Welcome to OSA Training 2015 Statistics Part II

By:

Greg Hinckson

Mitch Volk

Dr. Sybil DeVeaux

Iris Bishop

Course Summary

- Using data about a population to draw graphs
- Frequency distribution and variability within populations
- Bell Curves: What are they and where do we see them?
- Normal distribution
- Skewness in Curves
- Interpreting bell curves by their mean, variance, and standard deviation
- Inter-Quartile range
- Understanding and calculating Z scores
- Proportion: Calculating the area under the curve
- Correlation: What is the relationship between two variables?

Using data to draw graphs about a population

- A **statistic** is a way to represent or organize information in a way that helps you understand it better than simply looking at a series of numbers.

- You can use a set of data to draw a picture that will help you to understand and interpret that data.
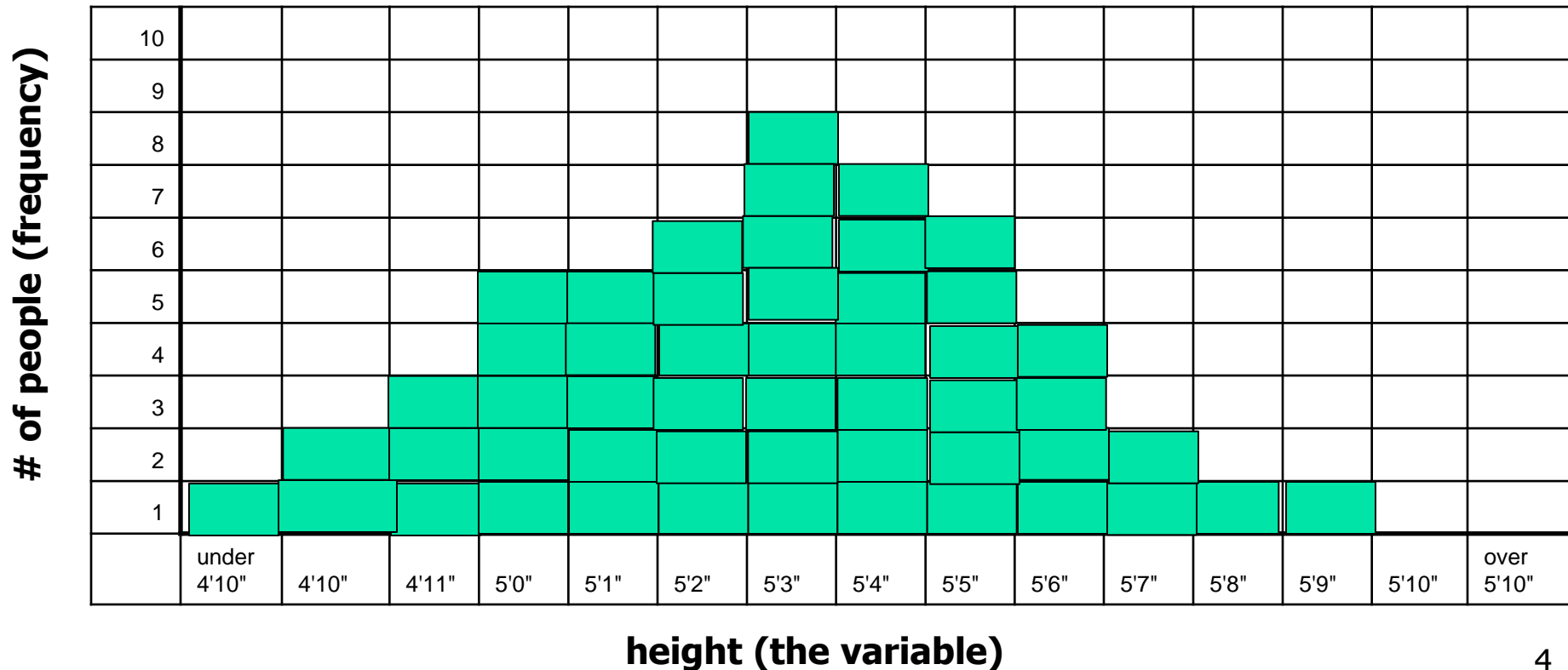
Using Data to Draw Graphs:
In-Class Exercise of Height Frequency Distribution

Instructions: Fill in the graph according to the results in class.

Frequency Distribution of Women's Height

height (the variable)

Using Data to Draw Graphs:
In-Class Exercise of Height Frequency Distribution

Instructions: Fill in the graph according to the results in class.

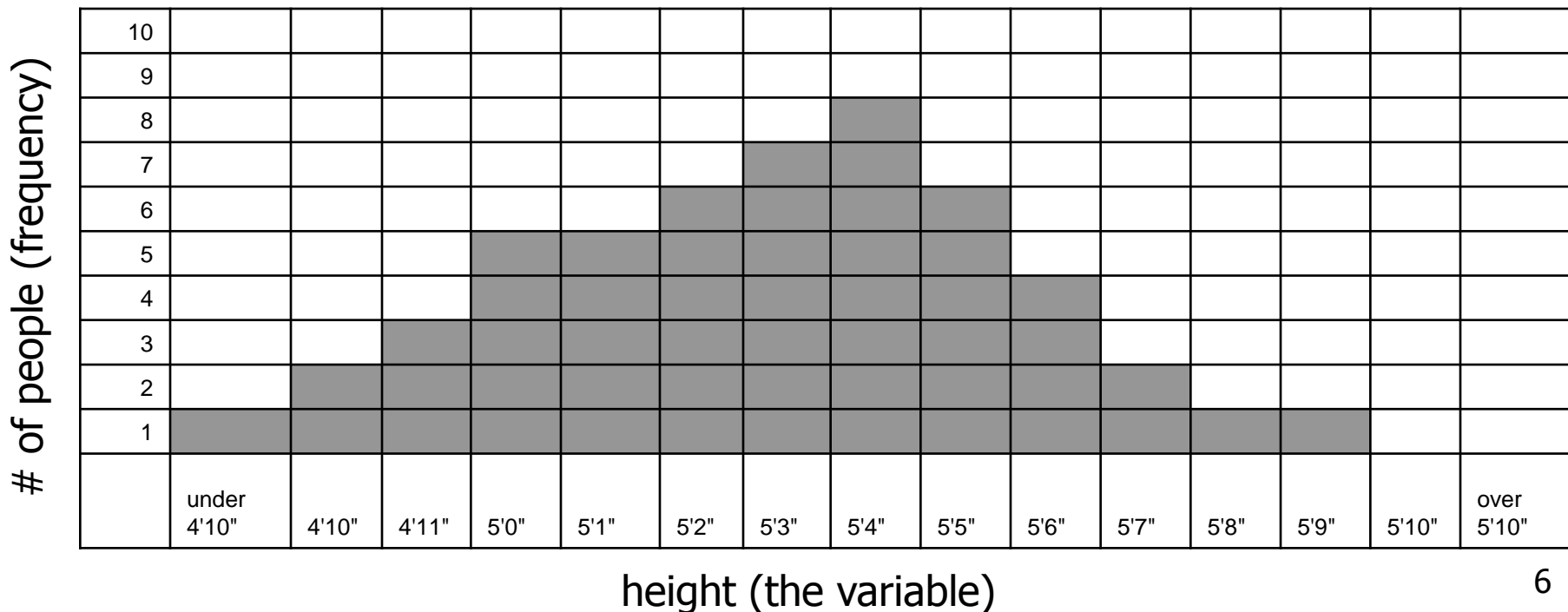**Frequency Distribution of Men's Height**

# of people (frequency)															
10															
9															
8															
7															
6															
5															
4															
3															
2															
1															
	under 5'3"	5'3"	5'4"	5'5"	5'6"	5'7"	5'8"	5'9"	5'10"	5'11"	6'0"	6'1"	6'2"	6'3"	over 6'3"

height (the variable)

Example: Height distribution among a group of 55 women

• The X axis (horizontal) refers to the variable, or the observation value that you are looking at in a population.

• The Y axis (vertical) reflects the frequency, or the number of times a particular value of X appears in a population.

Example of the Frequency Distribution of Women's Height

# of people (frequency)	under 4'10"	4'10"	4'11"	5'0"	5'1"	5'2"	5'3"	5'4"	5'5"	5'6"	5'7"	5'8"	5'9"	5'10"	over 5'10"
10															
9															
8									8						
7								7							
6							6			6					
5				5	5	5									
4											4				
3			3									3			
2		2													
1	1												1	1	

height (the variable)

6

Properties of Populations

Population

- A population is any group whose characteristics you look at. A population is different from a sample, which is a small portion of the population used to generalize about the whole population.

Central Tendency

- Large populations often tend to cluster towards their middle, or average, which is also known as the **mean.**

Variability

- In large populations, there is often a lot of diversity. For example, people come in a variety of heights and weights.

Example: The Hungry Softball Team

Situation

A softball team has just won a game. All 20 players on the team – the population – have gone to eat pizza.

Graph

A simple graph shows how many slices each of the 20 team members ate. For example, four people ate 5 slices of pizza, while only one person ate 8 slices.
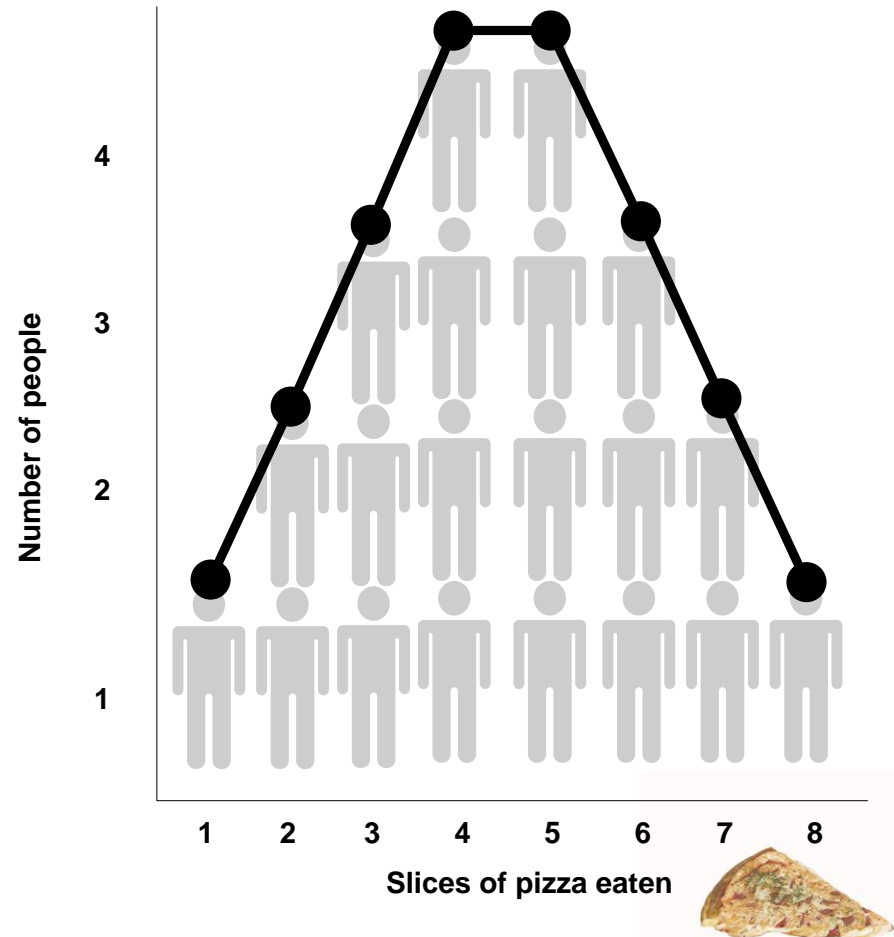

8
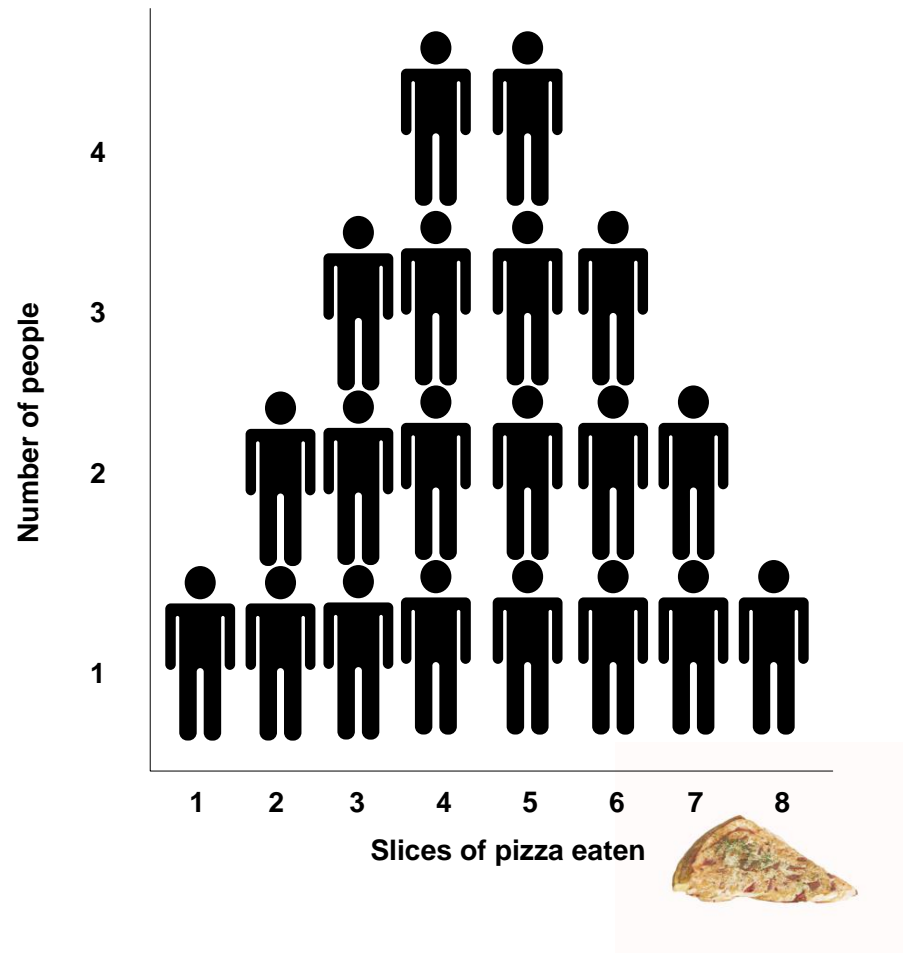
Example:
The Hungry Softball Team

Graph

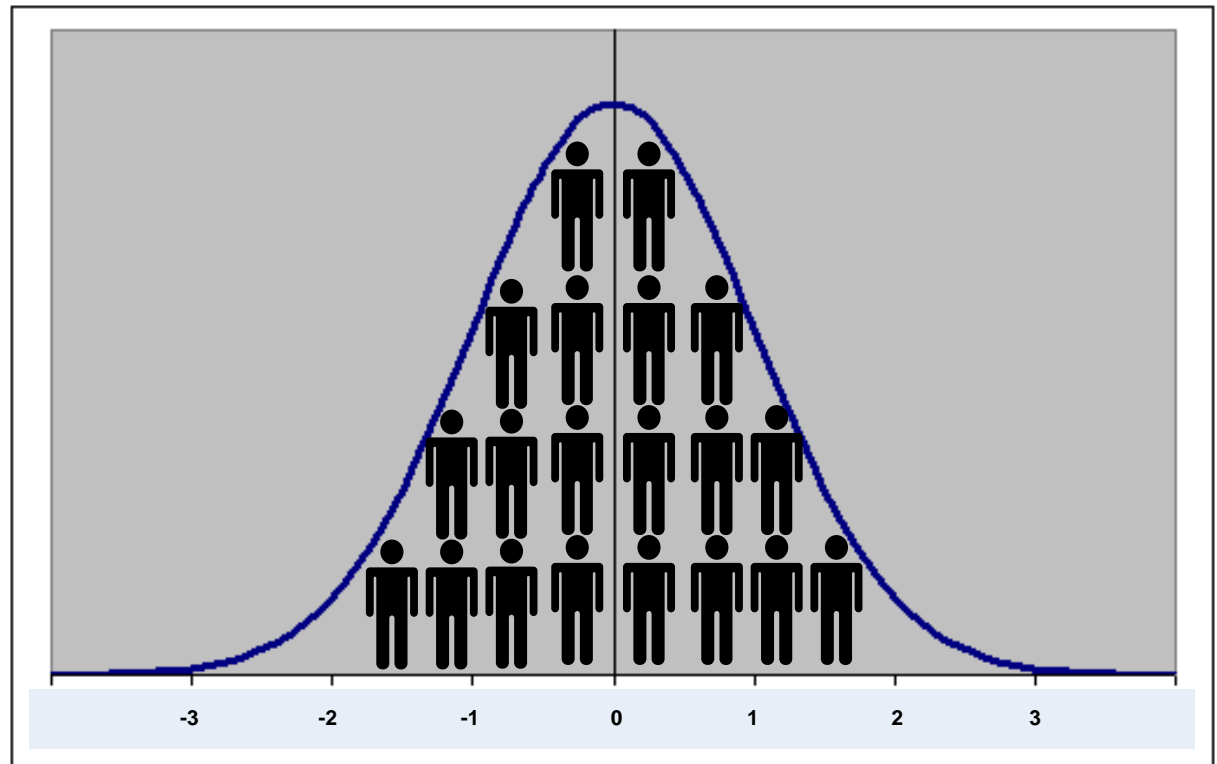A line shows how you could draw a simple graph using the tops of the heads of each group of players.

Example:
The Hungry Softball Team

Graph

This graph is a simplification of how you could graph pizza slices eaten into a bell curve.

Standard Deviations from the Mean

Bell Curves: What are they?

The Bell Curve
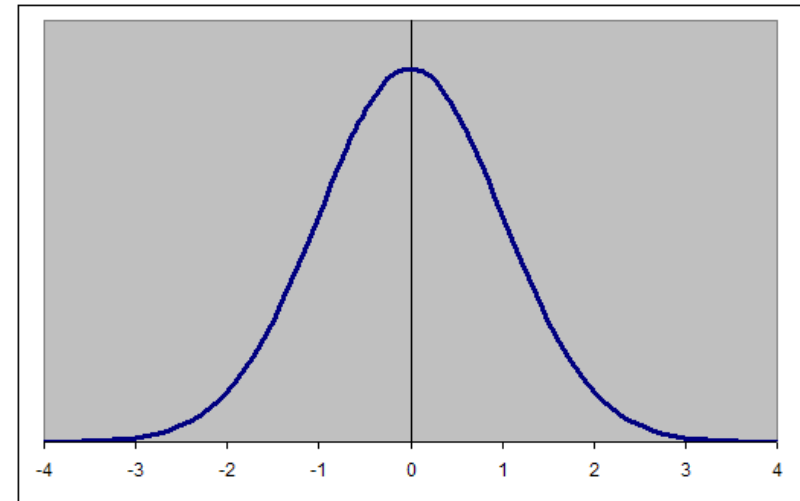
Basic Properties

- A bell curve is a very special kind of curve with unique properties.
- It is shaped like a bell.
- Also called a "normal curve" or "normal distribution," it shows how frequently different values recur in a population.
- It is symmetric and has a single peak <u>at its mean.</u>
- Its unique properties make it very useful in making statistical calculations.

Bell Curves:
Where do we see them?

- Normal distributions occur often, especially when a large group of data is concerned.
- Examples:
  - Height
  - Weight
  - SAT scores
  - IQ
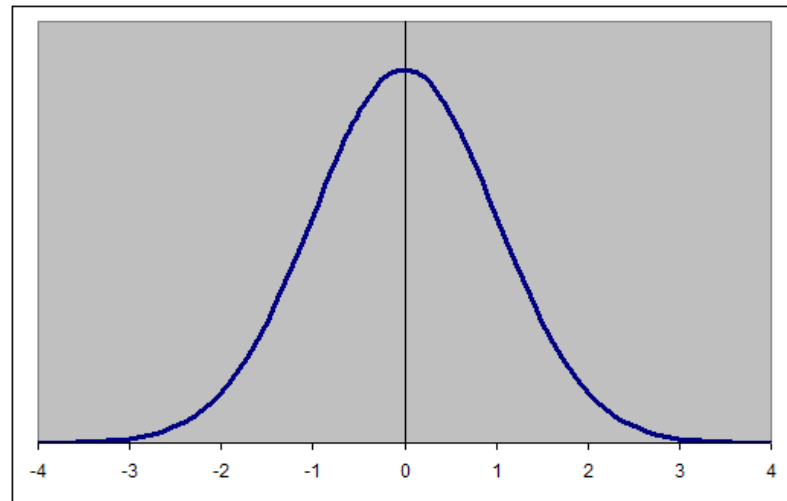
Bell Curves: Where do we see them?

- Example: fish size
- This diagram illustrates how MOST fish in a given species fall pretty close to the average
- Very small or large fish – called **outliers** because of their uncommon size – are much more rare and show up on one end of the bell curve.
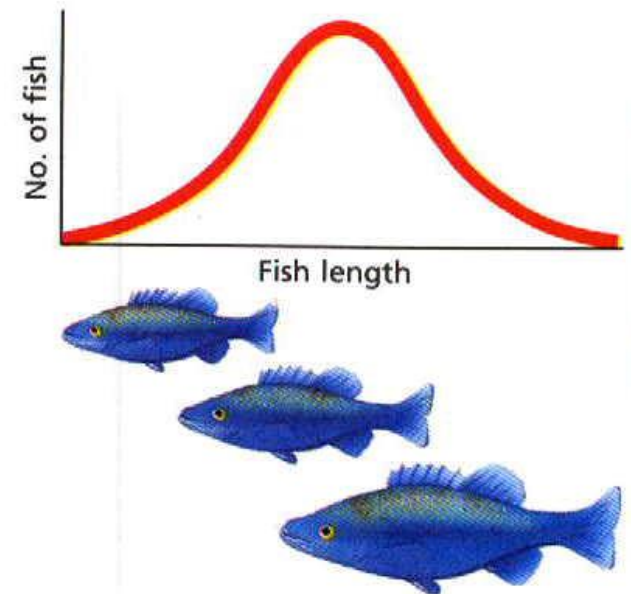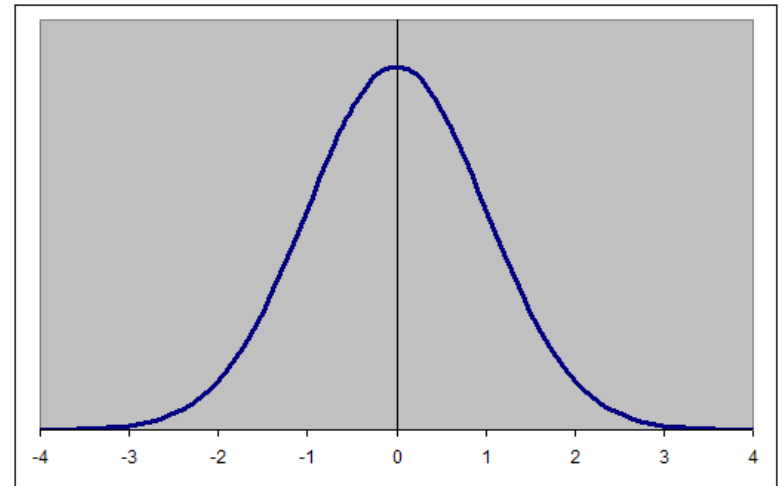
No. of fish

Fish length

Figure 16-2 A bell curve illustrates how most members of a population are grouped in an average range for a given trait while only a few are at the extreme ends of the range.

Bell Curves: Mean

Mean

- The mean shows the average of all the values in a population.

Mean = Addition of all the observations together
# of observations

The Mean

Sample Mean	Population Mean
$\bar{x} = \dfrac{\Sigma x}{n}$	$\mu = \dfrac{\Sigma x}{N}$

where Σx is sum of all data values

N is number of data items in population

n is number of data items in sample

The Mean Continued

- The arithmetic mean is a simple type of average. Suppose you want to know what your numerical average is in your math class. Let's say your grades so far are 80, 90, 92, and 78 on the four quizzes you have had.

The Mean Continued

- 80 + 90 + 92 + 78 = 340

- Then divide that answer by the number of grades that you started with, four: 340 / 4 = 85

- So, your quiz average is 85! Whenever you want to find a mean, just add up all the numbers and divide by however many numbers you started with.

On Average...........

- **The Mode a measure of the most frequently seen observation**

- **Q: What is the mode gender of the class?**

- **2, 4, 6, 0, 4, 1, 4,**

On Average……

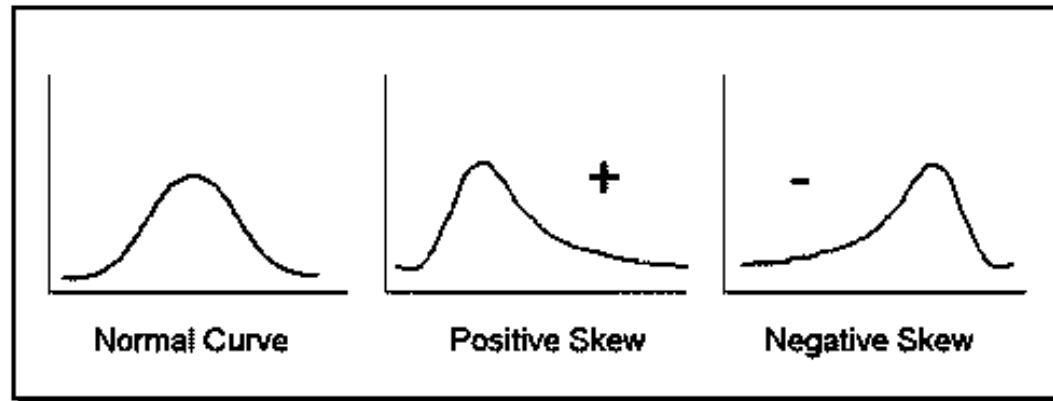- **Median** reflects the middle ranked value when observations are ordered least to greatest or vice versa

- **2, 2, 6, 7, 8**

  ↑

  median

- **1, 3, 3, 8, 8, 9**

  ↑

  median

  3+8/2=5.5

- The **skew** of a distribution refers to how the curve leans.

- When a curve has extreme scores on the right hand side of the distribution, it is said to be positively skewed. In other words, when high numbers are added to an otherwise normal distribution, the curve gets pulled in an upward or positive direction.

- When the curve is pulled downward by extreme low scores, it is said to be negatively skewed. The more skewed a distribution is, the more difficult it is to interpret.[1]
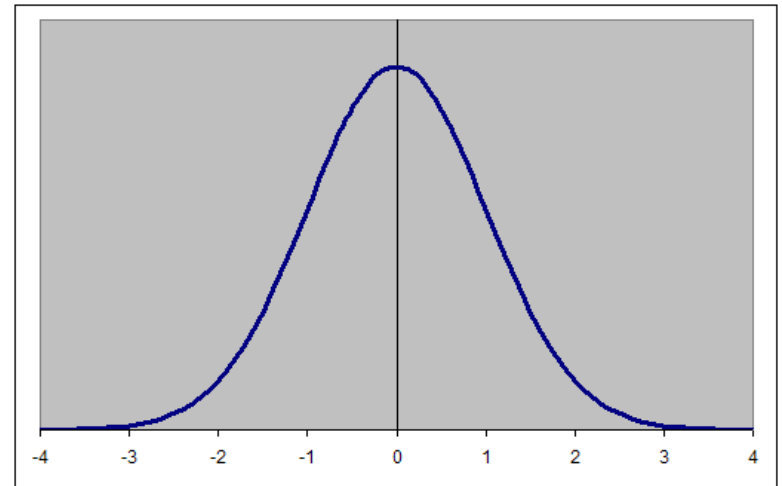

| Normal Curve | + Positive Skew | − Negative Skew |

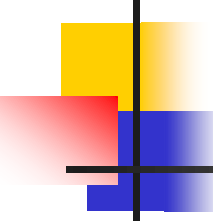[1] Text from: http://allpsych.com/researchmethods/distributions.html.

Bell Curves: Variance

Variance

- A measure of the variability of the population described by a bell curve.

- Calculated by adding together the square of the difference between EACH observation and the mean

$$\text{Variance} = \frac{\text{Sum of (each observation} - \text{mean)}^2}{\text{\# of all observations}}$$

- When you measure variability you are measuring the amount of difference among observations in a distribution such as differences in height among men.

- When you are looking at Standard Deviations you are asking how different is this observation from the mean. If the average height of women is 5'4" and Helen is 5'0 how far does she deviate from the mean?

Bell Curves Variance

$$s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n - 1}$$

$$s^2 - \text{Variance} - \frac{\sum_{i=1}^{n}(X_i - X)^2}{n} \quad \text{or} \quad \frac{\sum_{i=1}^{n}(X_i - X)^2}{n-1}$$

Variance cont'd…..

- 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 (Population M=2) and the variance is:

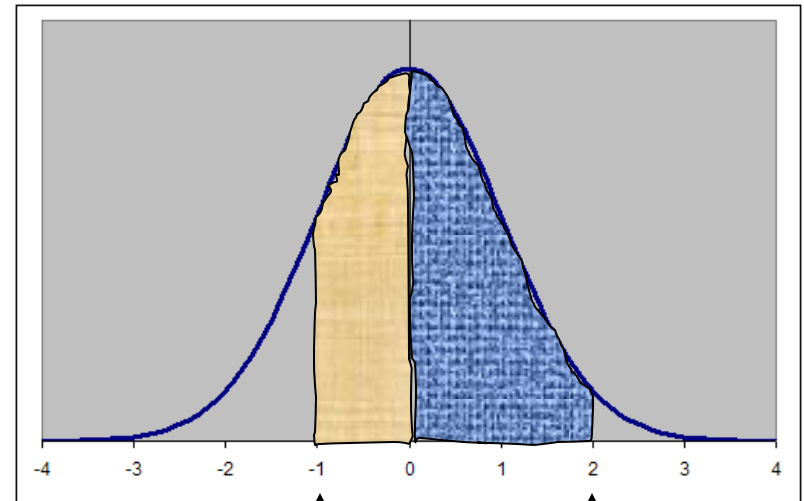$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Variance Calculation

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

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Bell Curves: Standard Deviation

Standard Deviation

- "A rough measure of the average amount by which observations deviate on either side of their means" (Witte & Witte, 2001)

- It's a way of measuring how far any observation is from the mean.

- In precise terms, it's the square root of the variance.

One standard deviation from the mean; Z = -1

Two standard deviations from the mean; Z = 2

Standard Deviation cont'd…

- The square root the variance is the standard deviation.

Standard Deviation Formula

$$\sigma = \sqrt{\frac{\sum (x - \overline{x})^2}{n}}$$

σ = lower case sigma

$\sum$ = capital sigma

$\overline{x}$ = x bar

Checking for Understanding

Two corporations hired 10 graduates. The starting salaries for each are shown in thousands of dollars. Find the deviation for the starting salaries of each corporation.

Corp A Salary	41	38	39	45	47	41	44	41	37	42

Corp B Salary	40	23	41	50	49	32	41	29	52	58

Inter-Quartile Range

The **inter-quartile range (IQR)**, measures the spread of the inner 50% of a data set.

Steps to find the *IQR:*

- Find Q_2 -the median
- Find Q_1 -the median of the lower half
- Find Q_3 -the median of the upper half
- $IQR = Q_3 - Q_1$

Checking for Understanding

Find the inter-quartile range for each corporation below.

Which corporation seems to have fairer starting salaries? Explain

Corp A Salary	41	38	39	45	47	41	44	41	37	42

| Corp B Salary	40	23	41	50	49	32	41	29	52	58

Bell Curves: What are they?

Advanced Properties

- They extend approximately 3 standard deviations above and below the mean.

- They have a total area under the curve of 1.00 (100%).

- The mean, median, and mode of a normal distribution are identical and fall exactly in the center of the curve.

- Do IQ's

Z Scores: What are they?

- Z-scores are a way to convert real data in the world into a form that fits on a bell curve.

- This only works if you have a normal distribution to begin with.

- IQ is a very standard example of a normal distribution that can be easily converted to Z scores.


Distribution of IQ scores
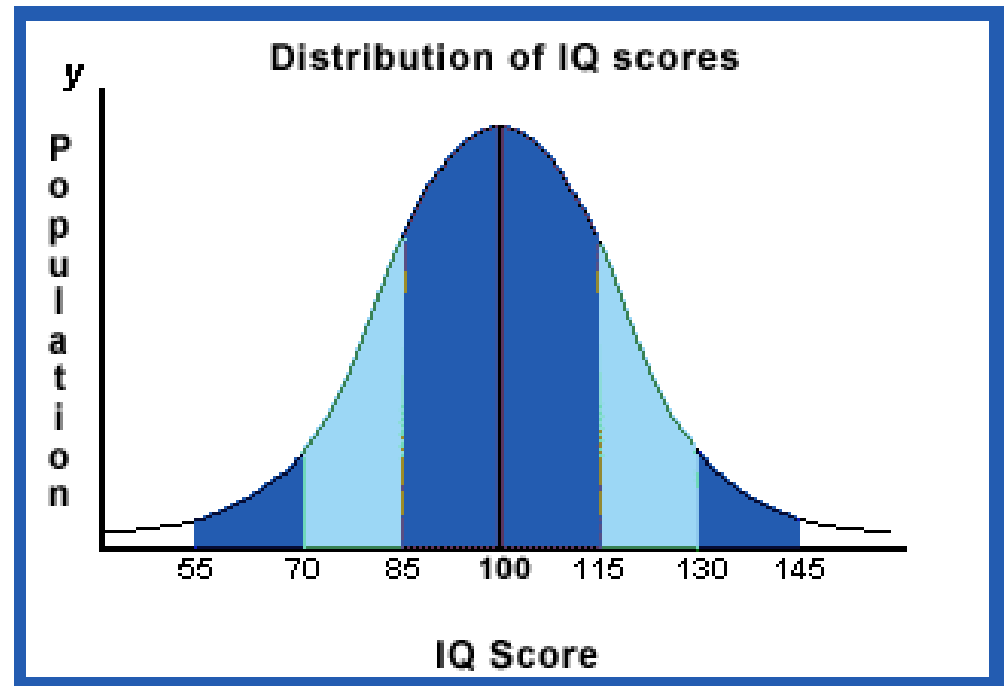
Z-Scores: What are they?

Mean and Standard Deviation

- The mean always has a z-score of 0.

- Other scores are converted to z-scores by their distance from the mean – how many standard deviations they are from the mean.

- The standard deviation is always equal to 1.

  - Example 1: z-score of -2 means that the value of an observation is two standard deviations from the mean (left).

  - Example 2: If a person's height is one standard deviation above the mean, the z-score for his or her height is equal to one (right).

34

Z-Scores:
How to Calculate Them
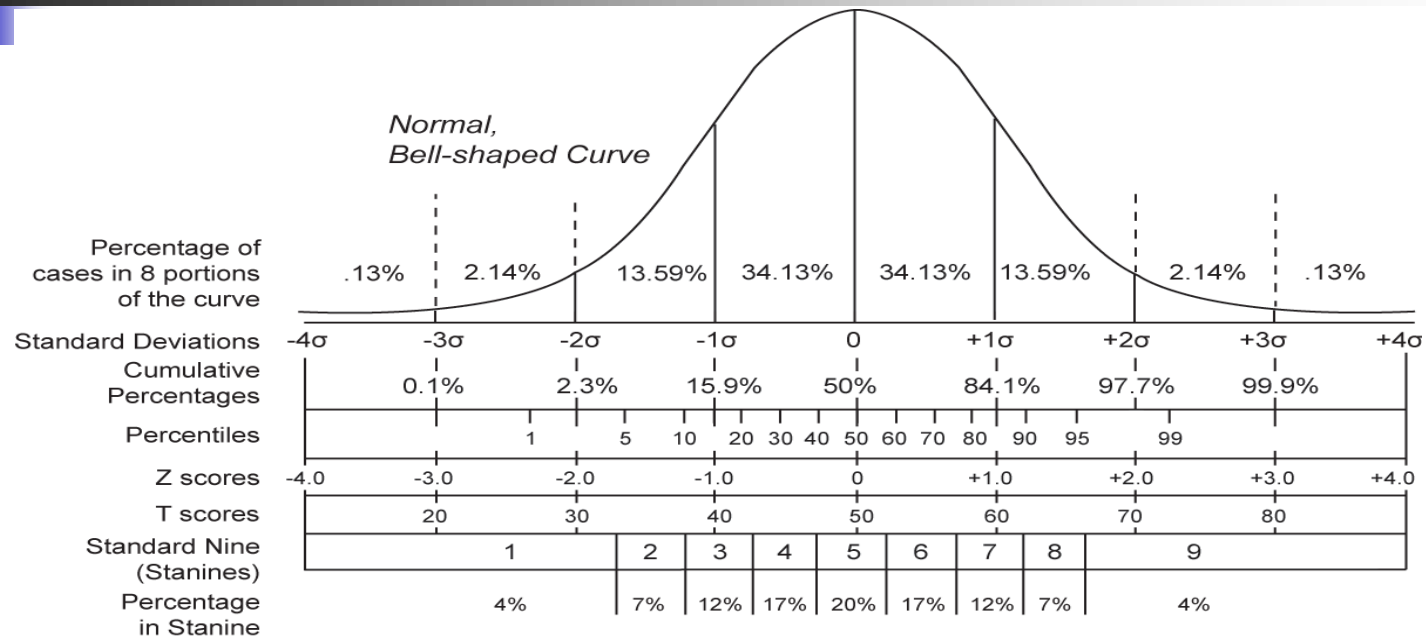
This bell curve reflects IQ (intelligence quotient). For IQ, the mean equals 100 and the standard deviation equals 15.

Z-score = **Observation Value – Mean**

Standard Deviation

Z-Scores

- Find the z-score corresponding to a raw score of 130 from a normal distribution with mean 100 and standard deviation 15.

Z-score = <u>Observation Value – Mean</u>

Standard Deviation

Z-Scores

- 130 - 100/15 = 2

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

$\mu =$ Mean

$\sigma =$ Standard Deviation

Z-Scores

- With an IQ score of 80 and a mean of 100, with a standard deviation of 15 what is the z-score?
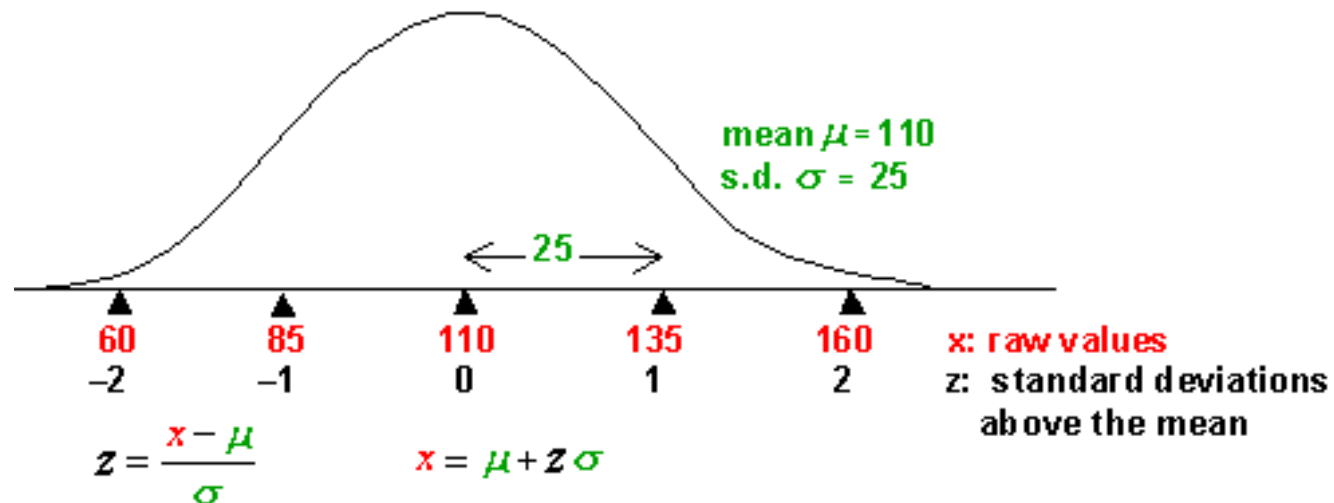
Z-Scores

$$\frac{80 - 100}{15} = -1.33$$

"Real" Data Compared with Z-Scores: Example

- The diagram below illustrates how you convert "real" numbers or "raw values" into z-scores.

- This example has a mean of 110 and standard deviation of 25.

- Again, when you have a normal ("bell") curve, you can always convert the numbers so that the mean is 0 and a standard deviation is equal to 1.

mean $\mu = 110$
s.d. $\sigma = 25$

$\longleftarrow 25 \longrightarrow$

60	85	110	135	160	x: raw values
–2	–1	0	1	2	z: standard deviations above the mean

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

Proportion

- The area under a bell curve tells you what percentage of ALL observations fall within that area.

- The total area under a bell curve is always equal to 1, or 100%.
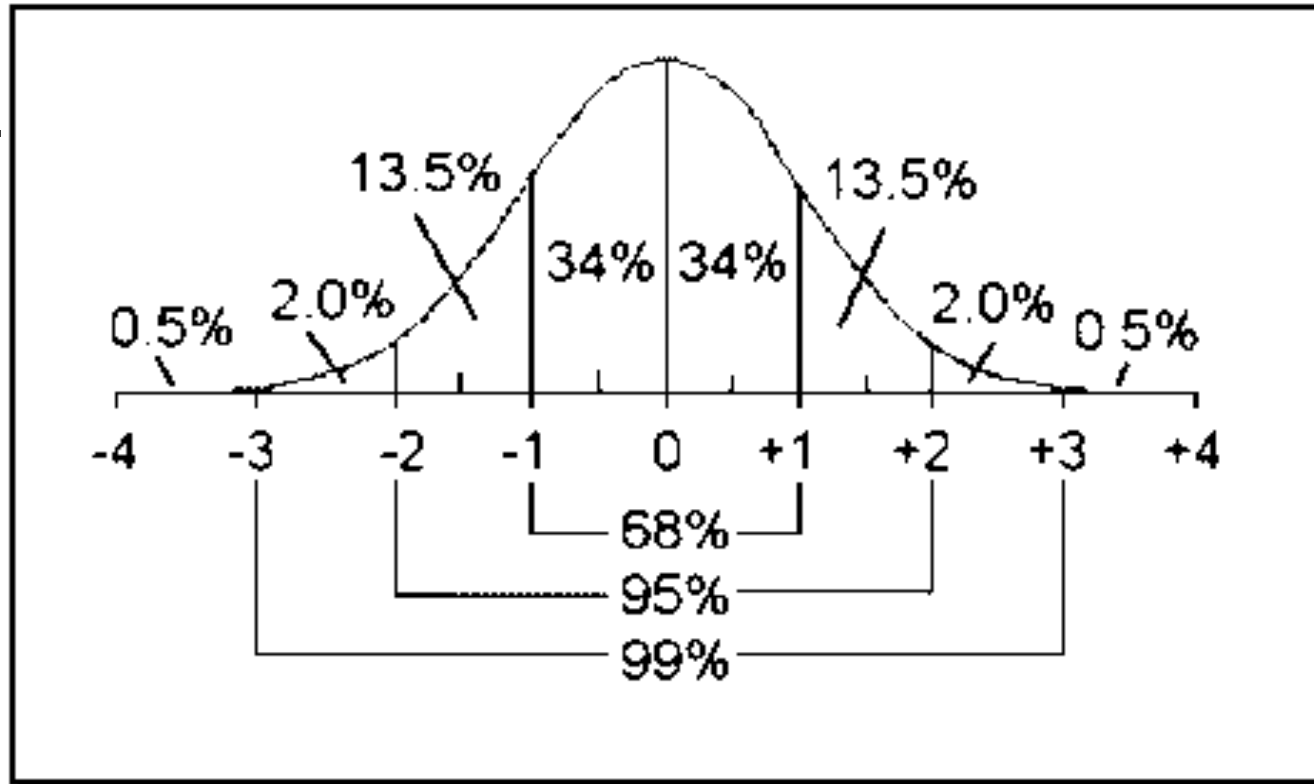
Proportion Example: The Hungry Softball Team Redux

Definition of Proportion

Think of proportion as counting the number of observations that would fit under a certain part of the curve.
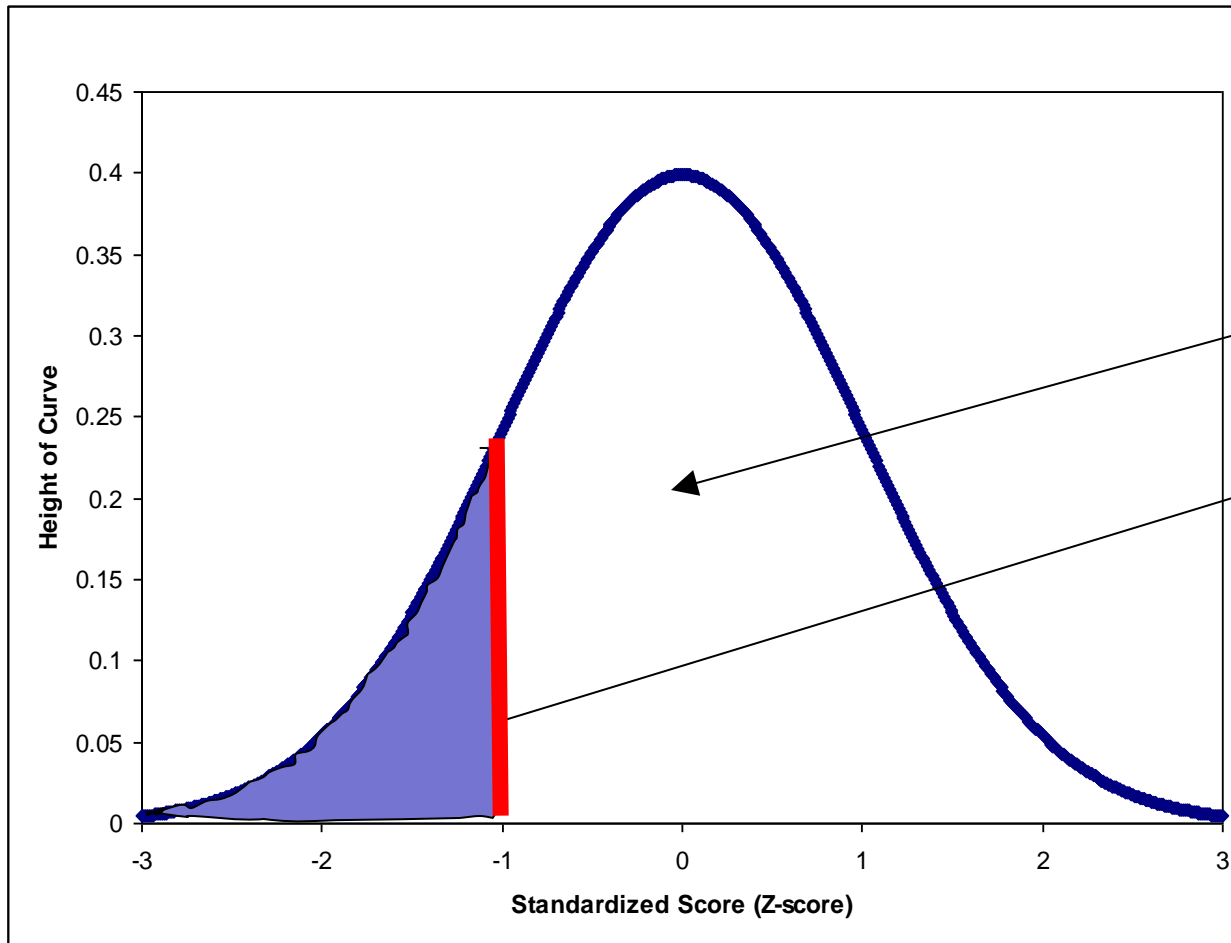

Standard Deviations from the Mean
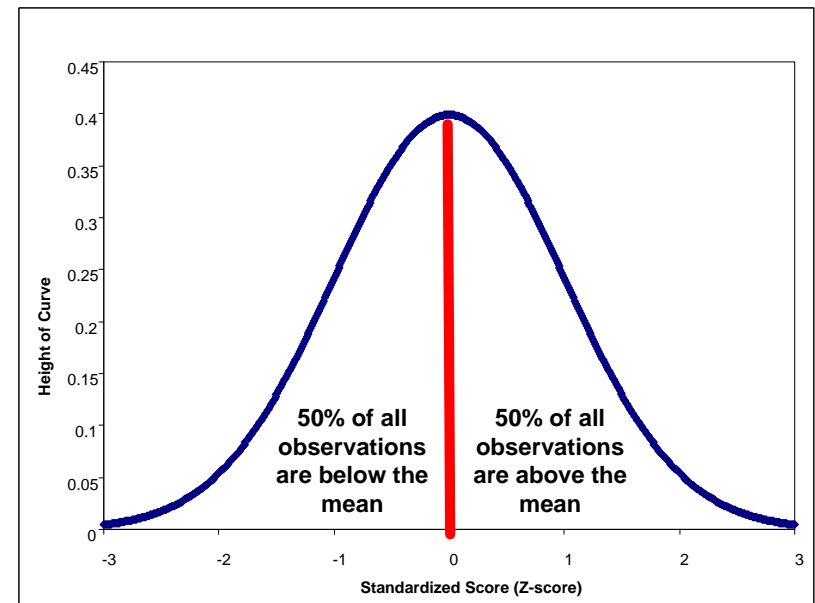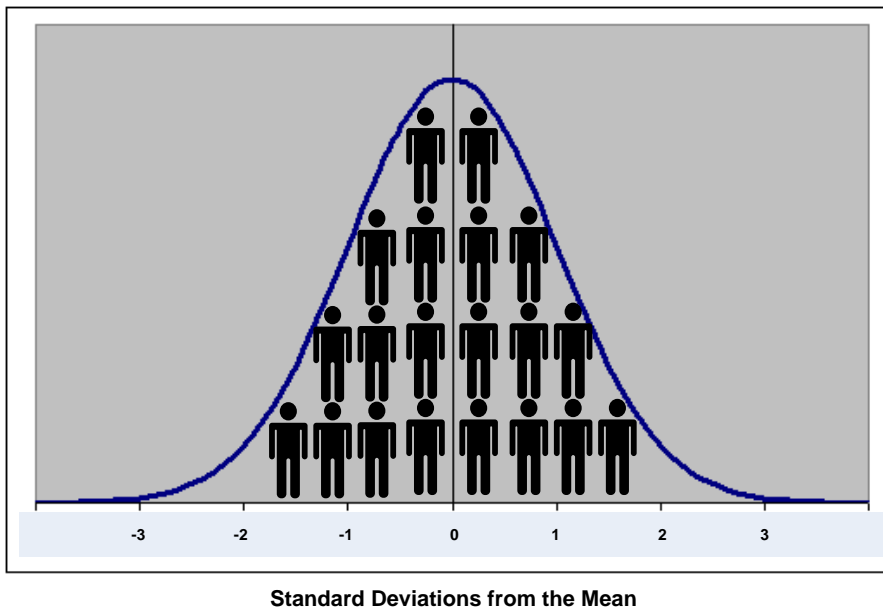
Proportion: Properties of All Normal Distributions

- 68% of observations fall within 1 standard deviation of the mean (34% on either side)
- 95% of observations fall within 2 standard deviations of the mean (47.5% on either side)
- 99% of observations fall within 3 standard deviations of the means (49.5% on either side)
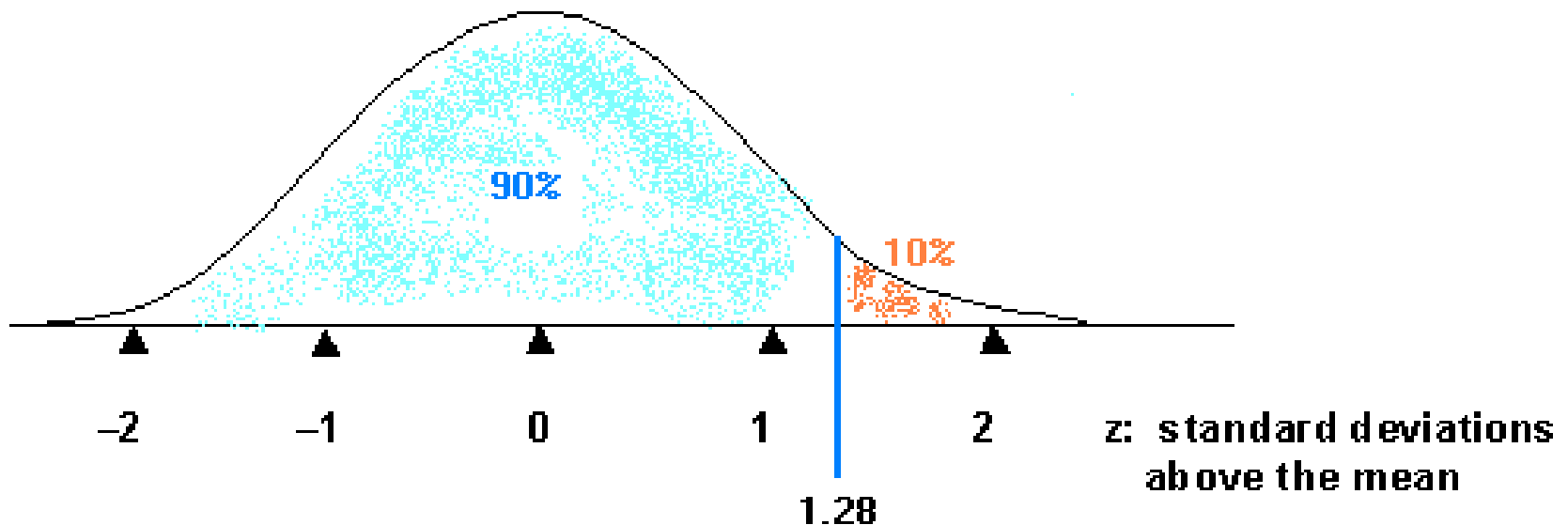
Proportion: Example

84% of observations fall to the right of the Z score of -1.

16% fall to the left of it.

Proportion

- Except at the mean, percentages are not "pretty" numbers (an even multiple of 10) when you have a whole number z-score

- Similarly, the z-score is usually not a "pretty" whole number when you have a "pretty" percentage.

Correlation: What is the relationship between two variables?

Overview

- Up to this point, the discussion has been focused on bell curves. Bell curves only really measure the distribution of one variable within a population.

- Correlation, by contrast, refers to the relationship between TWO variables within a population.

# Correlation: What is the relationship between two variables?

Direction

- Positive correlation: When you see an **increase** in one variable, you also tend to see an **increase** in the other variable.
  - Example: Income and SAT scores. As income rises, so, too, do SAT scores tend to rise for students.
- Negative correlation: When you see an **increase** in one variable, you tend to see a **decrease** in the other variable.
  - Example: alcohol consumption and manual dexterity. As the number of drinks someone has rises, his or her score on a manual dexterity test will tend to fall.
- No relationship: The two variables do not affect each other at all.
  - Example: Ice cream consumption and shark attacks.

Intensity ("r")

- How strong is the relationship between two variables?
- Values of r = 1 or r = -1 are the strongest, while r= 0 is the weakest.

Types of Correlation

no relationship

positive correlation

negative correlation

| r = +1.0 | r = 0.0 | r = -1.0 |

| r ≃ +0.6 | r ≃ -0.6 |
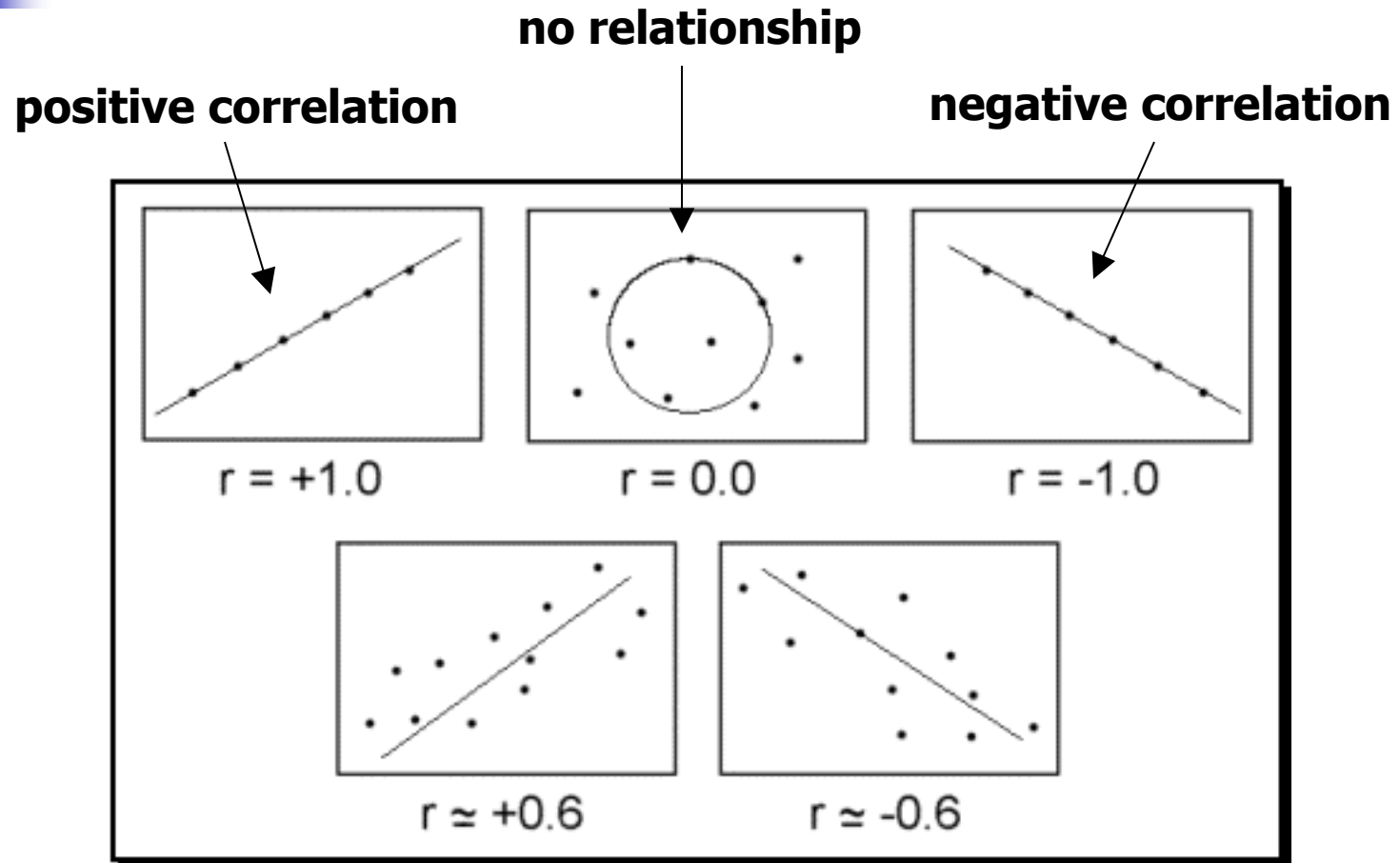
Correlation vs. Causation: Being Cautious with Conclusions

- **One common mistake is made by people interpreting a correlation as meaning that one thing causes another thing. When we see that depression and self-esteem are negatively correlated, we often surmise that depression must therefore cause the decrease in self-esteem. When contemplating this, consider the following correlations that have been found in research:**
  - **Positive correlation between ice cream consumption and drownings**
  - **Positive correlation between ice cream consumption and murder**
  - **Positive correlation between ice cream consumption and boating accidents**
  - **Positive correlation between ice cream consumption and shark attacks**

Correlation

- **If we were to assume that every correlation represents a causal relationship then ice cream would most certainly be banned due to the devastating effects it has on society. Does ice-cream consumption cause people to drown? Does ice cream lead to murder? The truth is that often two variables are related only because of a third variable that is not accounted for within the statistic. In this case, the weather is this third variable because as the weather gets warmer, people tend to consume more ice cream. Warmer weather also results in an increase in swimming and boating and therefore increased drownings, boating accidents, and shark attacks.**

Correlation vs. Causation: Conclusions

- So looking back at the positive correlation between depression and self-esteem, it could be that depression causes self-esteem to go down, or that low self-esteem results in depression, or that a third variable causes the change in both.

- When looking at a correlation, be sure to recognize that the variables may be related but that it in no way implies that the change in one **causes** the change in the other.[2]

[2] Correlation notes taken from from the following web site: http://allpsych.com/researchmethods/correlation.html

Notes
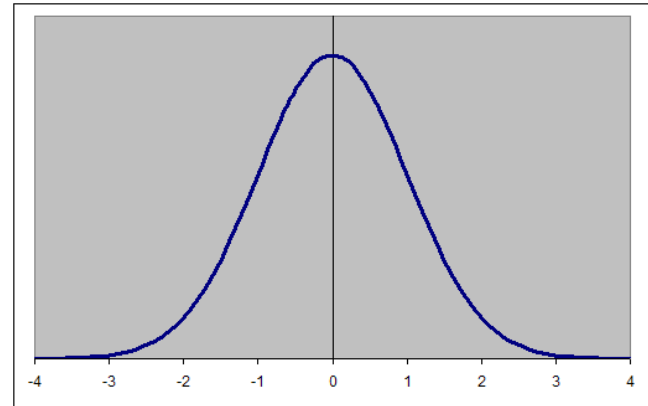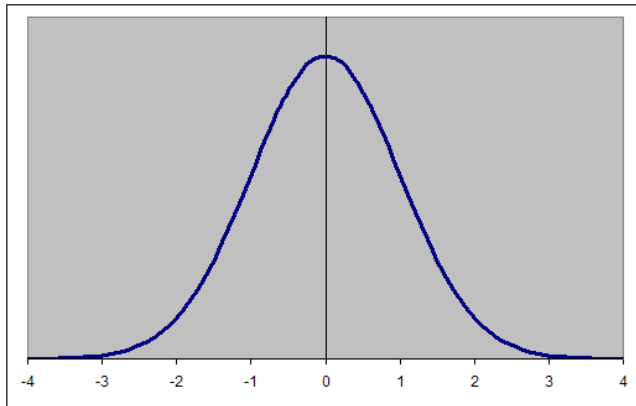
Practice Questions

• Take out your handouts

• Even numbers will be done in class.

• Odd numbers are left for you the students to practice with at home.

• Feel free to get email addresses from your instructor before you leave if you would like confirmation on your homework answers.

Practice Question 1

■1.A doctor wants to test the effectiveness of a new drug on her patients. She separates her sample of patients into two groups and administers the drug to only one of these groups. She then compares the results. Which type of study *best* describes this situation?

■A) census

■B) survey

■C) observation

■D) controlled experiment

Practice Question 2

Decide on a method of data collection you would use for each study. Explain

A) A study on the effect of low dietary intake of vitamin C and iron on lead levels in adults.

2) The age of people living within 500 miles of your home.

Practice Question 3

Frequency Distributions and Statistical Graphs

The following set of data is a sample of scores on a civil service exam:

58, 79, 81, 99, 68, 92, 76, 84, 53, 57, 81, 91, 77, 50, 65, 57, 51, 72, 84, 89

Practice Question 4

(a)   Complete the frequency distribution below for the data.

Classes/Intervals	Frequencies	Cumulative Frequencies
50 – 59		
60 – 69		
70 – 79		
80 – 89		
90 –99		

Practice Question 5

-
- **<u>Inter-Quartile Range</u>**

The test scores for 15 employees enrolled in a CPR training course are listed.
13, 9, 18, 15, 14, 21, 7, 10, 11, 20, 5, 18, 37, 16, 17

- Find the first, second, and third quartiles of the test scores.

Practice Question 6

Two corporations hired 10 graduates. The starting salaries for each are shown in thousands of dollars. Find the deviation for the starting salaries of each corporation.

Corp A Salary	41	38	39	45	47	41	44	41	37	42
Corp B Salary	40	23	41	50	49	32	41	29	52	58

A)Find the inter quartile range for the starting salaries of the two corporations above.

B) Based on your answer to parts (a) & (b), which corporation seems fairer with regards to starting salaries? Explain

Practice Question 7

A study shows that 80% of the selling prices for houses in an area are within two standard deviations of the mean. Is this a normal distribution? Explain.

Practice Question 8

The mean price of houses in Canarsie BK is $482,156, with a standard deviation of $30,000. The data set has a bell shaped distribution. Between what two prices do 95% of the houses fall?
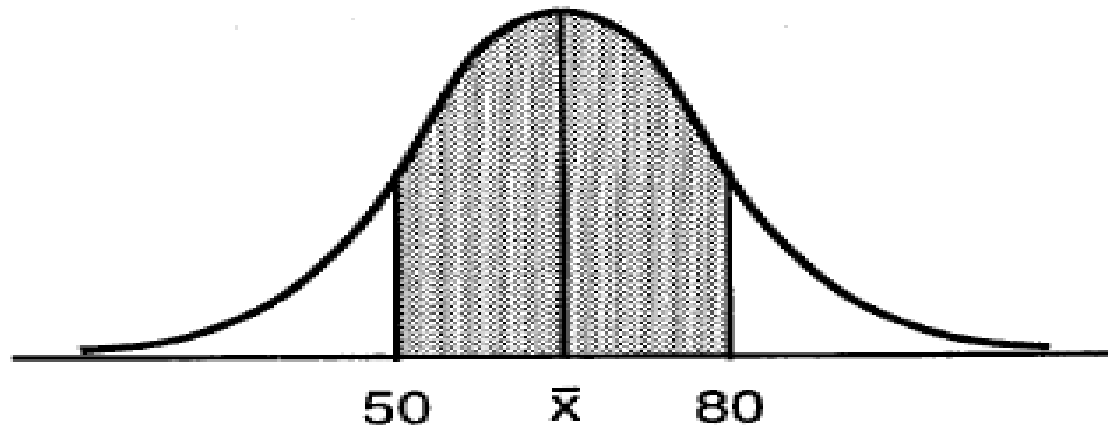
Practice Question 9

On a standardized test, Cathy had a score of 74, which was exactly 1 standard deviation below the mean. If the standard deviation for the test is 6, what is the mean score for the test?

A) 68
B) 71
C) 77
D) 80

Practice Question 10

In the accompanying diagram, about 68% of the scores fall within the shaded area, which is symmetric about the mean, . The distribution is normal and the scores in the shaded area range from 50 to 80.

50   x̄   80

What is the standard deviation of the scores in this distribution?

Practice Question 11

In a normal distribution
$X + 2\sigma = 80 \ and \ X - 2\sigma = 40$ when X
represents the mean and σ represents the
standard deviation. What is the standard
deviation?

Practice Question 12

On a standardized test with normal distribution, the mean is 75 and the standard deviation is 6. If 1200 students took the test, approximately how many students would be expected to score between 69 and 81?

A) 408
B) 600
C) 816
D) 1140

Practice Question 13

Lester, a statistician, measured the mean speed of vehicles on the Belt highway at 7:30am and got 56mph with a standard deviation of 4mph.
Amanda, a highway patrol clocks three cars with speeds of 62mph, 42mph and 56mph at the same time.

(a) Find the z-scores for each speed
(b) Which speed should be issued a ticket? Explain

Practice Question 14

On Kyana's statistics test, the mean score was 79 with a standard dev of 7. On her ELA test the mean was 43 with a standard dev 3. Determine which test Kyana performed better on comparatively with respect to her peers:

Practice Question 15

Jim's score on a national math assessment exceeded the scores of 95,000 of the 125,000 students who took the assessment. What was Jim's percentile rank?

Practice Question 16

In a New York City high school, a survey revealed the mean amount of cola consumed each week was 12 bottles and the standard deviation was 2.8 bottles. Assuming the survey represents a normal distribution, how many bottles of cola per week will approximately 68.2% of the students drink?

A) 6.4 to 12
B) 6.4 to 17.6
C) 9.2 to 14.8
D) 12 to 20.4

Practice Question 17

The number of minutes students took to complete a quiz is summarized in the table below.

Minutes	14	15	16	17	18	19	20
Number of Students	5	3	X	5	2	10	1

If the mean number of minutes was 17, which equation could be used to calculate the value of *x*?

1) $17 = \dfrac{119 + x}{x}$

3) $17 = \dfrac{446 + x}{26 + x}$

2) $17 = \dfrac{119 + 16x}{x}$

4) $17 = \dfrac{446 + 16x}{26 + x}$

Practice Question 18

The air conditioner priced at $480 is discontinued at a local department store. What is the median price of the remaining air conditioners?

$500, $840, $470, $480, $420, $440, $440

Sources/Additional Resources

- Basic explanation of bell curves:
  http://allpsych.com/researchmethods/distributions.html

- Understanding Proportions:
  http://www.utah.edu/stat/bots/game7/Game7.html

- Basic explanation and proportions:
  http://www1.hollins.edu/faculty/clarkjm/Stat140/normalcurves.htm