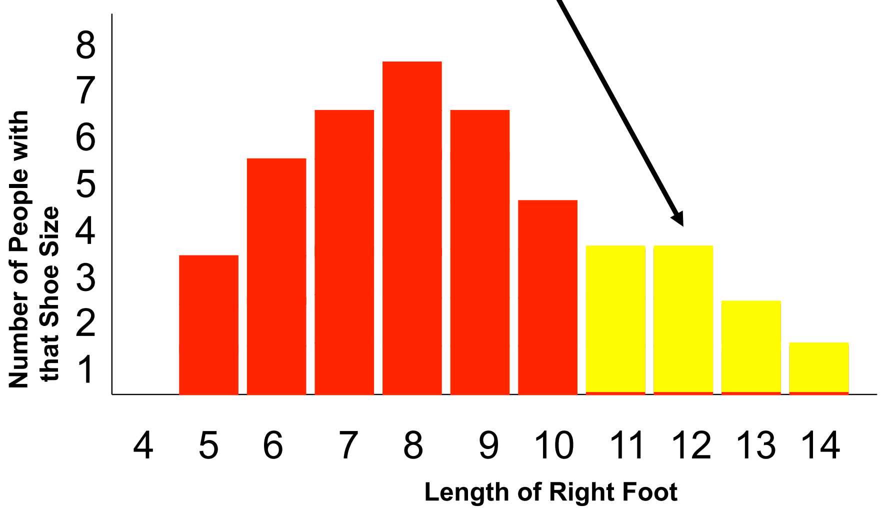
Skew refers to the tail of the distribution.

Because the tail is on the negative (left) side of the graph, the distribution has a negative (left) skew.

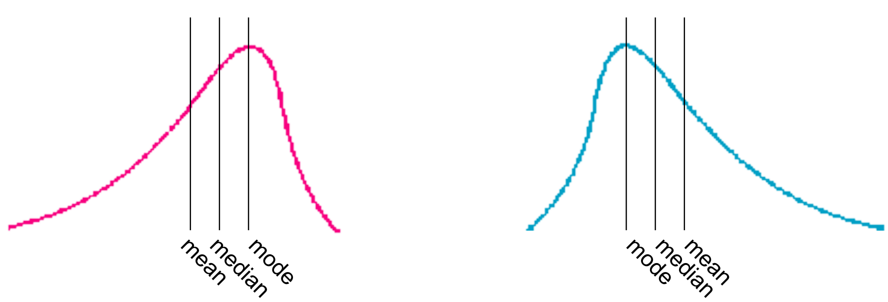


When most of the scores are low, the distributions is not normal, but positively (right) skewed.

Because the tail is on the positive (right) side of the graph, the distribution has a positive (right) skew.



When data are skewed, **they do not possess the characteristics of the normal curve (distribution)**. For example, 68% of the subjects do not fall within one standard deviation above or below the mean. The mean, mode, and median do not fall on the same score. The mode will still be represented by the highest point of the distribution, but the mean will be toward the side with the tail and the median will fall between the mode and mean.



Negative or Left Skew Distribution

Positive or Right Skew Distribution

Mathematical Formula for Height of a Normal Curve

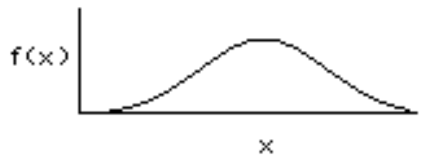
The height (ordinate) of a normal curve is defined as:

$$\frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-(x-\mu)^2/2\sigma^2}$$

where μ **is the mean** and σ **is the standard deviation**, π **is the constant 3.14159**, and **e is the base of natural logarithms and is equal to 2.718282**.

x can take on any value from -infinity to +infinity.

f(x) is very close to 0 if **x is more than three standard deviations** from the mean (less than -3 or greater than +3).



Don't need to memorize formulas.