Organization of Staff Analysts

TRAINING COURSE

FOR THE

ANALYST EXAM

.....

STATISTICS

Welcome to OSA Training 2015
Statistics Part I

By:
Greg Hinckson
Mitch Volk
Dr. Sybil DeVeaux
Iris Bishop

QUALITATIVE DATA

- DEALS WITH DESCRIPTIONS
- DATA CAN BE OBSERVED BUT NOT MEASURED
- COLORS, TEXTURES, SMELLS, TASTES, APPEARANCE, BEAUTY, ETC.

QUANTITATIVE DATA

- DEALS WITH NUMBERS
- DATA WHICH CAN BE MEASURED
- LENGTH, HEIGHT, AREA, VOLUME, WEIGHT, SPEED, TIME, TEMPERATURE, SOUND LEVELS, COSTS, MEMBERS, AGES, ETC.

ORGANIZING AND SUMMARIZING DATA

- **NOMINAL SCALE:** CATEGORIES SUCH AS MALE, FEMALE; DEMOCRAT, REPUBLICAN, INDEPENDENT
- **ORDINAL SCALE:** RANKING MEASURE SUCH AS PRIVATE, CORPORAL, SERGEANT; FIRST, SECOND, THIRD PLACE
- **INTERVAL SCALE:** TRUE NUMERICAL MEASUREMENT SUCH AS DEGREES IN TEMPERATURE, POUNDS FOR WEIGHT, FEET FOR HEIGHT

CENTRAL TENDENCY

- MEAN
- MEDIAN
- MODE

MEAN

- MEAN IS THE AVERAGE
- MEAN = SUM OF THE SCORES
- TOTAL NUMBER OF SCORES

- FIND THE MEAN FOR THESE TEST SCORES:

70, 90, 80, 85, 75, 60, 75, 95, 90, 80, 85

885/11= 80.45

MEDIAN

- MEDIAN IS THE NUMBER IN THE MIDDLE. PUT THE VALUES FROM LOWEST TO HIGHEST. THEN FIND THE NUMBER EXACTLY IN THE MIDDLE.
- WHEN IT IS AN **ODD** NUMBER OF VALUES IT IS IN THE MIDDLE.

  FIND THE MEDIAN: 100, 70, 60, 85, 90

WHEN IT IS AN **EVEN** NUMBER OF VALUES, TAKE THE TWO MIDDLE MOST NUMBERS AND ADD THEM UP AND DIVIDE THE SUM BY 2.

FIND THE MEDIAN: 90, 90, 100, 80, 90, 85

90

MODE

- MODE IS THE VALUE THAT OCCURS THE MOST OFTEN.
- FIND THE MODE FOR THESE WEIGHTS: 110, 140, 130, 160, 120, 180, 140
- FIND THE MODE FOR THESE TEST SCORES: 90, 80, 60, 75, 90, 100, 85

RANGE

- THE DIFFERENCE BETWEEN THE HIGHEST VALUE AND THE LOWEST VALUE.
- FIND THE RANGE FOR TEMPERATURES IN NYC: 68, 55, 72, 49, 53, 64, 58

DREAM AGES

- FIND THE MEAN, MEDIAN, MODE, AND RANGE FOR OUR SET OF DREAM AGES.

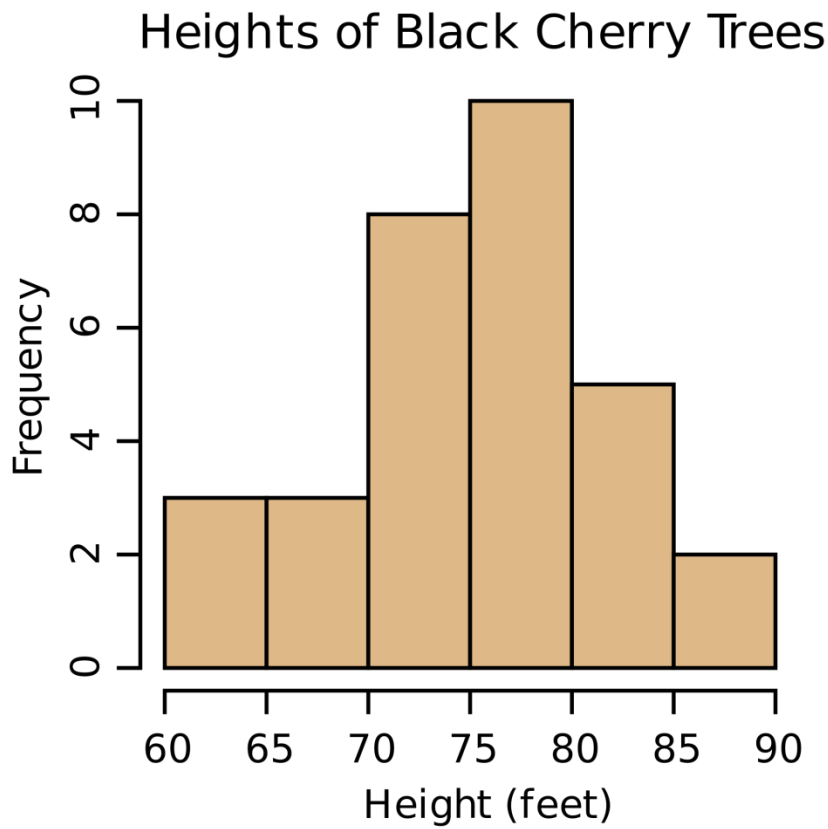
FREQUENCY DISTRIBUTION

- THE NUMBER OF TIMES A GIVEN QUANTITY OCCURS IN A SET OF DATA. THE NUMBER OF METROCARDS SOLD AT THE 23RD STREET AND LEXINGTON AVENUE SUBWAY STATION OVER THE LAST 5 DAYS: 200, 350, 200, 175, 350

METRO CARDS SOLD	**FREQUENCY**
100	0
150	0
175	1
200	2
350	2

Histograms

A Histogram is a bar-type graph without spaces between the bars



- Using the histogram to the left, how many trees are taller than 75 feet?

Example: Histograms

The Fahrenheit temperature readings on 30 April mornings in Stormville, New York, are shown below.

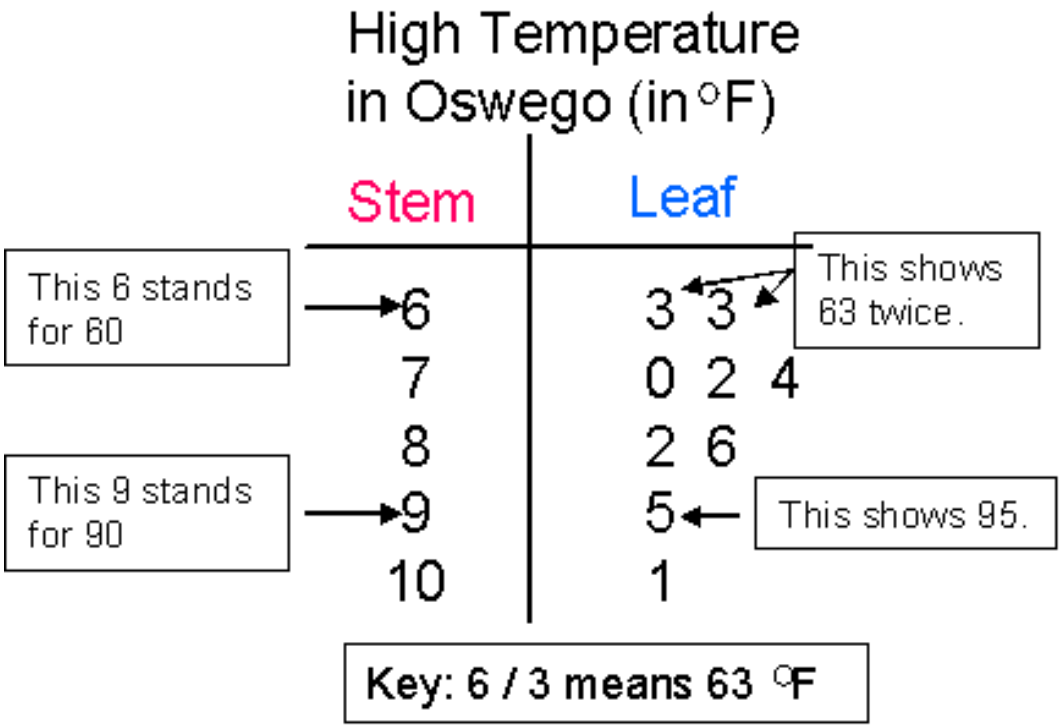
41° , 58° , 61° , 54° , 49° , 46° , 52° , 58° , 67° , 43° ,
47° , 60° , 52° , 58° , 48° , 44° , 59° , 66° , 62° , 55° ,
44° , 49° , 62° , 61° , 59° , 54° , 57° , 58° , 63° , 60°

- Using the data, complete the frequency table below.

Create a frequency Histogram

Interval	Tally	Frequency
40–44		
45–49		
50–54		
55–59		
60–64		
65–69		

Stem and Leaf Plots



Sample STANDARD DEVIATION & VARIANCE

Sample Variance

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

Sample Standard Deviation

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Example

- I need ten volunteers state their ideal weights.
- We are now going to find the standard Deviation of your ideal weights of the ten of you.

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

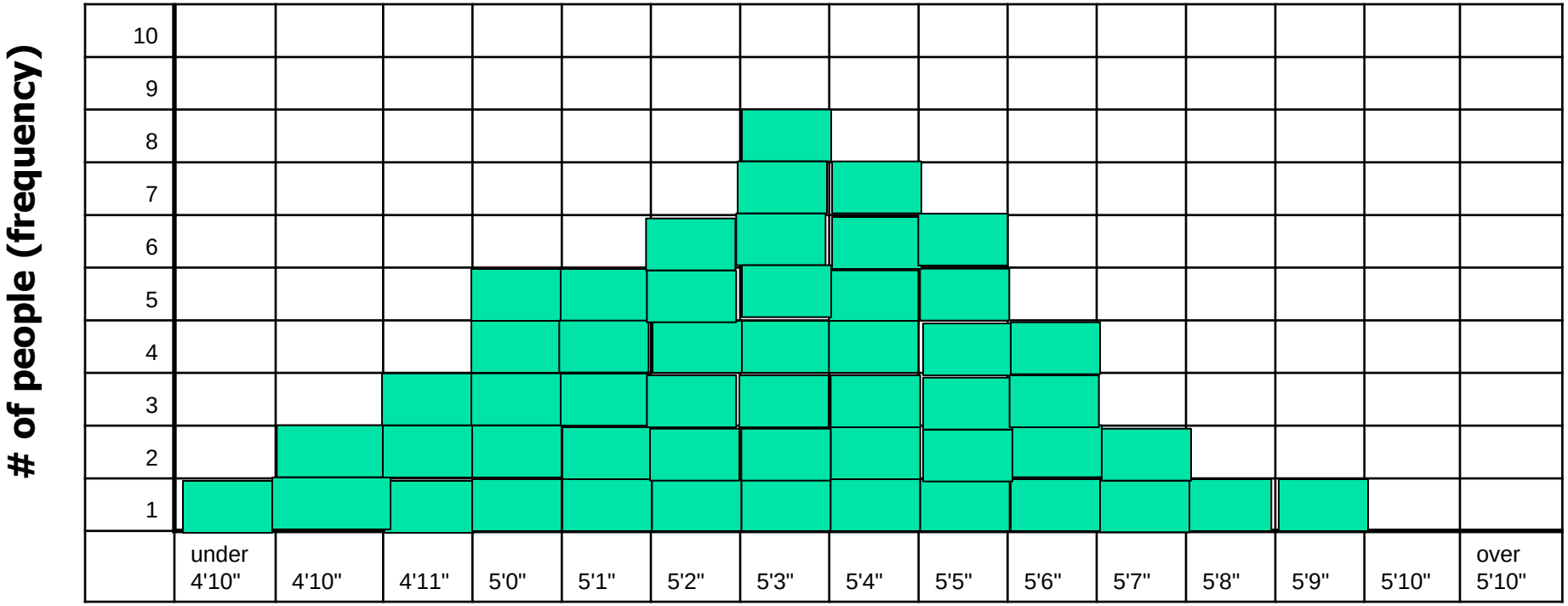
- A **statistic** is a way to represent or organize information in a way that helps you understand it better than simply looking at a series of numbers.
- You can use a set of data to draw a picture that will help you to understand and interpret that data.

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



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)

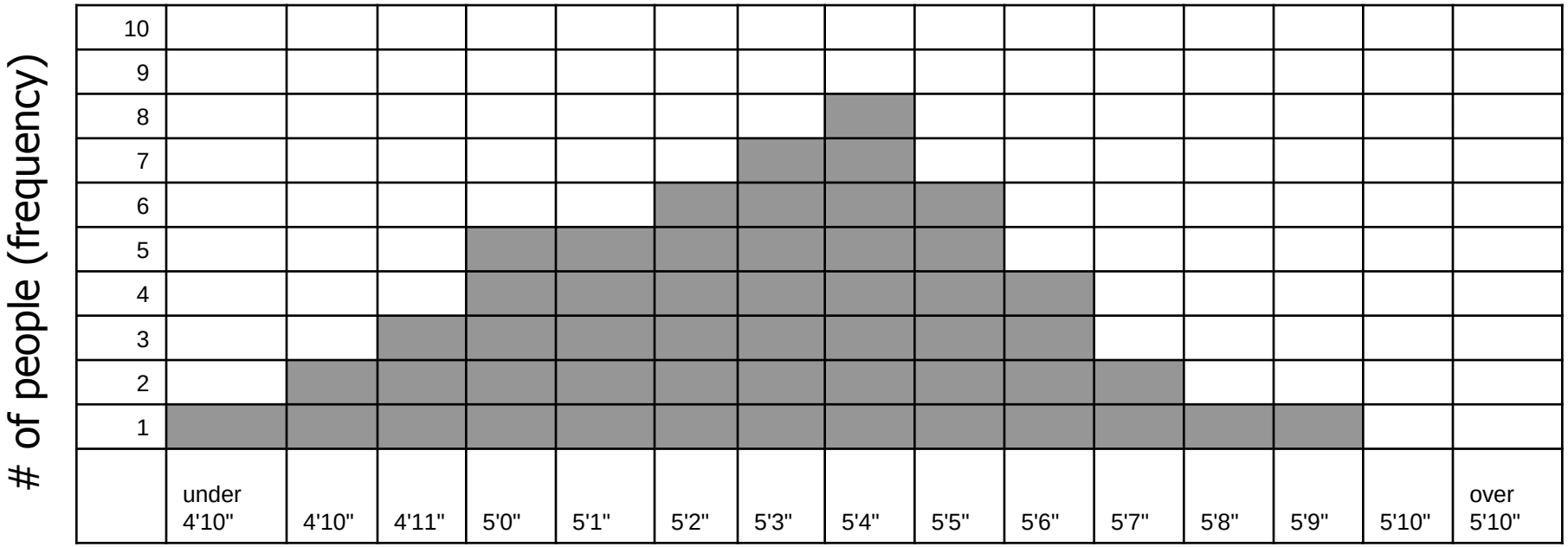
10														
9														
8														
7														
6														
5														
4														
3														
2														
1														
	under 5'3"	5'3"	5'4"	5'5"	5'6"	5'7"	5'8"	5'9"	5'10"	5'11"	6'0"	6'1"	6'2"	6'3"	over 6'3"

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



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.

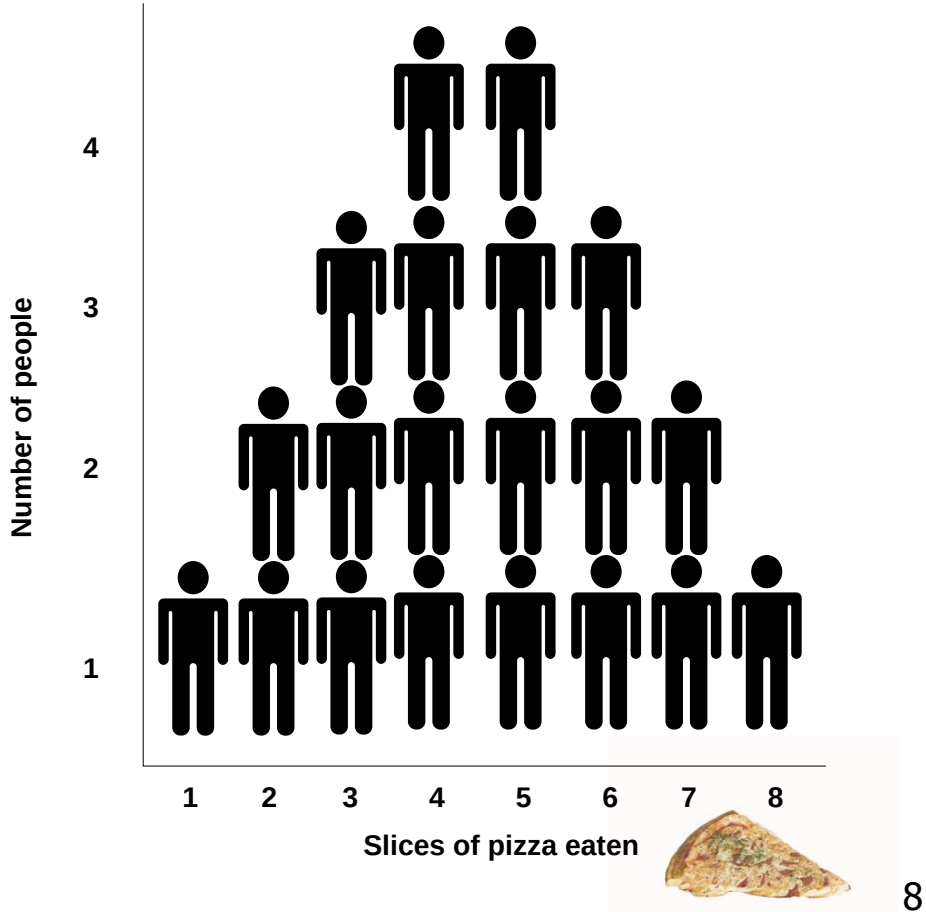
Example: The Hungry Softball Team

Situation

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

Graph

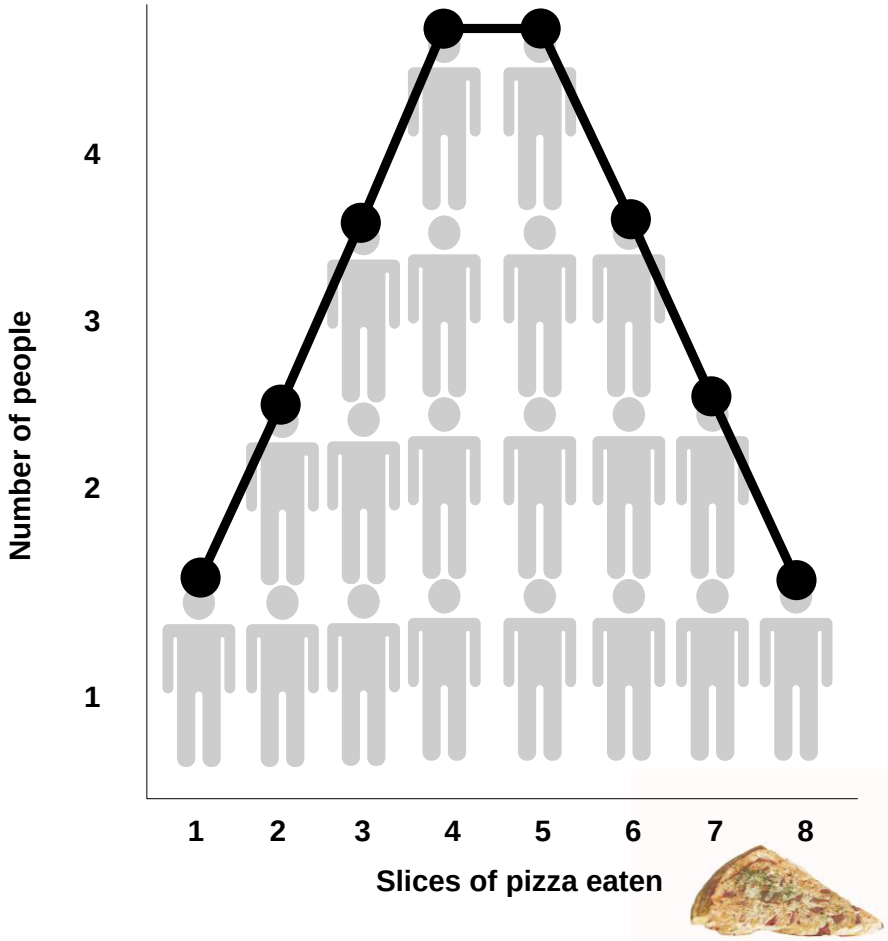
A simple graph shows how many slices each of the 20 team members ate. For example, four people ate 5 slices of pizza, while only one person ate 8 slices.



Example: The Hungry Softball Team

Graph

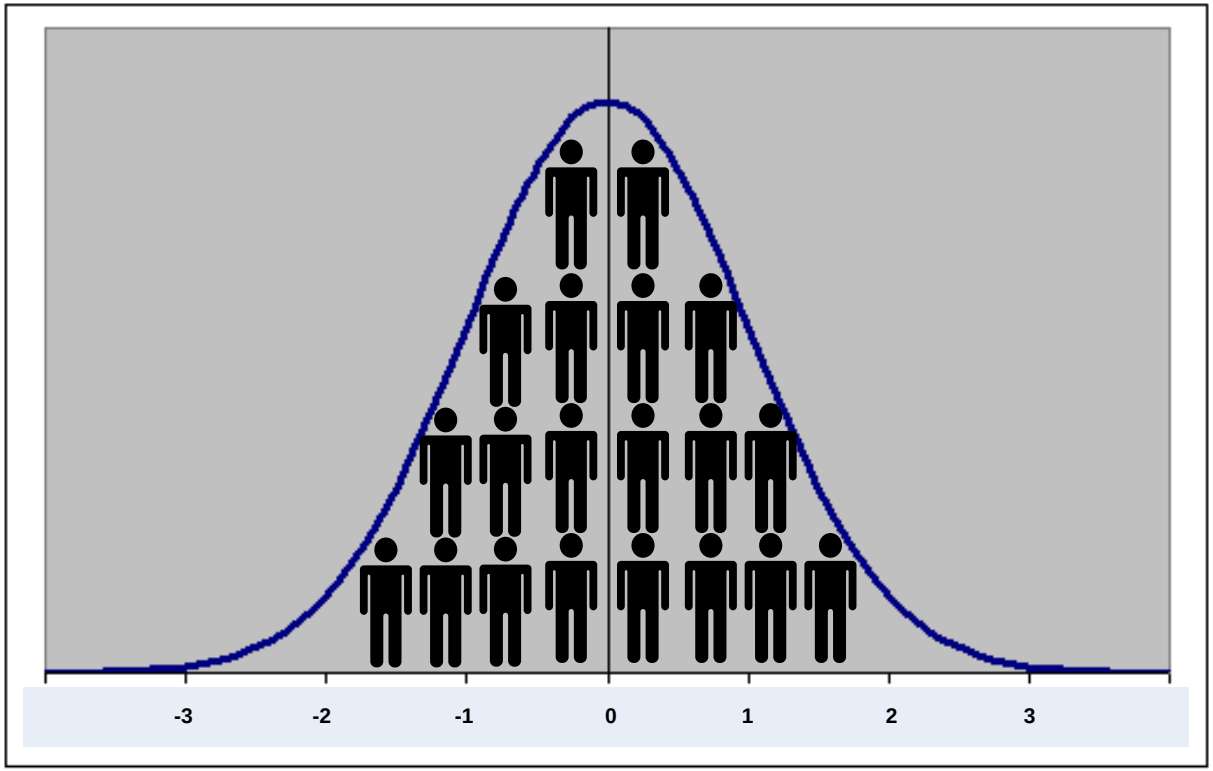
A line shows how you could draw a simple graph using the tops of the heads of each group of players.



Example:
The Hungry Softball Team

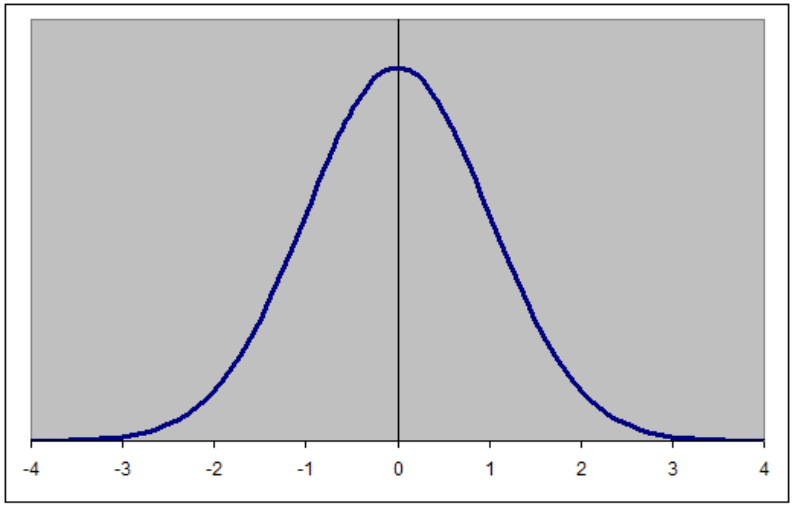
Graph

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



Bell Curves: What are they?

The Bell Curve

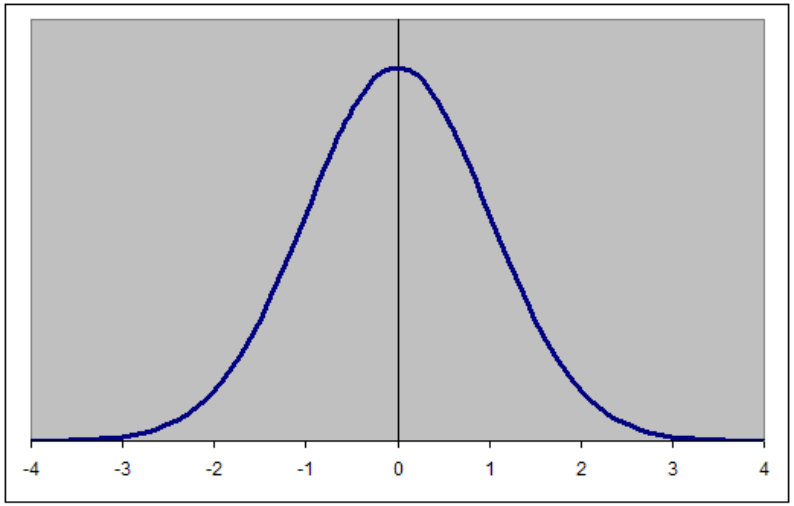


Basic Properties

- 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.

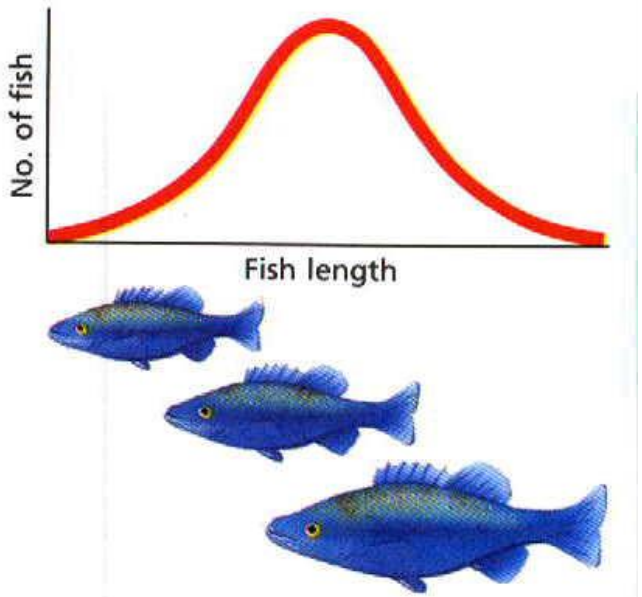
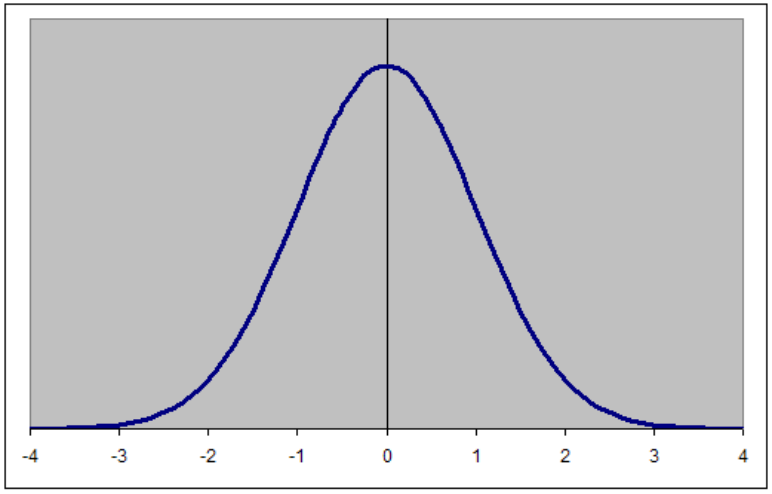


Figure 16-2 A bell curve illustrates how most members of a population are grouped in an average range for a given trait while only a few are at the extreme ends of the range.

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
$\bar{x} = \frac{\Sigma x}{n}$	$\mu = \frac{\Sigma x}{N}$

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**

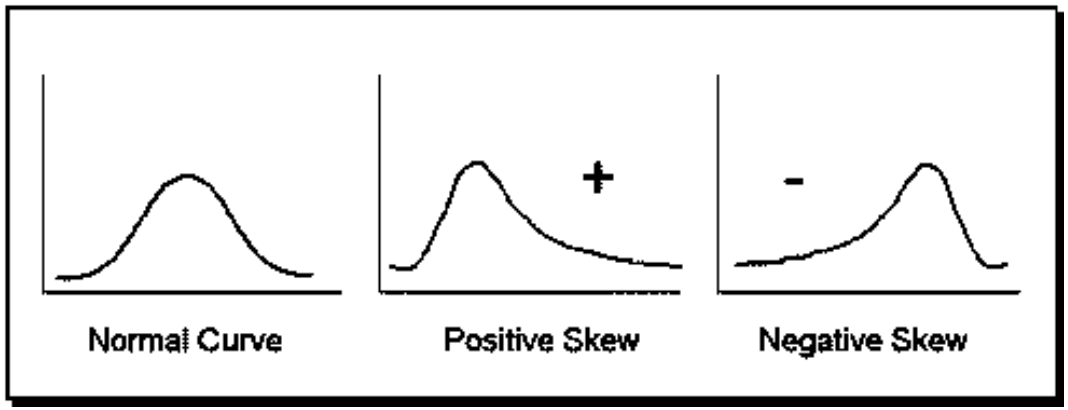
median

- **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]

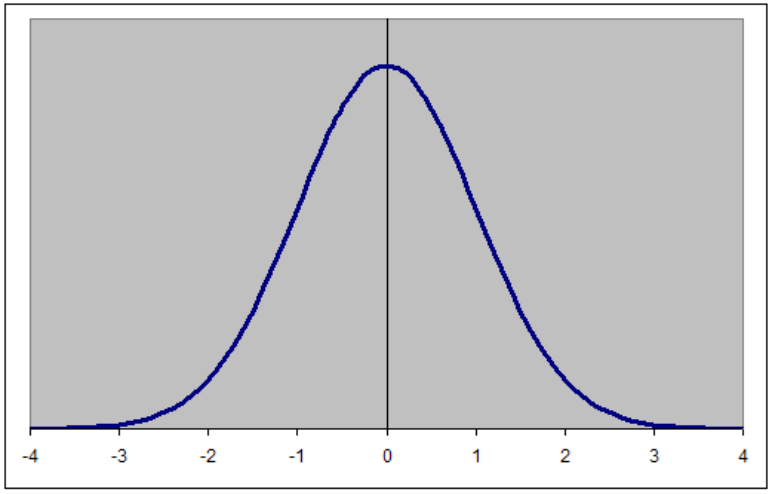


[1] Text from: http://allpsych.com/researchmethods/distributions.html.

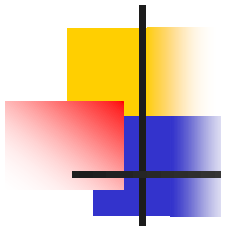
Bell Curves: Variance

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

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

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

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

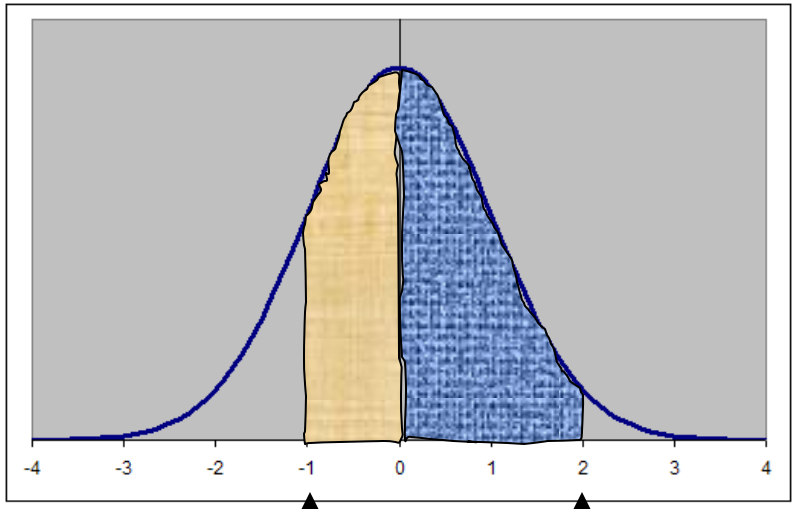
$$\sigma^2 = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = 0.667$$

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

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.



Standard Deviation cont'd…

- The square root the variance is the standard deviation.

Standard Deviation Formula

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

σ = lower case sigma

$\sum$ = capital sigma

$\bar{x}$ = 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**

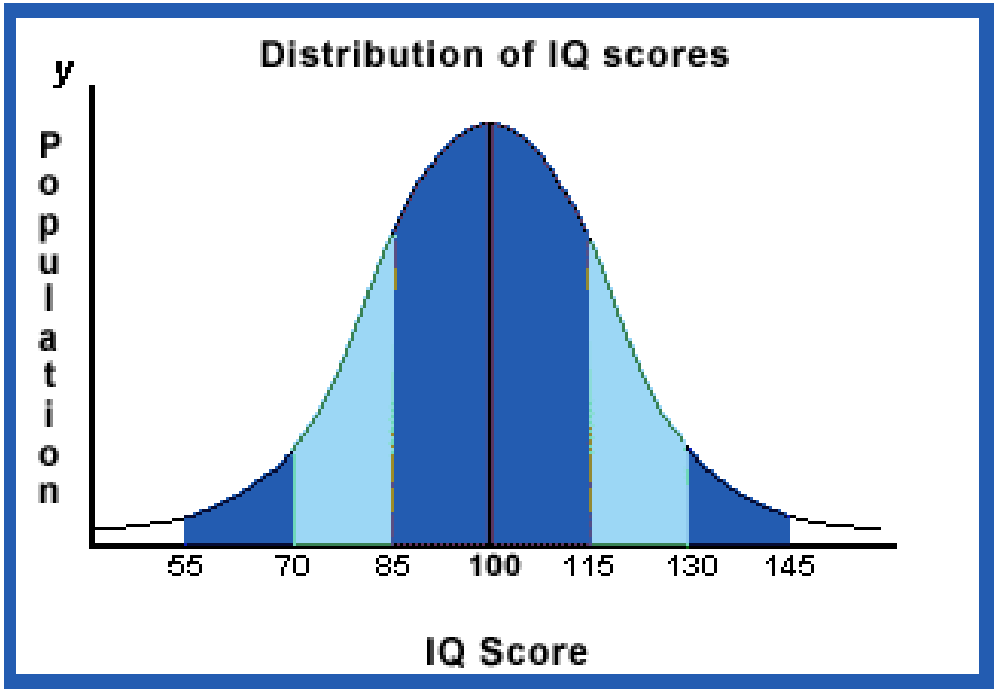
Bell Curves: What are they?

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.

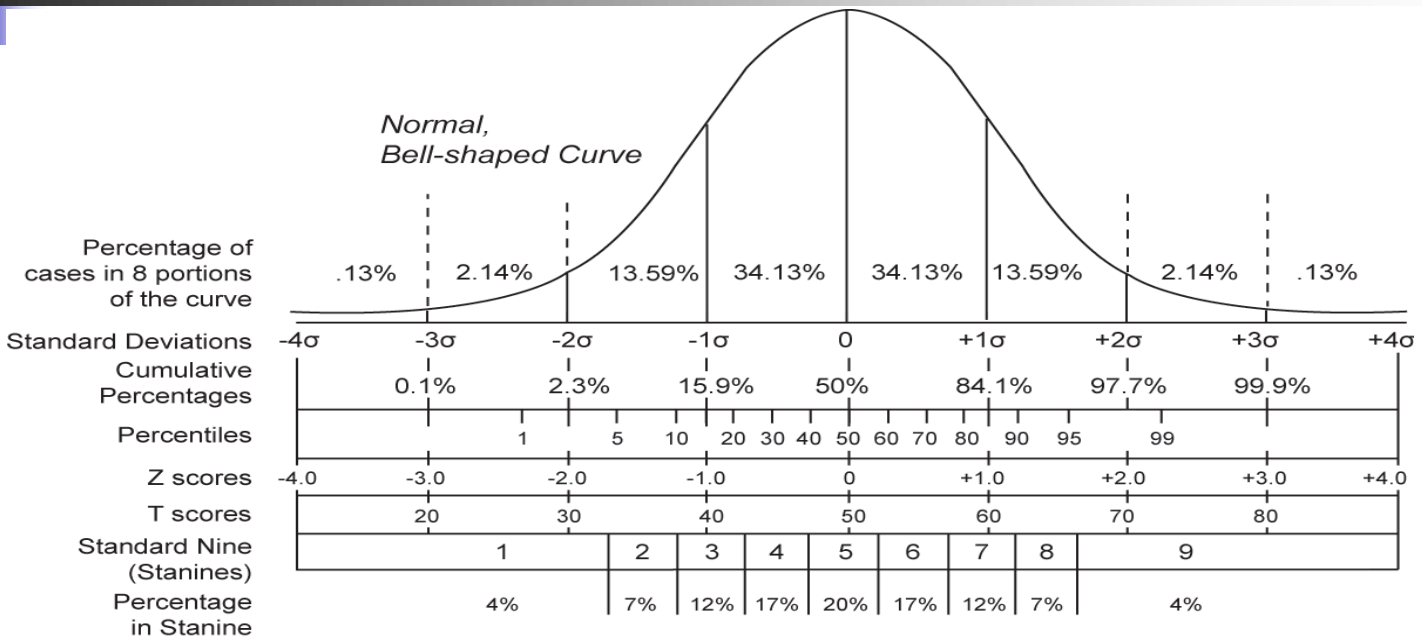


Z-Scores: What are they?

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).

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.

$$\textbf{Z-score} = \frac{\textbf{Observation Value} - \textbf{Mean}}{\textbf{Standard Deviation}}$$

Z-Scores

- 130 - 100/15 = 2

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

μ = Mean

σ = Standard Deviation

Z-Scores

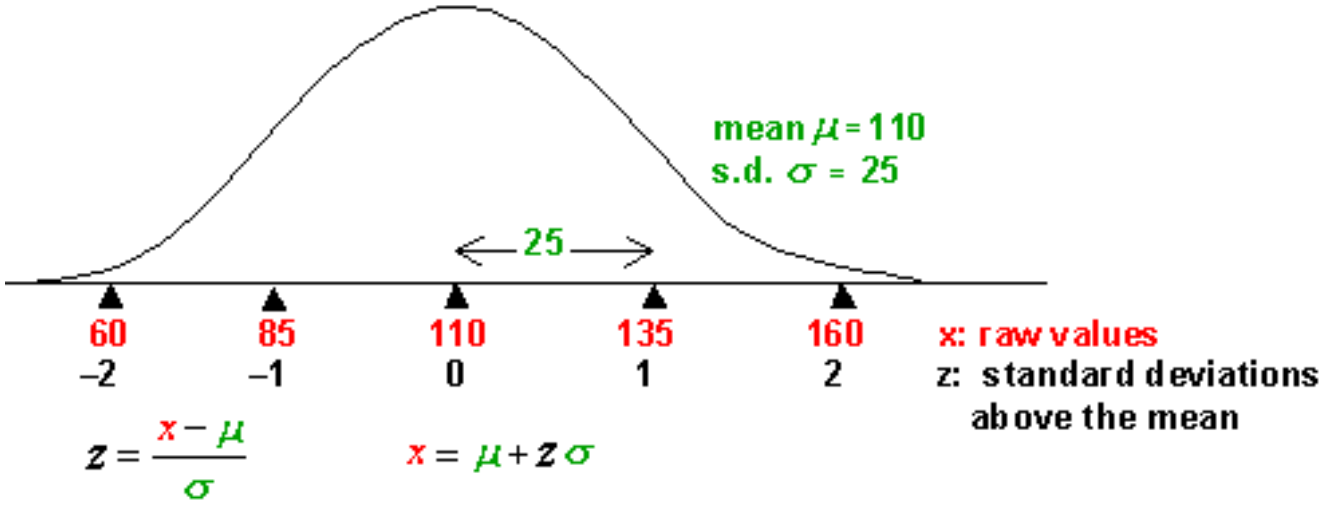
- With an IQ score of 80 and a mean of 100, with a standard deviation of 15 what is the z-score?

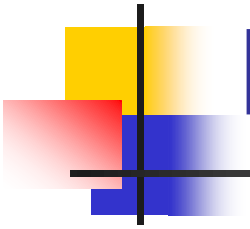
Z-Scores

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

"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.





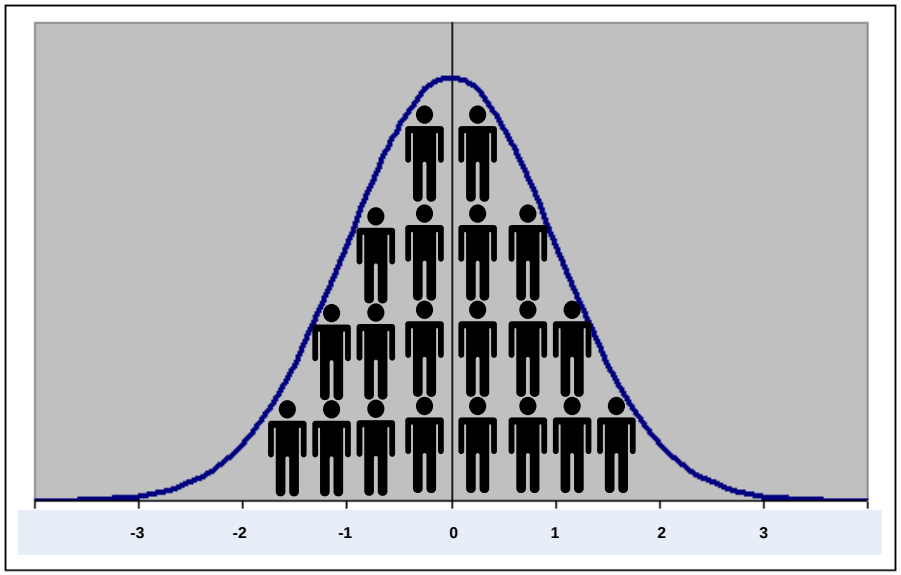
Proportion

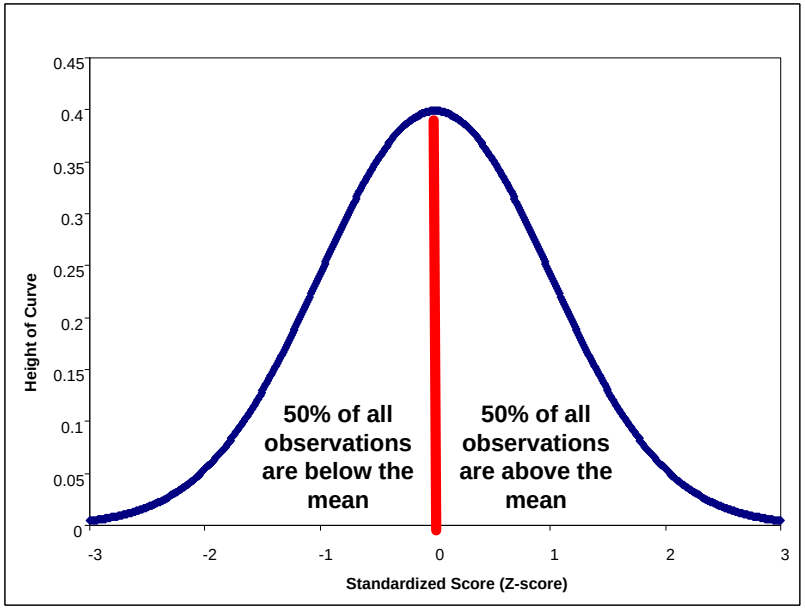
- The area under a bell curve tells you what percentage of ALL observations fall within that area.
- The total area under a bell curve is always equal to 1, or 100%.

Proportion Example: The Hungry Softball Team Redux

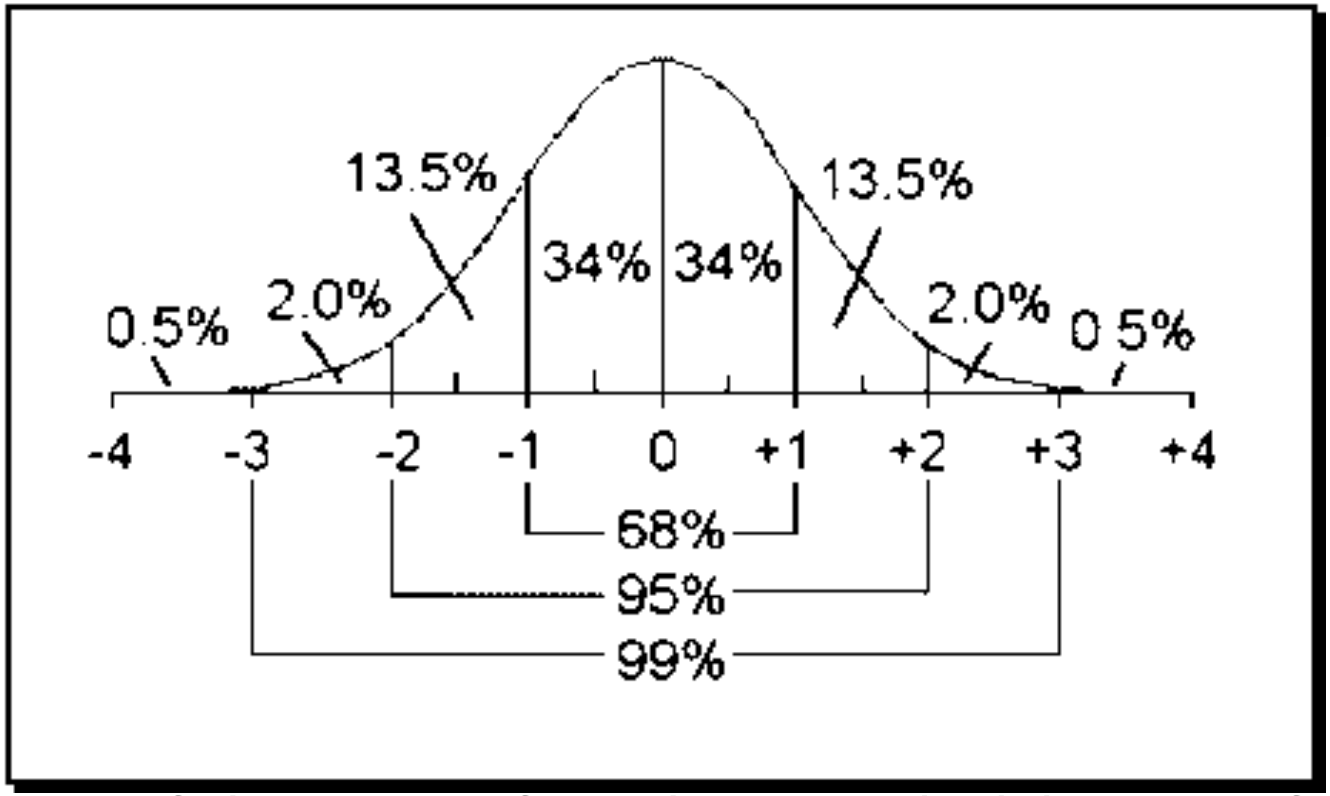
Definition of Proportion

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



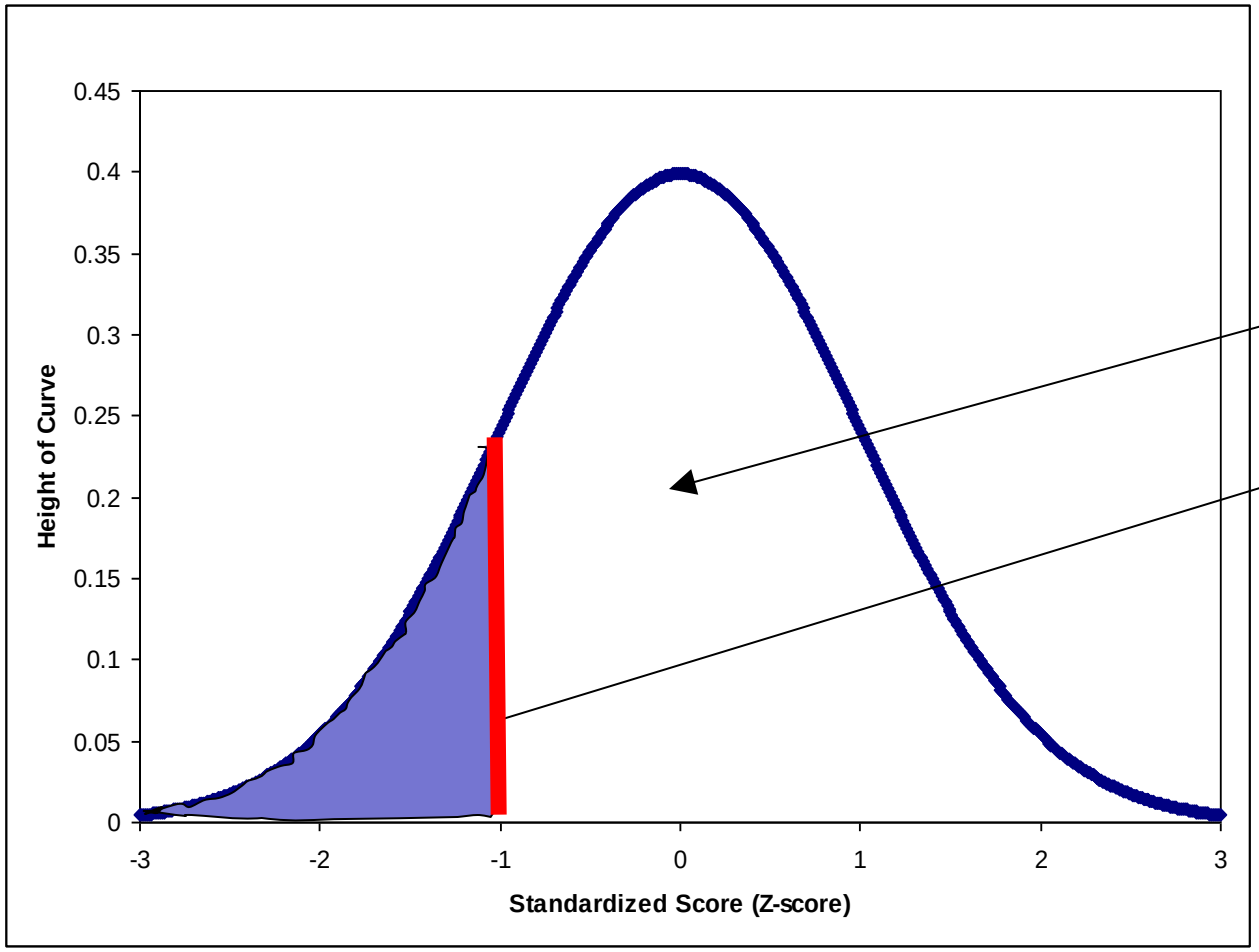


Proportion: Properties of All Normal Distributions



- 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

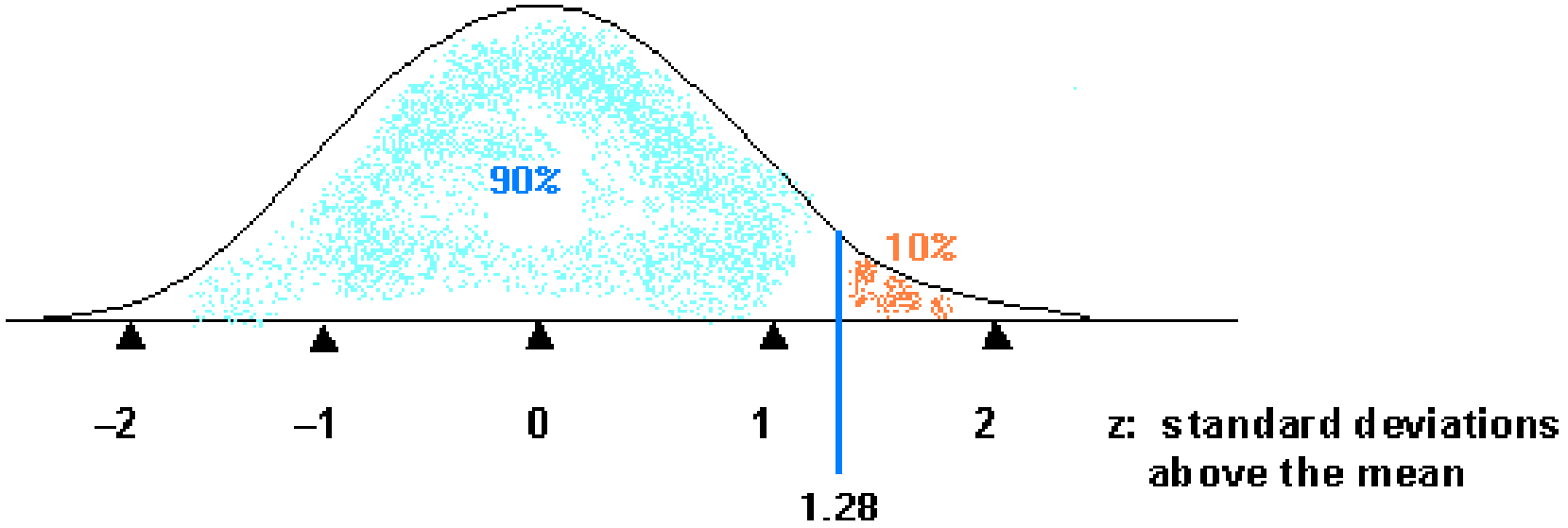


84% of observations fall to the right of the Z score of -1.

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?

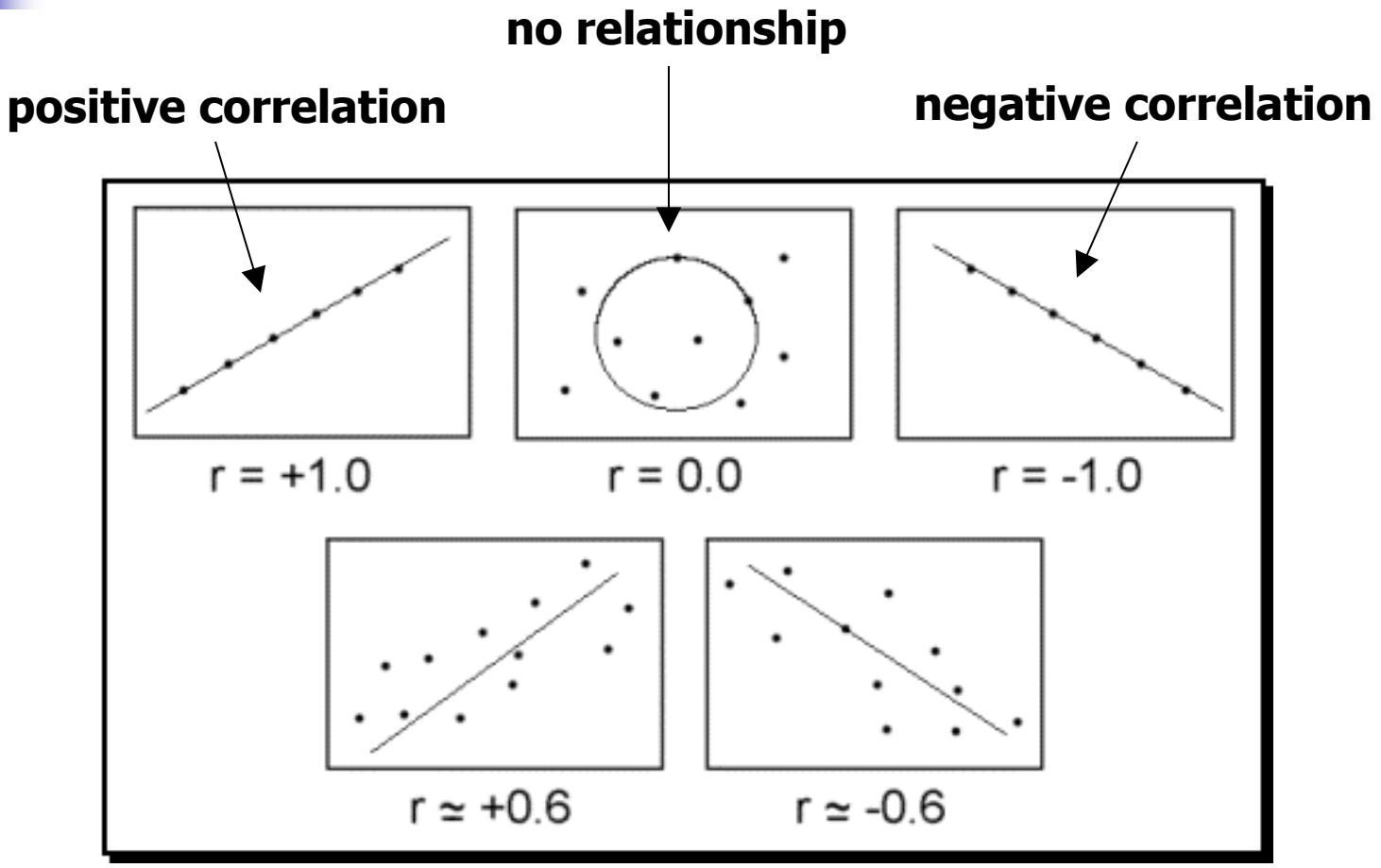
Direction

- 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



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

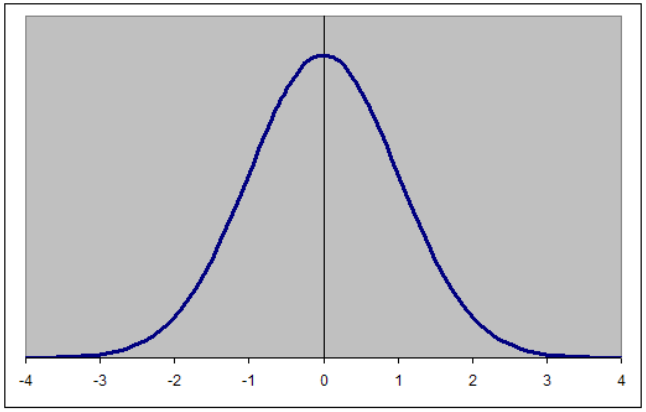
- **If we were to assume that every correlation represents a causal relationship then ice cream would most certainly be banned due to the devastating effects it has on society. Does ice-cream consumption cause people to drown? Does ice cream lead to murder? The truth is that often two variables are related 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.**

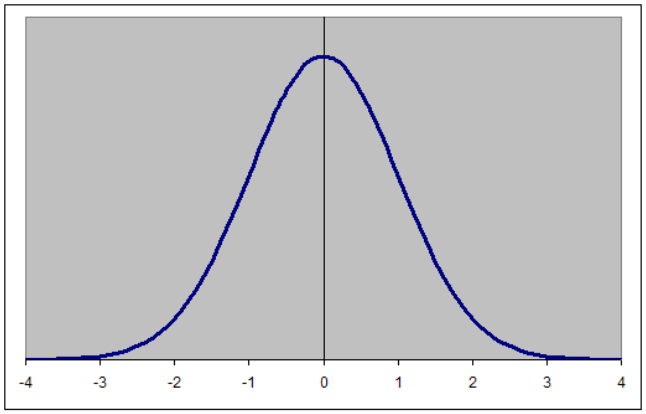
Correlation vs. Causation: Conclusions

- 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

- **Inter-Quartile Range**

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?

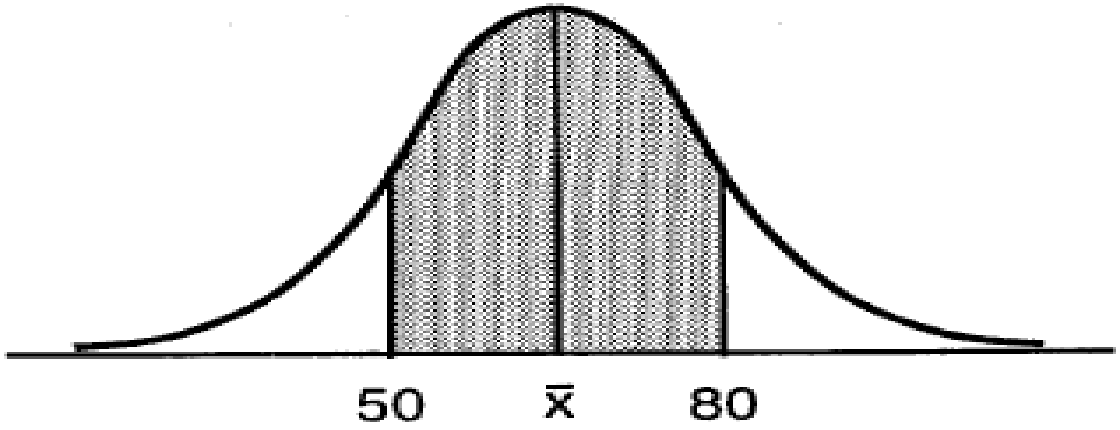
Practice Question 9

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.



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) $17 = \frac{119 + x}{x}$

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

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

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

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

OSA Training 2015

Statistics Part 1&2 **Practice Questions**

The Basics of Statistics

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? **(4)**

1) census

2) survey

3) observation

4) controlled experiment

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.

Controlled Experiment

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

Survey

Frequency Distributions and Statistical Graphs

3. 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

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

Classes/Intervals	Frequencies	Cumulative Frequencies
50 – 59	**6**	**6**
60 – 69	**2**	**8**
70 – 79	**4**	**12**
80 – 89	**5**	**17**
90 –99	**3**	**20**

(b) Construct a frequency histogram for the data set.

(c) Construct a cumulative frequency histogram for the data set

(d) Compare and contrast the two histograms in pats (b) and (c).

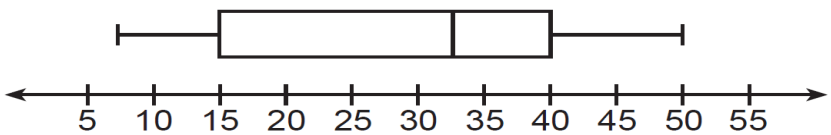
The bars in the regular frequency histogram fluctuates (go up and down), The bars in the cumulative frequency histogram however, consistently rise.

(e) Construct a stem-and-leaf plot for the data set.

Stem	Leaf
5	0 1 3 7 7 8
6	5 8
7	2 6 7 9
8	1 1 4 4 9
9	1 2 9

4. The box-and-whisker plot below represents the ages of 12 people. What percentage of these people is age 15 or older?

75%



Inter-Quartile Range

5. 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

(a) Find the first, second, and third quartiles of the test scores.

$Q_1 = 10$ $Q_2 = 15$ $Q_3 = 18$

(b) Create a box-and-whisker plot for the scores

(c) Find the *IQR* for the scores

IQR = 8

Standard Deviation and The Normal Distribution

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**

6.(a) 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 = 2.974

Corp B = 10.5

(b) Find the inter quartile range *(IQR),* for the starting salaries of the two corporations above.

Corp A *IQR* = 5

Corp B *IQR* = 18

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

Corp A because the starting salaries are more clustered than those of Corp B.

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

No. If the data set was normally distributed, then about 95% of the data would be within two standard deviations of the mean.

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?

Between $422,156 & $542,156

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? **(4)**

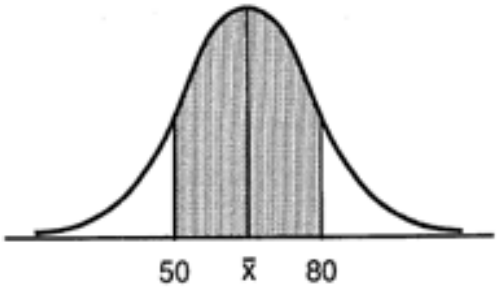
1) 68

2) 71

3) 77

4) 80

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



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

15

11. In a normal distribution, $\bar{x}+2\sigma=80$ and $\bar{x}-2\sigma=40$ when $\bar{x}$ represents the mean and σ represents the standard deviation. What is the standard deviation?

10

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? **(3)**

1) 408

2) 600

3) 816

4) 1140

Comparing z-scores

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

62mph = 1.5 **42mph = -3.5** **56mph = 0**

(b) Which speed should be issued a ticket? Explain

42mph, because that speed poses a greater threat to the flow of traffic.

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:

STATS: 90 ELA: 50

z = 1.57 **z = 2.33**

94% **99%**

Kyana did better on the ELA test because her percentile rank is higher

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?

76%

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? **(3)**

1) 6.4 to 12

2) 6.4 to 17.6

3) 9.2 to 14.8

4) 12 to 20.4

Measures of Central Tendency

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? **(4)**

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

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

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

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

18. The air conditioner priced at $480 is discontinued at a local department store. What is the median price of the remaining air conditioners? **(455)**

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

STATISTICS QUESTIONS

1. The term central tendency refers to

   a) a central statistic
   b) clusters towards the middle
   c) a tendency in a central area
   d) the middle of a dispersion

2. Another name for the bell curve is

   a) normal curve
   b) Z score
   c) standard deviation
   d) looped curve

3. A mean is

   a) a sample
   b) the most frequent number
   c) the average of all values
   d) histogram

4. When a curve has extreme scores on the right hand side of the distribution, it is said to be

   a) an intensity
   b) negatively skewed
   c) positively skewed
   d) a correlation

5. What does variability measure

   a) the amount of difference among observations in a distribution
   b) the amount of distance in a bell curve
   c) the amount of similarity between two observations
   d) the amount of difference between Z scores

6. What information does the standard deviation give you?

a) it measures the difference between two means
b) the difference from the mean
c) the difference between a raw score and a Z score
d) variability of observations

7. What is a Z score?

a) a measure of the variability described by a bell curve
b) the square root of the variance
c) a way to covert real data in the world into a form that fits on a bell curve
d) the average of all the values

8. What is a proportion?

a) the relationship between two variables
b) the comparison between the mean and the mode
c) the difference between the Z score and the mode
d) the area under a bell curve which tells you what percentage of ALL observations fall within that area

9. What is the relationship between two variables within a population?

a) reliability
b) proportion
c) correlation
d) standard deviation

10. Which statement is NOT true?

a) a correlation in no way implies that the change in one causes the change in the other
b) the skew of a distribution refers to how the curve leans
c) 68% of observations fall within one standard deviation
d) none of the above

1. b 2. a 3. c 4. c 5. a 6. b 7. c 8. d 9. c 10. d

STATISTICS QUESTIONS

1. Of the following, the formula which is used to calculate the arithmetic mean from data grouped in a frequency distribution is

   a) $M = \frac{N}{fx}$

   b) $M = N\ (fx)$

   c) $M = \frac{fx}{N}$

   d) $M = \frac{x}{fN}$

2. Arranging large groups of numbers in frequency distributions

   a) gives a more composite picture of the total group than a random listing
   b) is misleading in most cases
   c) is unnecessary in most instances
   d) presents the data in a form whereby further manipulation of the group is eliminated

3. The value of statistical records in MAINLY dependent upon the

   a) method of presenting the material
   b) number of items used
   c) range of cases sampled
   d) reliability of the information used

4. A set of ordered scores and their corresponding frequencies is called a

   a) frequency distribution
   b) a central tendency
   c) mean
   d) none of the above

5. The mean global temperature for the years that encompass 1970 through 1980 is _____.

6. The mean global temperature for the years that encompass 1980 through 1990 is _____.

7. The mean global temperature for the years that encompass 1990 through 2000 is _____.

8. The median of the below listed numbers is _____.

   24, 38, 39, 40, 42, 50, 55, 70, 80, 90

9. The median of the below listed numbers is _____.

   24, 38, 39, 40, 41, 42, 50, 55, 70, 71, 94

10. The median of the below listed numbers is _____.

    21, 23, 43, 45, 49, 53, 57, 61, 67, 73, 79, 85, 91

11. The mode of the below listed numbers is _____.

2,2,2,2,2,5,5,5,5,5,5,5,5,7,7,7,7,7,7,7,9,9,9,9,9

12. The term bimodal refers to

a) a frequency distribution which has two peaks
b) a histogram which has two peaks
c) a graph which has two separate concentrations
d) all of the above

13. In the formula for the mean, N is referred to as

a) the total number of scores
b) normal curve
c) the sum of the scores
d) the central tendency

14. An essay test was administered to 100 applicants and was then rated in multiples of 5. The results were as follows:

score	number of applicants
90	10
85	10
80	25
75	30
70	15
65	10

The mean is

a) 75 b) 77 c) 80 d) 82

15. One could chart a frequency distribution by either using a

a) central tendency or a histogram
b) normal curve or a central tendency
c) line graph or a histogram
d) line graph or a normal curve

16. The point on the scale at which the concentration is greatest or that value which occurs the greatest number of times and which might be taken as typical of the entire distribution is called

a) mean
b) median
c) mode
d) bell curve

1. c 2. a 3. d 4. a
8. 46 9. 42 10. 57 11. 5
12. d 13. a 14. b 15. c 16. c

STATISTICAL METHODOLOGIES

FREQUENCY DISTRIBUTION (1)

1. Such things as test scores, class rank, weight, and income, are called variables. Income, for instance, is called a variable because different income values are possible. In general, things that vary in value from case to case or time to time are called _____.

VARIABLES

2. The number of times a particular value of a variable occurs is referred to as the *frequency* of that value. If 17 students receive a score of 70 on a test, then the score of 70 has a _____ _____ of 17.

FREQUENCY

3. A distribution is a series of separate values such as scores which are ordered to magnitude. A group of ordered scores can make up a *distribution*. For example, a group of scores ranging from the lowest to the highest score is a _____ (see table).

Scores
13
11
11
9
9
9
8
5

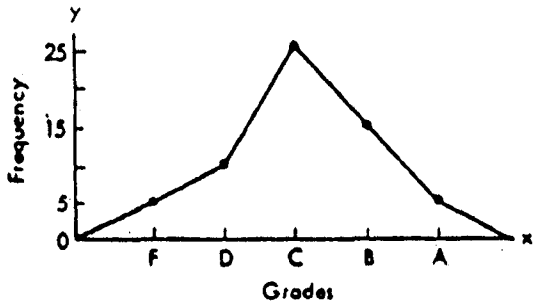
DISTRIBUTION

4. A set of ordered scores and their corresponding frequencies is called a *frequency distribution*. This can be represented in table or graph form. The table below shows the number of times a score occurs in its group. This table is a frequency _____.

Scores	Frequency
13	I
11	II
9	III
8	I
5	I

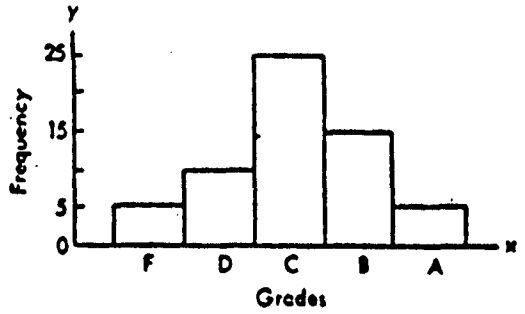
DISTRIBUTION

5. Frequency distributions can also be graphically illustrated. The two most common graphs used to illustrate frequency distributions are the *frequency polygon* and the *histogram*. If scores and their frequencies are illustrated with points connected by lines, it is called a *frequency polygon*. Because the illustration below shows the frequency of particular scores by the height of points that are connected by lines, it is called a frequency _____.



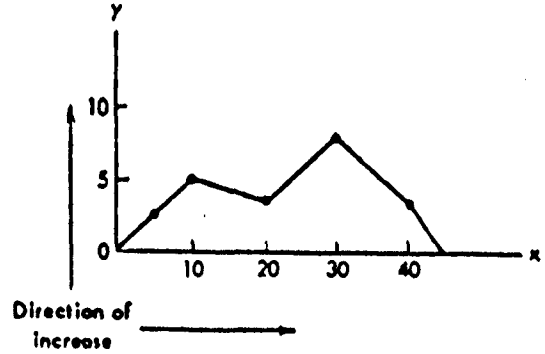
POLYGON

6. In a histogram frequency distribution, the scores and their frequencies are designated by the use of rectangular boxes. In the frequency distribution below, the height of the rectangular boxes indicates the frequency of students that received that particular score. It is called a _____.



HISTOGRAM

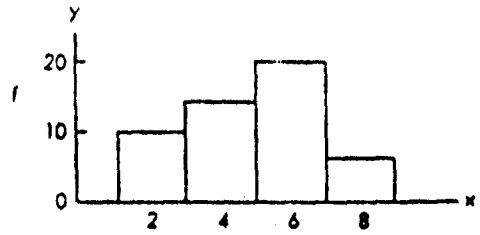
7. It is the accepted practice for the vertical side of a graph, called the *ordinate* axis, to be used to designate the frequency. The horizontal side, called the *abscissa* axis, is used for the scores. Direction of increase is upward for the frequency on the ordinate axis. Direction of increase for the variable is from left to right on the _____ axis.



ABSCISSA

2

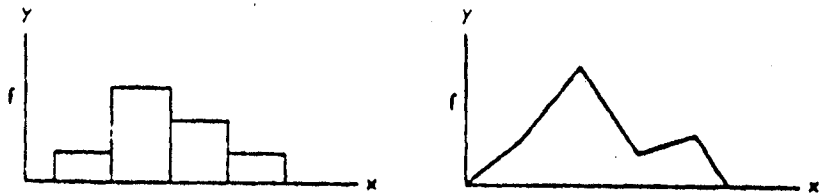
8. On this graph the *f*, which designates the frequency, is the _____ axis, and the *x*, designating the variable, is the _____ axis.



ORDINATE

ABSCISSA

9. The two most common graphs used to illustrate frequency distributions are the frequency polygon and the _____. Graph A. is a _____ Graph B. is a _____



Graph A Graph B

HISTOGRAM

HISTOGRAM

FREQUENCY POLYGON

10. After scores have been tabulated into a frequency distribution, a measure of *central tendency*, or central position is often calculated. Central tendency gives us a concise description of the average or typical performance of the group as a whole. Measures of _____ tendency allow us to compare two or more groups in terms of typical performance.

CENTRAL

11. In statistics there are several "averages" or measures of _____ _____ in common use. Three of these are (a) the mean, (b) the median, and (c) the mode.

CENTRAL TENDENCY

12. The mean is generally the most familiar and most useful to us. The mean is computed by dividing the *sum of the scores* by the *total number of scores*. The formula for the mean would be

$$\text{Mean} = \frac{\text{sum of the scores}}{?}.$$

TOTAL NUMBER OF SCORES

13. Instead of stating that the mean is the sum of the scores divided by the total number of scores, it is easier to use the following symbols:
    a. Mean = $\bar{X}$ read "X bar") or *M*. (The symbols $\bar{X}$ or *M* are used when referring to the mean of a sample from the total population.)

3

b. Sum of the scores = ΣX (Σ = sum; X = each score).
c. Total number of scores = N.
Thus the formula for the mean would be $\bar{X} = ?/?$

$\Sigma X/N$

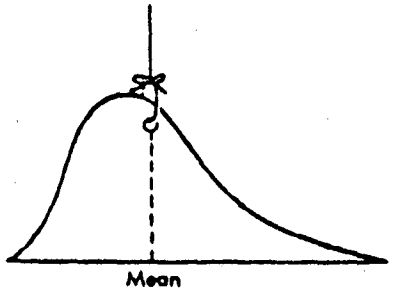
14. **Compute the mean ($\bar{X} = \Sigma X/N$) from the given information:**

Scores (X):	7
	3
	6
	4
Sum of scores (ΣX):	20
Number of scores (N):	4

$\bar{X} = ?/? = ?$

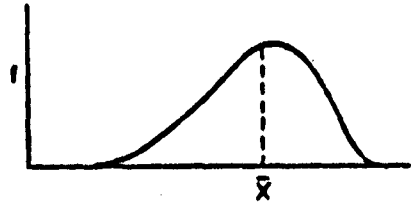
$20/4 = 5$

15. Finding the arithmetic mean of a distribution is analogous to finding the center of moment, or the balance point, in a solid block. If a distribution were suspended by the mean it would hang level or balanced. The mean, whose symbol is _____, is the center of moment in a frequency distribution.



$\bar{X}$ or M

16. Thus, if extremely high or extremely low scores are added to a distribution, the mean tends to shift toward those scores. If the center of balance of the distribution is shifted to one side or the other of the curve, the curve becomes "skewed." The following curve has a few extremely low scores. Consequently, this distribution is _____



SKEWED

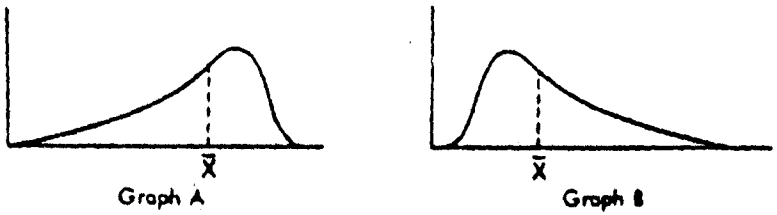
17. Extreme scores, either high or low, tend to _____ a distribution.

SKEW

18. If a distribution is massed so that the greatest number of scores is at the right end of the curve and a few scores are scattered at the left end, the curve is said to be *negatively* skewed. If the massing of scores is at the left end of the curve with the tail extending to the right end, then the curve is *positively* skewed.

4

Graph A illustrates _____ skewness. Graph B illustrates _____ skewness.

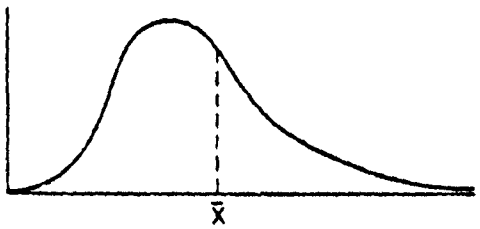


Graph A Graph B

NEGATIVE

POSITIVE

9. This graph's tail is extending to the right because of a few extremely high scores, therefore, it is _____ skewed.



POSITIVELY

A curve is *symmetrical* when one half of the curve is a reproduction of the other half. If you folded a frequency polygon at the mean and the two halves were similar, then the frequency distribution represented by the polygon would be said to be _____

SYMMETRICAL

21. According to the formula for computing the mean ($\bar{X} = \Sigma X/N$), we can define the mean as the arithmetic average of the scores in a distribution. If we added extreme scores to one end of a previously symmetrical curve, the mean would shift towards those extreme scores. Would the curve be symmetrical or not symmetrical? _____

NOT SYMMETRICAL (OR ASYMMETRICAL)

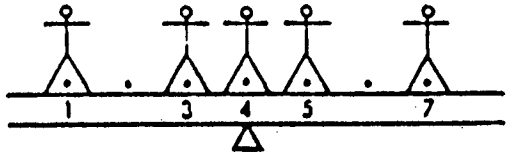
22. Regardless of whether the curve is symmetrical or asymmetrical, the mean is always the center of balance. Does this imply that the mean is centrally located in asymmetrical curves? _____

NO

23. Let us illustrate this point by placing a distribution along an interval scale such as that below. Each figure represents one person. The scale would obviously balance if a fulcrum were

under the middle number, 4. Calculate the mean by the formula $\bar{X} = \Sigma X/N$ to verify this. Was this distribution symmetrical?

_____

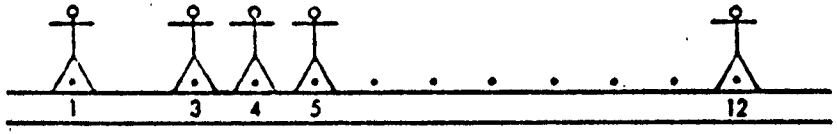


$\bar{X} = 20/5 = 4$

YES

1. If the person with a score of 7 had gotten 12, what would be the mean? _____ Place a fulcrum (i.e., Δ) at the balance point of the scale below. Is the fulcrum centrally located? _____ Is this distribution symmetrical?

_____



5

FULCRUM SHOULD BE UNDER NUMBER 5

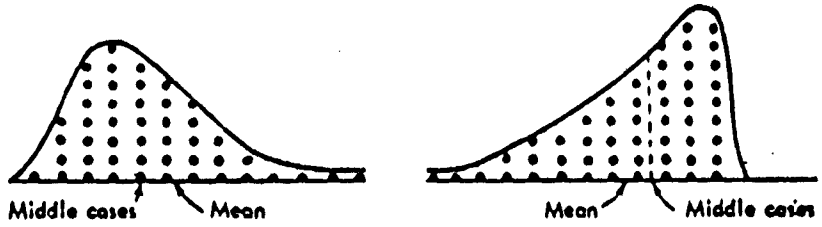
NO

NO

25. What would be the mean for the above distribution if the person who scored 12 had instead scored 22? _____

7

26. When a curve is positively skewed (see graph A) the mean is located to the _____ (right or left) of most of the cases. When a curve is negatively skewed (see graph B) the mean is located to the _____ of most of the cases. (Each dot is one case.)



RIGHT

LEFT

27. If the measures 2, 2, 3, 3, 15 were a frequency distribution, the curve would be _____ skewed.

POSITIVELY

6

28. The preceding distribution of 2, 2, 3, 3, 15 has a mean of 5. What will the mean be if 10 points are added to the score of 15 (making it a score of 25)? _____

7

29. We saw that by adding 10 points to the score of 15, the mean of the distribution 2, 2, 3, 3, 15 was raised by 2 points. The reason for this is that the mean is an arithmetic average and each score contributes to its value. When 10 was added it was averaged or distributed equally among the five scores. This has the same effect as adding a constant of 2 points to each score (10 points/5 scores = 2 points per score). When 2 points are added to each score, the mean is raised by _____ points (from 5 to 7).

2

30. When a constant is added to each score of a distribution, that constant is added to the previous mean to find the new mean. If each score of a distribution is *multiplied* by a constant, the new mean is found by multiplying the old mean by that _____

_____

CONSTANT

31. The distribution 0, 2, 2, 3, 13 has a mean of 4. What would the mean be if each score was *multiplied* by a constant of 2?

_____

8

MEDIAN

32. By adding or substituting an extreme score to a distribution, the mean no longer represents a centrally located score but represents a measure that is more typical of the extreme score.

This causes us to rely on another measure of central tendency which is called the median or the middle score. The median is abbreviated Md or Mdn. The measure of central tendency that is less affected by the addition of an extreme score is the

_____

MEDIAN

33. The median is a *point* on a scale of measurement above which are exactly half the cases and below which are the other half of the cases. The student should note that the median is defined as a *point* and not as a specific measurement, e.g., a score or a case. From the distribution 4, 6, 8, 10, 12, it is easy to see that 8 is the middle score. The score of 8 is at the *point* where there are two scores above and two scores below, hence, 8 is the median. What is the median of 11, 11, 14, 19, 19? _____

_____

14

34. To obtain the median, the measures are arranged in ascending order from the lowest to the highest measure. Then by count-

ing up this scale, the point is selected where there are an equal number of cases above this point and an equal number of cases below this point. The value of this point is the middle or the _____ case.

MEDIAN

35. The median of the distribution 3, 2, 0, 1, 6 can be found by first arranging the measures from the lowest to the _____ _____ number (0, 1, 2, 3, 6). Then we find the middle score or case, which is _____, and that is the median.

HIGHEST

2

36. It is not too difficult to determine the median of a distribution with an odd number of cases—if providing the score at the midpoint has a frequency of 1. In the distribution 3, 5, 7, 9, 11, 13, 13, the median score, which is _____, has a frequency of 1. The score of 13 has a frequency of _____ _____.

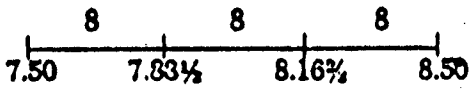
9

2

37. A distribution whose midpoint score has a frequency greater than 1 (e.g., 5, 6, 9, 9, 9, 11) presents a special problem. To overcome this, the student needs to know what is meant by "the interval of a score." For our purposes, the interval of a score ranges from .5 unit below a given score up to .5 unit above a given score. For example, the score of 9 includes all values within the limits of 8.5 up to 9.5. The exact midpoint of the interval whose lower and upper limits are 8.5 and 9.5, respectively, is 9. The score of 17 would represent the interval from 16.5 up to _____

17.5

38. In the distribution of 5, 6, 6, 7, 8, 8, 8, 11, 13, 15, the score of 8 has a frequency of _____ The score of 8 has an interval range from 7.5 up to _____ It is assumed that the unit of three scores (8, 8, 8) is spread equally through the interval of 7.5 to 8.5. Each 8 occupies ⅓ (0.33⅓) of a unit. For example:

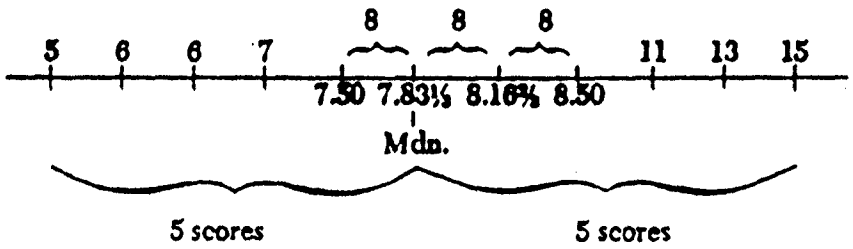


3

8.5

8

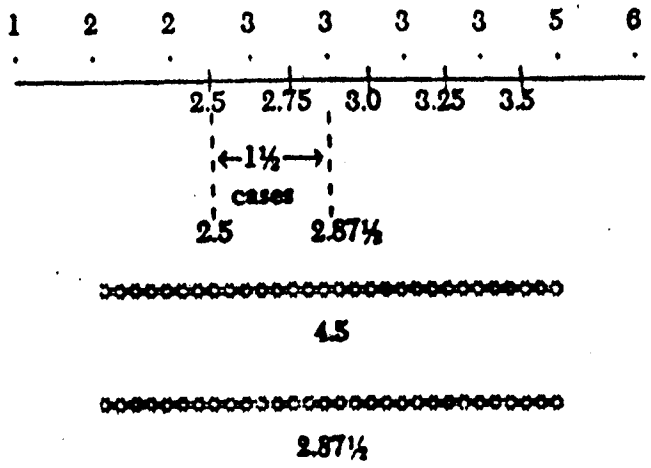
39. The midpoint of the distribution 5, 6, 6, 7, 8, 8, 8, 11, 13, 15, where half the scores are on one side and half the scores are on the other, is between the fifth and _____ score. Below the interval 7.5 to 8.5 there are four scores, consequently the fifth score extends ⅓ of the way into the three score unit. Thus, the point between the fifth score and the sixth, which is the _____, is found by adding ⅓ of the unit to 7.5 (7.5 + .33⅓ = 7.83⅓). Note the illustration:



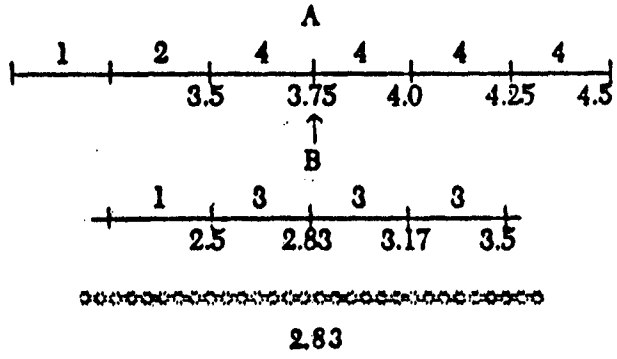
SIXTH

MIDPOINT OR MEDIAN

40. The distribution 1, 2, 2, 3, 3, 3, 3, 5, 6 has nine cases. The median of these nine cases is a point which has 4.5 cases below it and _____ cases above it. This midpoint falls within a four-case unit. Below the lower limit of 3 (which is 2.5) there are three cases; therefore by extending one and one-half cases into the interval of 2.5 to 3.5 we can locate the median. Each 3 accounts for one-fourth of the four-case unit, hence one and one-half cases is equal to ¼ plus ⅛ of a unit (0.25 + 0.12½). The value of the median is 2.5 + 0.37½ = _____. It is illustrated as follows:



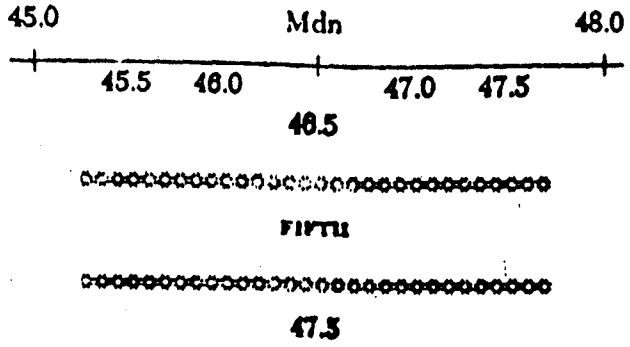
41. The same principle applies to distributions with even numbers of cases except that the median falls midway between the two middle cases. For example in distribution A, shown below, the arrow indicates the median. Draw an arrow to indicate the median of distribution B.



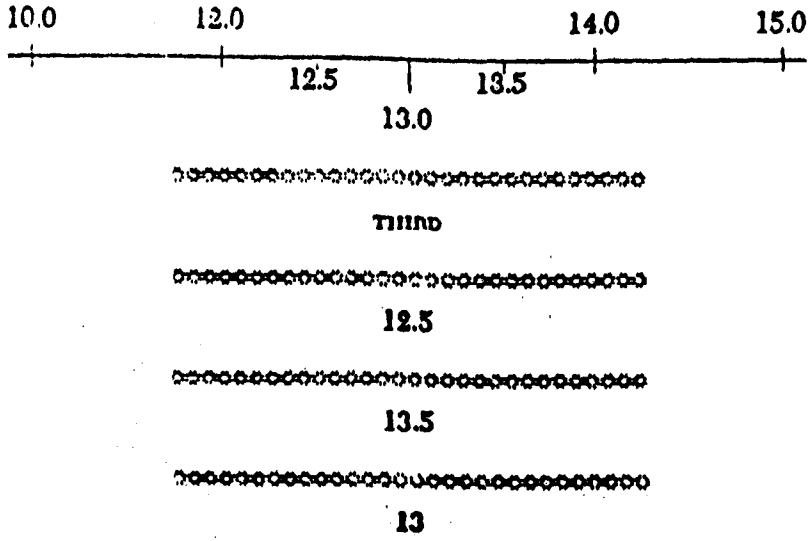
42. The distribution 40, 43, 44, 45, 48, 53, 56, 60 has eight cases. The median is a point midway between the fourth and _____

9

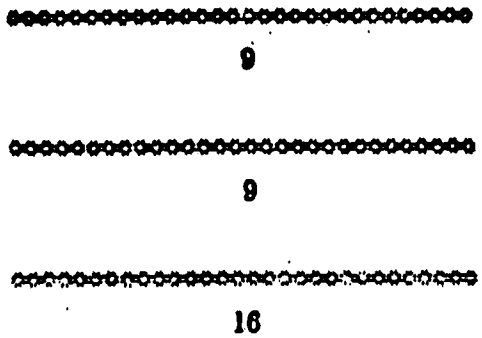
_____ scores. The upper limit of 45 (the fourth score) is 45.5 and the lower limit of 48 (the fifth score) is _____ and midway between these two limits is the point 46.5, which is the median (as noted in the illustration below)..



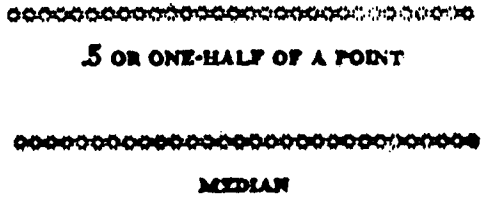
43. The distribution 10, 12, 14, 15 has four cases. The median is a point midway between the second and _____ case. The upper limit of 12 is _____, and the lower limit of 14 is _____ and midway between these two limits is the point _____, which is the median. Note the illustration:



44. The values of the median as a method for obtaining the most typical measure of central tendency is a skewed distribution becomes even greater the more extreme the end score is. For example, the median of 0, 6, 9, 10, 10 is _____ The mean of this distribution is 7 (i.e., 35/5). If one of the 10's is substituted by the extreme number 55 (the new distribution would be 0, 6, 9, 10, 55), the median remains _____ but the mean is now _____



45. The mean and median will *both* be affected if the score of 55 is *added* to the distribution 0, 6, 9, 10, 10. The new distribution would be 0, 6, 9, 10, 10, 55. The mean is 90/6 = 15. The addition of the extreme score shifted the value of the mean 8 points to the right. How far to the right was the median shifted? _____ Which measure of central tendency was affected the least? _____



46. When we want to minimize the effect of one or more extreme scores, we should use the _____ to represent the average score of the distribution.

MEDIAN

47. The median, for both odd and even number of cases, is the point on a distribution where there are an equal number of cases above and _____ that point.

BELOW

MODE

48. A third measure of central tendency is the mode. It may be defined as the one value or score which occurs with the most frequency. The mode of the series 2, 3, 4, 4, 4, 5, 5 is 4. The mode of the series 7, 8, 10, 10, 10, 11, 11 is _____. The median is _____

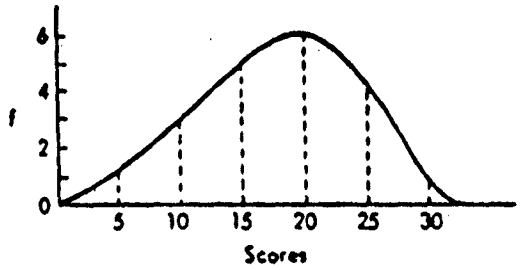
10

10

49. Is it possible for a distribution to have a median and a mode of the same value? _____ (yes or no)

YES

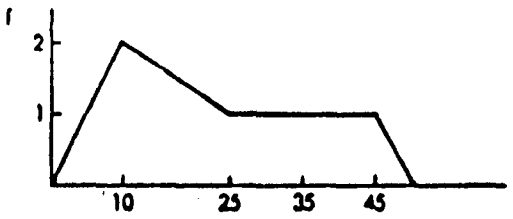
50. The mode is used as a simple, inspectional "average" to show, in a hurry, the center of concentration of a frequency distribution. What is the mode or the rough average of the frequency polygon shown below? _____



20

51. The mode is not generally used unless there are a large number of cases in a distribution. When the number of cases in a distribution is small, there is a good possibility of several scores having the same frequency. The frequency polygon shown below is an extreme example. It is evident that the mode is 10 but it does not give a close approximation of the average case. The mean is 25 ($\Sigma X/N = 125/5$). The cases, in ascending order, are

10, 10, 25, 35, 45, with the number 25 at the midpoint; thus _____ is the median.



25

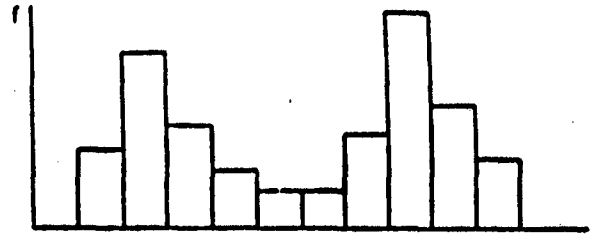
52. The mode is used, in preference to either the median or the mean, when a measure of the most characteristic value of a group is desired. What is meant by "the most characteristic value" can be exemplified by clothing fashions. The _____ is what is being worn the most.

MODE

53. The mode is used also to be sure that the average you obtain exists in actuality. In finding the average size of automobile tire that is purchased, the mean or median size might be a tire that doesn't exist. Therefore, one would want to know the size of tire bought most often. This would be the _____.

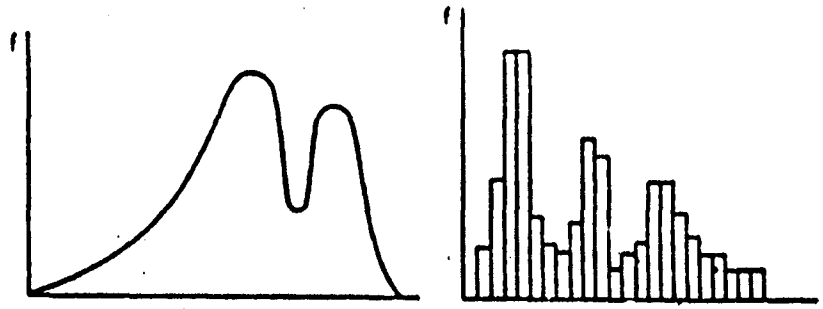
MODE

54. In addition to serving as a measure of central tendency, the concept of modality is useful in describing the shape of some distributions. If a histogram or a frequency distribution has two peaks, it is referred to as a *bimodal* distribution. If a distribution has more than two peaks, it is called *multimodal*. The following histogram appears to have two separate concentrations of frequencies; consequently it is called _____.



BIMODAL

55. The distribution of the frequency polygon illustrated below is _____. The distribution of the histogram is _____. The frequency pologon is negatively skewed, whereas the histogram is _____ skewed.



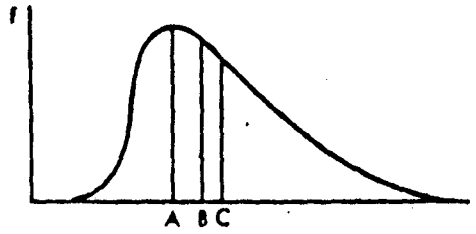
BIMODAL

MULTIMODAL

POSITIVELY

12

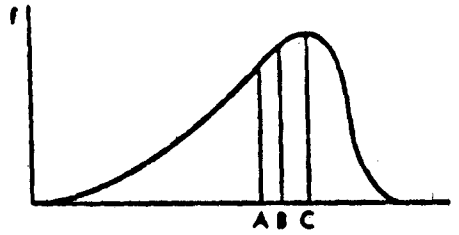
56. The score which occurs with the most frequency is the mode; hence the mode is totally uninfluenced by extreme scores. The mean is greatly influenced by extreme scores. On the basis of these two statements and the preceding exercises on the median, it is evident that line A is the mode (it is totally uninfluenced by the extreme scores). Line B is not affected as much as line C, thus it must be the _____. Line C is the _____ _____; it was influenced the most by the extreme scores.



MEDIAN

MEAN

57. The frequency distribution below is _____ skewed. Line A is the _____. Line B is the _____ _____. Line C is the _____. The mean of a negatively skewed distribution is located left of the center. The mean of a positively skewed distribution is located _____ _____ of the center.



NEGATIVELY

MEAN

MEDIAN

MODE

RIGHT

58. What are the mean, median, and mode of the distribu

1 2 2 4 6 ?_____

3

2.25

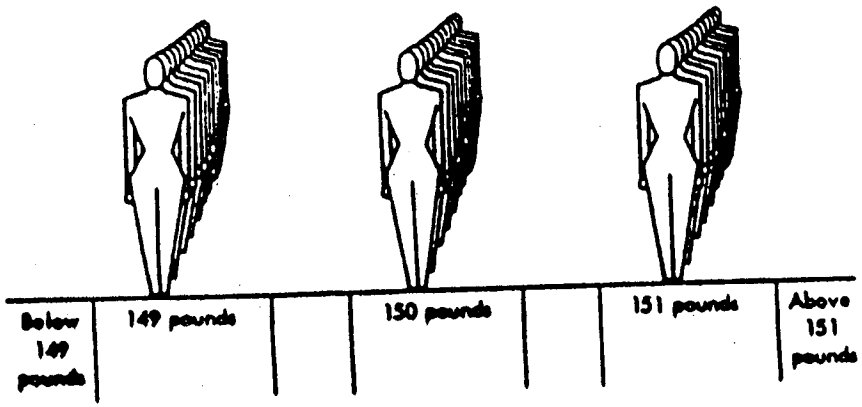
2

59. The mean, median, and mode have been discussed as averages. It should now be evident why these three statistical tools are called measures of _____

CENTRAL TENDENCY

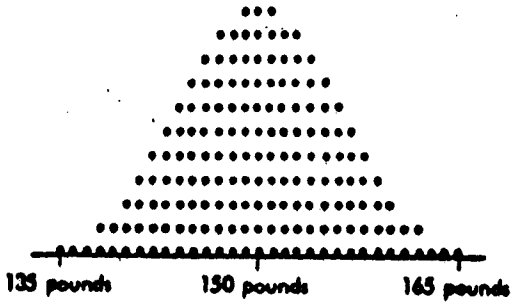
THE NORMAL CURVE

60. Let us suppose that each of 260 students lined up in front of a sign that gave his weight. The sign running from left to right in order of increasing weight are from 135 to 165 pounds. The numbers of persons in any one line is the frequency of that weight. The number of 150 pounders is the _____ of 150 pounders.



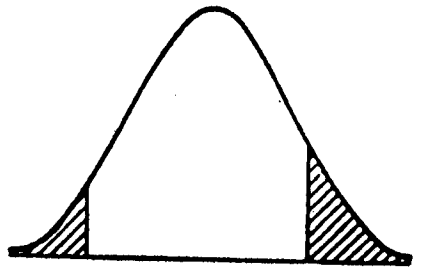
_____

61. From an airplane, the place where this odd event was occurring might look like the diagram below. Each dot represents a _____



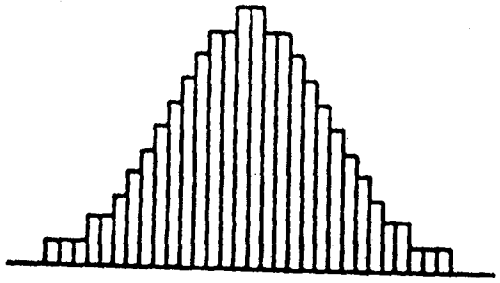
PERSON OR STUDENT

62. Assuming that the students are separated from one another by the same amount of space, the *number of cases would be indicated by the area.* For example with 260 cases, the 26 heaviest students would occupy the extreme right 10% of the crowd. The 13 lightest people would occupy the extreme left _____ _____% of the crowd.



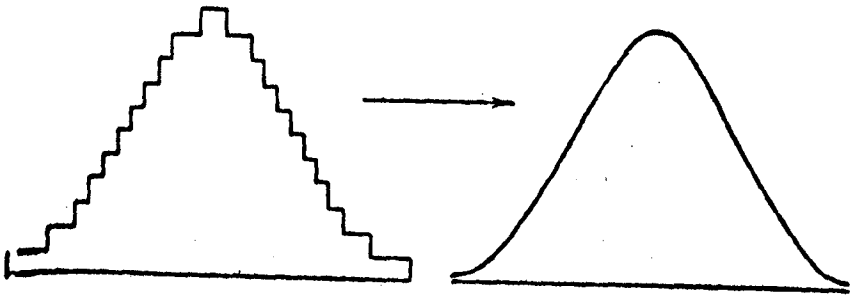
5%

63. If each column of students is represented by a rectangular box, we have our old friend the _____



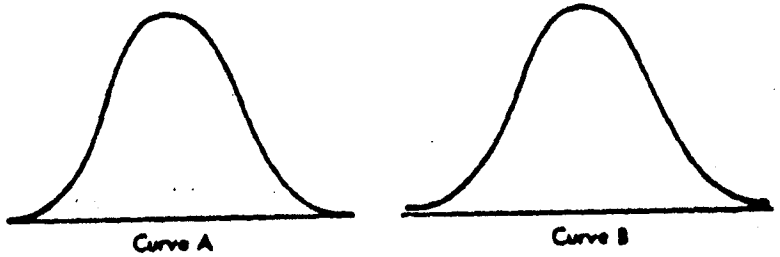
HISTOGRAM

64. If we have a very large number of people and use very small weight categories, the irregular step-like curve would become smooth and continuous. The resulting figure approaches a special type of curve called the normal curve. In frequency distributions normality is not associated with small groups of people but rather with very _____ groups of people.



LARGE

65. In a normal curve (which by definition describes an infinite number of cases) the tails of the curve never touch the baseline. Which curve below could be a true normal curve? _____



B

66. It has been found that quantitative data gathered from high-frequency random measurements of natural phenomena and of many mental and social traits, even though not precisely normal in distribution, can be closely described by the normal _____

CURVE

67. The distributions of such diverse properties as achievement test scores, I. Q., and height and weight of people form approximately _____ _____

NORMAL CURVES

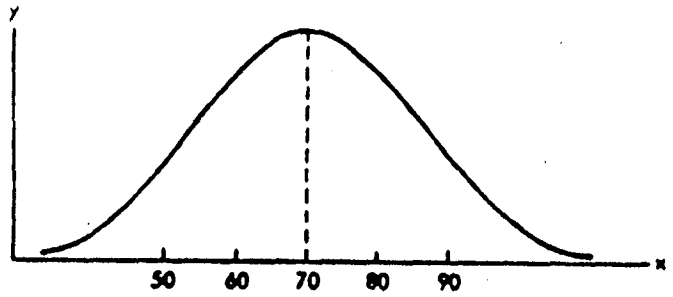
68. The end points of a normal curve remain open and recede indefinitely so as never to touch the abscissa or base line because the number of cases has to be _____

INFINITE

69. When a line approaches infinitely close to another line but does not touch that line, the lines are said to be asymptotic. The end points of a normal curve are _____ to the base line.

ASYMPTOTIC

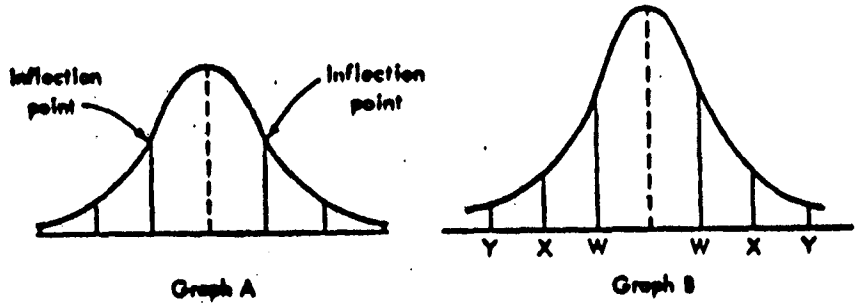
70. The bell-shaped curve illustrated below approximates what the statisticians call a normal curve. Note the following properties:
   a. It is symmetrical.
   b. The mean, median, and mode have the same value (in this instance, 70).
   c. There are thus an equal number of scores on either side of the mean (central axis).
   d. It is composed of infinitely large numbers of _____ _____.
   e. The end points of the curve are _____ to the abscissa (base line).



CASES

ASYMPTOTIC

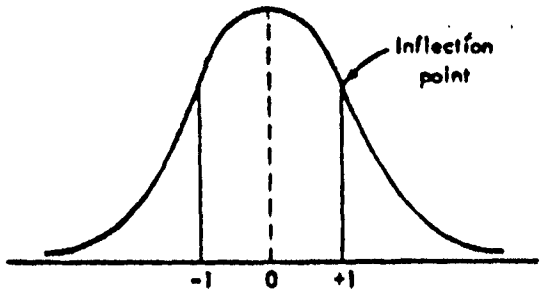
71. Another identifying characteristic of the normal curve is its mathematical construction. There are two points on the normal curve where the curve changes direction from convex to concave. These points are points of inflection (see graph A). Are the inflection points on graph B at line W, line X, or line Y? _____ _____



Graph A Graph B

LINE W

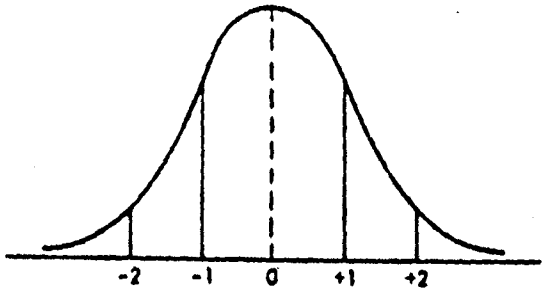
72. By the methods of calculus it can be shown that a line drawn perpendicular from the point of inflection to the abscissa is *one unit* of distance or deviation from the central axis. If one uses this distance as a *standard*, a uniform method of dividing the base line into equal segments can be established. If the central axis is designated as zero, the line one unit of distance to the right would be plus and the line one unit of distance to the left would be _____



MINUS

16

73. Mathematically, the lines −1 and +1 are situated one unit of distance or deviation from the central axis or values (the mean, median, and mode). These two lines are designated as ±1 (read as plus and minus one). Two units of distance or deviation from the central axis are labeled as +2 and _____



−2

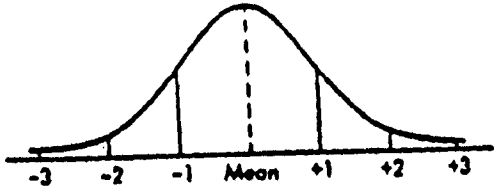
74. Using the unit of distance established by constructing a perpendicular line from the point of inflection to the abscissa as a standard, we can divide the base line into several equal segments. Since the normal curve is asymptotic with respect to the abscissa, one could divide the base line into equal parts indefinitely. All segments would be a uniform or standard distance. The unit of distance was established by constructing a perpendicular line from the point of _____ to the abscissa.

INFLECTION

75. The proportion of cases beyond ±3 units from the center of the normal curve is so small that they are generally ignored. It is thus common practice to illustrate only those cases contained between the arbitrary limits of +3 and _____ units of deviation.

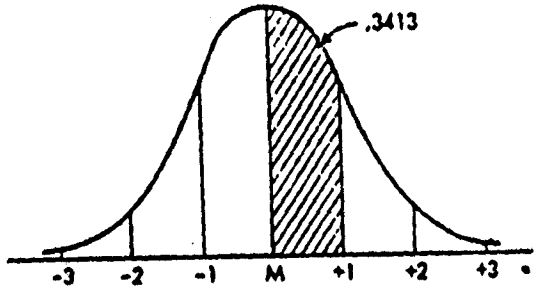
−3

76. Notice that in the graph below each divided segment is equal to the distance from (or the deviation from) the mean to the perpendicular line drawn from the _____ point.



INFLECTION

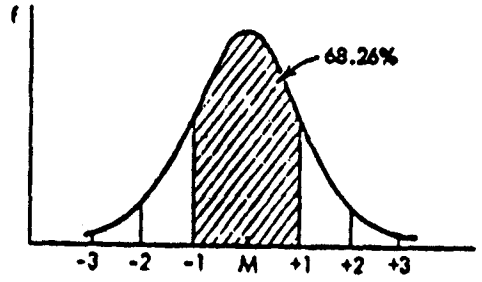
77. The total area under the normal curve may be set to equal 1 or unity. Between the mean and +1 unit of deviation above (to the right of) the mean is .3413 (about ⅓) of the total area, i.e., from the mean to +1 unit of deviation lies 34.13% of the total cases. Since −1 unit of deviation is proportionate to +1 unit of deviation, _____% of the total cases lie between −1 unit of deviation and the mean.



34.13%

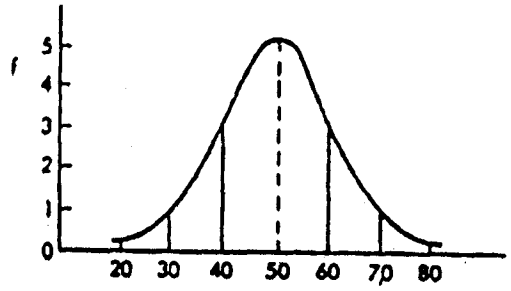
78. Because of the massing of scores around the central values, a little more than ⅔ (2 × 34.13% = 68.26%) of the total frequencies are between +1 and −1 deviations. If a normal distribution has

a total frequency of 1000 scores, approximately 341 scores (34.13% × 1000) are located between the mean and −1 unit of deviation and approximately 341 scores are located between the mean and +1 unit of deviation. How many scores are located between −1 deviation and +1 unit of deviation? _____



683 (632.6)

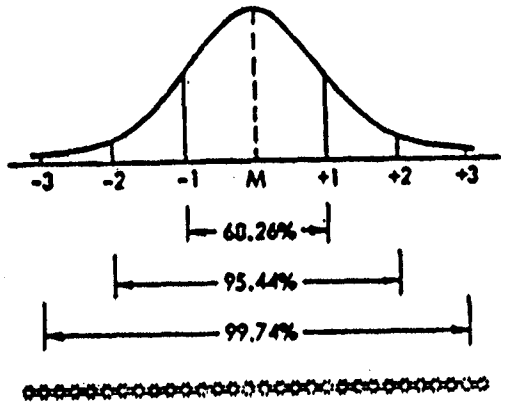
79. In this frequency distribution the deviation points, −1 and +1, mark off the middle _____% of the total scores. They occur at the scores of 40 and _____



68.26% OR 68%

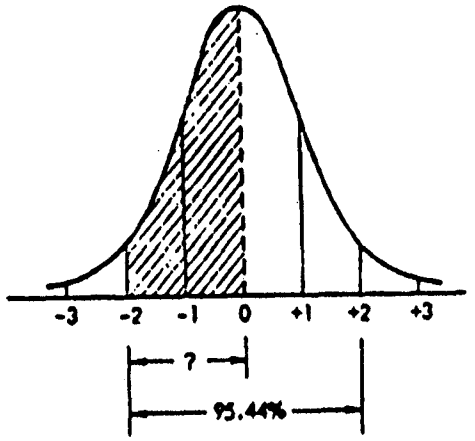
60

80. Although the normal curve extends indefinitely to the left and to the right, the end points of the curve approach the base line so closely that over 95.44% (see graph below) of the area or frequencies are included between the limits −2 and +2 and 99.74% of the cases are included between the limits −_____ and +_____



−3 AND +3

81. The percentage of cases contained between the mean (central axis) of a normal curve and +3 units of deviation is 49.87% (one-half of 99.74%). The percentage of cases between the central axis (the mean) of a normal curve and −2 units of deviation is _____%



47.72%

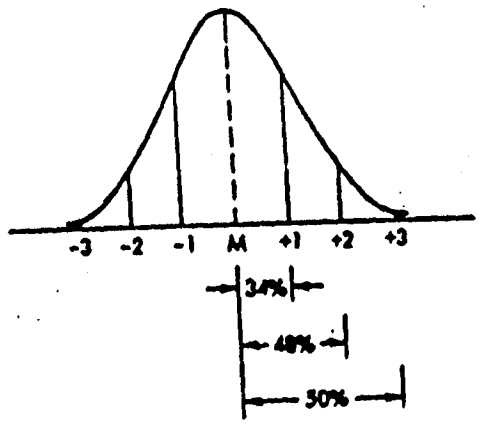
82. As we stated before, for practical purposes the limits of the frequencies of the normal curve rarely exceed those of ±3 units of deviation from the mean. The approximate twenty six hundredths of one percent (.0026) of the total cases existing outside the limits of ±3 are so slight in amount that the unity of the curve is generally assumed to be unaffected. Approximately thirteen hundredths of one percent (.0013) of the total cases extend beyond +3 and approximately _____ hundredths of one percent of the total cases extend beyond −3.

THIRTEEN

83. Though the percentage of cases is very small and insignificant at a considerable distance from the mean (beyond ±3) the proportion of frequencies approaches zero but never equals zero. The reason is that the normal curve is _____ with respect to the abscissa or base line.

ASYMPTOTIC

84. Since the proportion of the total cases that exist beyond the limits of ±3 is so slight, it may be plausible to treat data as if 100% of the total cases fall within ±3 deviations. If one makes this assumption, the percentage of cases is rounded off to the nearest whole percent. Thus (note graph below) the percentages of cases from the mean to +1, +2, and +3 are 34%, 48%, and 50%, respectively. The percentages of cases *from the mean* to −1, −2, and −3 deviations are _____%, _____%, and _____%, respectively.



34%

48%

50%

85. The percentage of cases below the mean is _____%.

50%

86. The percentage of cases between the mean and +1 is _____ _____% of the total cases.

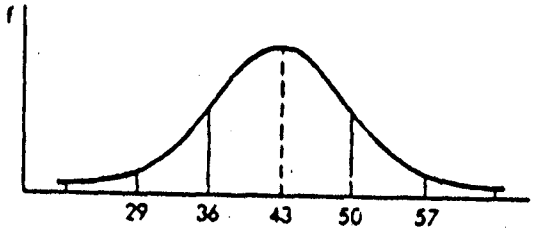
34%

87. The percentage of cases below +1 deviation is 84% (50% plus 34%) of the total cases and the percentage of cases above +1 deviation is _____% of the total cases. The percentage of cases below −1 deviation is _____%.

16% (100% − 84%)

16% (50% − 34%)

88. In relation to the scores on the graph below, about what percentage of cases lie below 43? _____ Between 43 and 57? _____ Below 57? _____ Above 57? _____ Below 29? _____



50%

48%

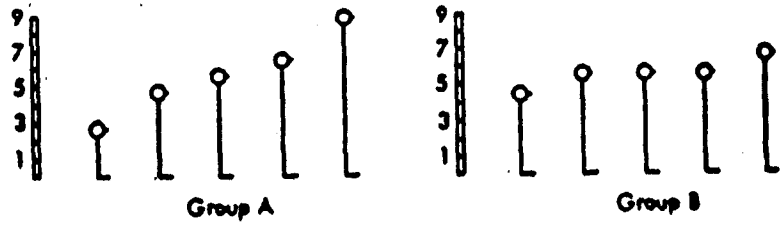
98%

2%

2%

VARIABILITY

89. Descriptions of groups by frequency distributions, central tendency, and normality have been discussed. Another way of describing a group is to have some index of how much variety exists. Consider the height of the two groups of people below. Both groups have a mean and median of 6 feet but the most variable is group _____



GROUP A

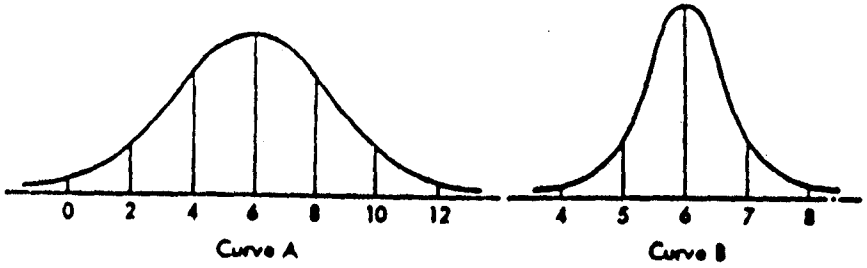
90. One common measure of variability is the range. The range of a set of scores is the distance between the midpoints of the lowest and highest scores. To find the range, subtract the lowest score from the highest score. The range of group A whose heights are 3, 5, 6, 7, 9 is 9 minus 3 or 6. The range for less variable group B whose heights in fact are 5, 6, 6, 6, 7 is _____

2 (OR 7 MINUS 5)

91. If the normal curves below, in which the vertical deviation lines are one standard unit apart, represent large populations,

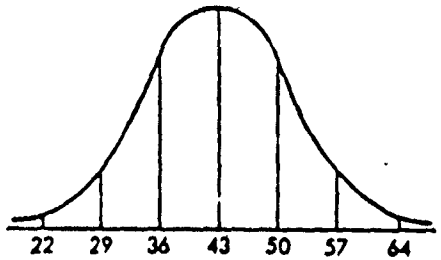
20

which curve represents the most variable group? _____



CURVE A

92. The distance from one deviation line to another is the other major index of variability. It is called a *standard deviation*. In the diagram below, 29 differs from 43 by two _____



STANDARD DEVIATIONS

93. When members of a group deviate very little from each other, the standard deviations are very small. The reverse is true for highly variable groups. Consequently, the variability or diversit of two groups can be compared by the relative size of thei _____

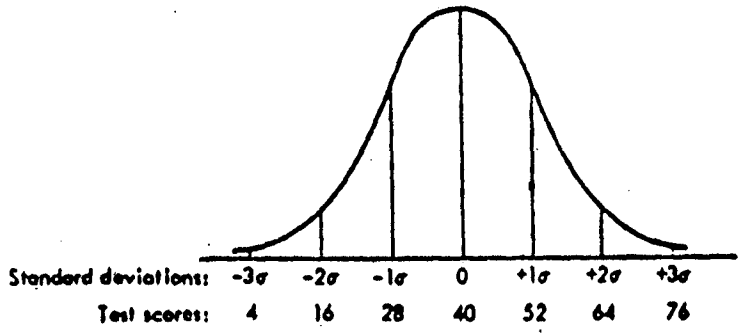
STANDARD DEVIATIONS

94. The capital letters S.D. are used when referring to the standa deviation of a sample of a population. It is common practic to symbolize the standard deviation by the small Greek lette sigma (σ) when referring to the population values. To abbreviat a sample's standard deviation, one could use the capital letter _____. To abbreviate the standard deviation for population, use the small Greek letter _____

S.D.

σ OR SIGMA

95. Suppose that a population of scores is distributed so that th mean is 40 and the distance of 12 points is 1σ (read as on standard deviation). The students one standard deviation abo the mean received the score of _____



52

96. What is the difference, in score points, between the scores of $+1\sigma$ and -2σ of the above illustration? _____

36 (3×12)

97. The standard deviation is a kind of average of all the deviations from the mean score. The amount a score (X) deviates from the mean ($\bar{X}$) is symbolized by the small letter x, that is, $X - \bar{X} = x$. Give the symbols for the following:

Raw score _____

Mean _____

Deviation _____

X

X

x

98. To calculate a standard deviation, the deviations from the mean have to be squared. To square a number, you multiply it by itself. For example, 5 squared, or 5^2, $= 5 \times 5 = 25$; $2^2 =$ ____

4

99. A minus times a minus equals a plus, therefore, $(-4)^2$ or $-4 \times -4 = 16$. Complete the following:

$(-7)^2 =$ _____

$(-3)^2 + (-1)^2 + (-1)^2 + 5^2 =$ _____

49

36

100. The opposite of squaring a number is taking a square root. The square root of 25 or $\sqrt{25} = 5$; $\sqrt{16} =$ _____

4

101. When some arithmetic occurs inside a square-root sign, work the arithmetic before taking the square root.

$\sqrt{1+3} = \sqrt{4} = 2$

$\sqrt{\frac{5^2+7}{2}} = \sqrt{\frac{32}{2}} = \sqrt{16} =$ _____

$\sqrt{\frac{(-3)^2+(-1)^2+(-1)^2+5^2}{4}} = \sqrt{\frac{?}{4}} = \sqrt{?} =$ ____

Note: When we take the square root of a number, both positive and negative roots occur; but we are concerned only with the positive roots.

4

$\sqrt{36/4} = \sqrt{9} = 3$

102. Squaring, dividing and taking the square root are used in solving the formula for the standard deviation: S.D. $= \sqrt{\Sigma x^2/N}$. The symbol $\sqrt{}$ directs a person to take the square _____

root

22

103. The formula for the standard deviation, $\sqrt{\Sigma x^2/N}$, can be executed in six steps:

1. Compute the mean ($\bar{X}$).
2. Subtract the mean from each score to find the deviations from the mean ($X - \bar{X} = x$).
3. Square each deviation from the mean (x^2).
4. Add the squares of the deviations (Σx^2).
5. Divide the Σx^2 by N ($\Sigma x^2/N$).
6. Take the square root of $\Sigma x^2/N$ ($\sqrt{\Sigma x^2/N}$).

The steps above illustrate why the standard deviation is defined as the square root of the average of the squared deviations from the _____.

MEAN

104. To see how to calculate a standard deviation, let's assume a distribution of 0, 2, 2, 8. The mean ($\bar{X}$) is 3.

$X - \bar{X}$ or x (deviations) = (0–3), (2–3), (2–3), (8–3)
= −3, −1, −1, ____
x^2 (squared deviations) = 9, 1, 1, ____
Σx^2 (sum of squared deviations) = 9 + 1 + 1 + ____ = ____

5

25

25 = 36

105. We have found the sum of the squared deviations (Σx^2) to be 36. With an N of 4, the average of the squared deviations from the mean is $\Sigma x^2/N = 36/4 = 9$. The square root of the average of the squared deviations from the mean equals 1 standard deviation. That is, S.D. = $\sqrt{\Sigma x^2/N} = \sqrt{36/4} =$ _____

3

106. Calculate the standard deviation for the distribution 1, 3, 4, 5, 7.

a. $\bar{X} = 4$
b. $(X - \bar{X})$ or $x =$ (1–4), (3–_), (_–_), (_ –), (_ –)
c. $x^2 =$ —, —, —, —, —
d. $\Sigma x^2 =$ — + — + — + — + —
e. $\Sigma x^2/N =$ ____
f. S.D. = $\sqrt{\Sigma x^2/N} =$ ____

4, (4–4), (5–4), (7–4)
−3, −1, 0, 1, 3
9, 1, 0, 1, 9
4

2

107. Subtracting a constant from or adding a constant to all the raw scores of a distribution does not change the value of the stand-

ard deviation. If the standard deviation of 50, 52, 52, 58 is 3, the standard deviation of 50–50, 52–50, 52–50, 58–50 or 0, 2, 2, 8 is _____

Note: The formula used here for computing standard deviations is based directly on the definition of standard deviation. This formula was chosen because it best helps one understand the basic concept of standard deviation. If you have to calculate standard deviations of various data, consult a statistical methods book for the standard deviation formula most appropriate to your data.

3

108. The range is easier to understand and easier to calculate than the standard deviation but it has some serious disadvantages. Not much else can be done with the range. The standard deviation (and its square called the variance) is the basis of a whole branch of statistics. The measure of variation having the greater versatility is the _____.

STANDARD DEVIATION

109. The size of the range depends a good deal upon the size of the sample. There is more chance of simultaneously drawing a very high score and a very low score when the sample is larger. Consequently, range generally increases with an increase in the size of the _____

[illegible]

110. Because all the scores are used in computing the standard deviation while only two scores (the highest and lowest) are used in computing the range, the standard deviation is much more stable than the range. The most stable measure of variability is the _____

STANDARD DEVIATION

111. For example, a sample of 20 scores could be drawn at random from a population of 200 scores. The standard deviation and the range could now be calculated and the 20 scores returned to the population pile. If this process were repeated many times, the standard deviations would vary in size much less than would the _____.

RANGE

INTERPRETING TEST SCORES

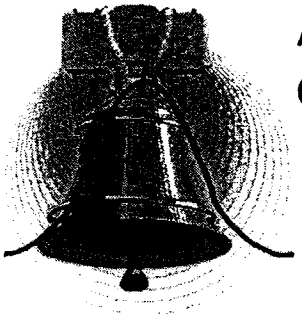
112. The number of correct answers that a person acquires on a test is called his raw score. Assuming that each question on a test counted one point, a raw score of 12 would mean that an individual answered _____ questions of the test correctly.

12

NOMINAL, ORDINAL, AND INTERVAL SCALES

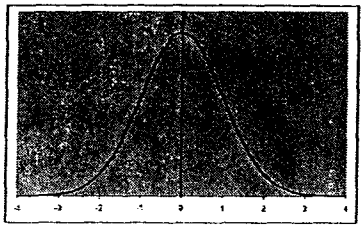
Often the methods used in analyzing or interpreting experimental data differ according to the measurement scale used. The nominal scale merely names categories, such as male-female, or success-failure, or Democrat-Republican-Independent. There may be any number of categories, but every item must be uniquely classified as belonging to one or another of them. If numbers are assigned, such as 1 = male and 2 = female, the numbers have no meaning as an actual numerical measure; in other words, it makes no sense to talk about 1.5 as being half-male and half-female. The ordinal scale is a ranking measure, such as private-corporal-sergeant, or first-second-third place in a race, or A-B-C-D-E in a grading scale. There still may be categories named, but with ordinal measure the categories now are ranked as being "better" or "worse," "higher" or "lower," etc. The interval scale is a true numerical measurement scale in which numerical differences have meaning, such as degrees for temperature, pounds for weight, feet for height, or counting measure for the number of redheads in a class. Sometimes it is difficult to classify certain observations as to measurement scale, and sometimes the classification will vary according to the context of the experiment. For example, the group assignment of children to reading groups would be nominal if the choice of any particular child for Group A, Group B, and so on were made arbitrarily. However, if the children were assigned according to their reading ability, it would be ordinal, since it would rank them. If, after the children are assigned, we wished to count the number of children in each group, our counts would involve the interval scale.

Statistics Part II: Understanding the Bell Curve



Administrative Staff Analyst

OSA Training



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
- Interpreting bell curves by their mean, variance, and standard deviation
- Understanding and calculating Z scores
- Proportion: Calculating the area under the curve
- Skewness in Curves
- Correlation: What is the relationship between two variables?

2

27

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.

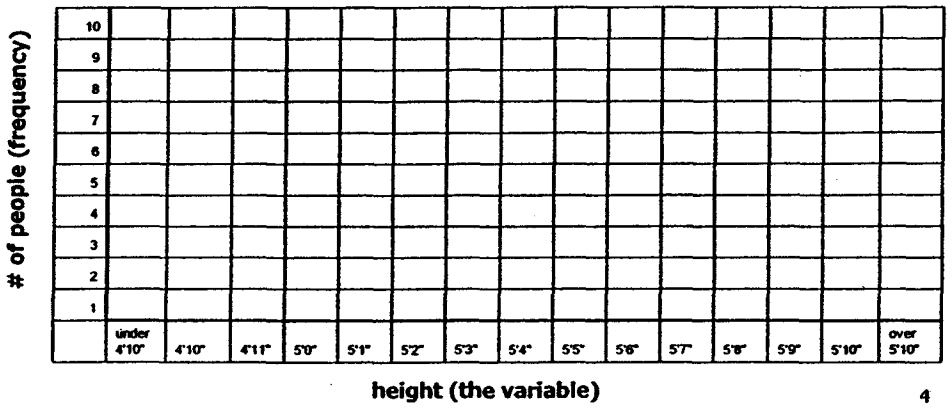
3

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



4

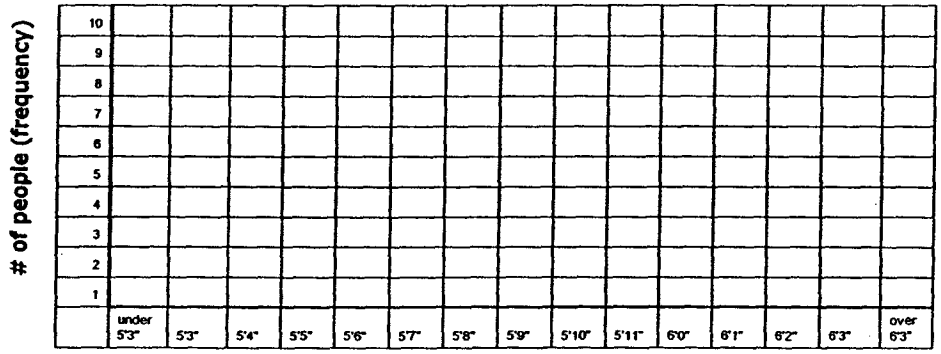
28

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

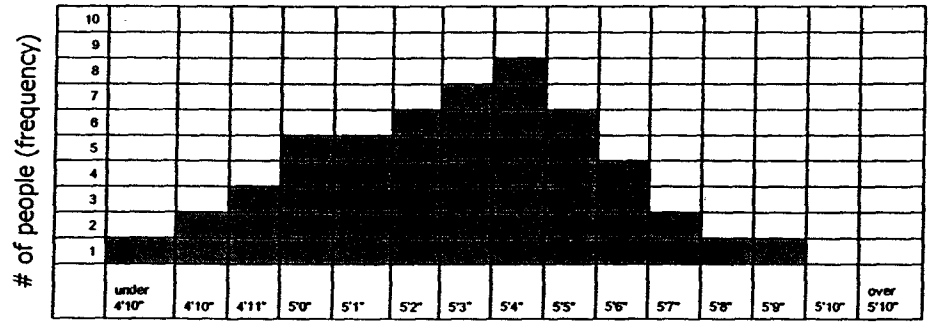


5

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



6

Properties of Populations

Definition of a 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.

7

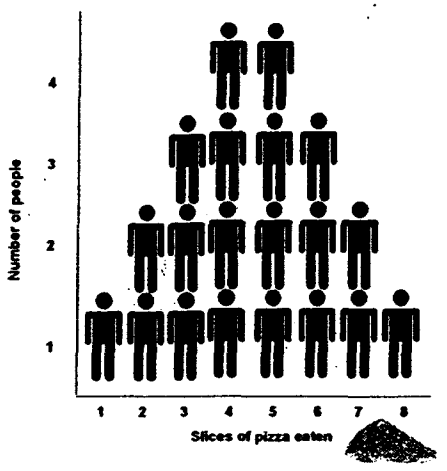
Example: The Hungry Softball Team

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.



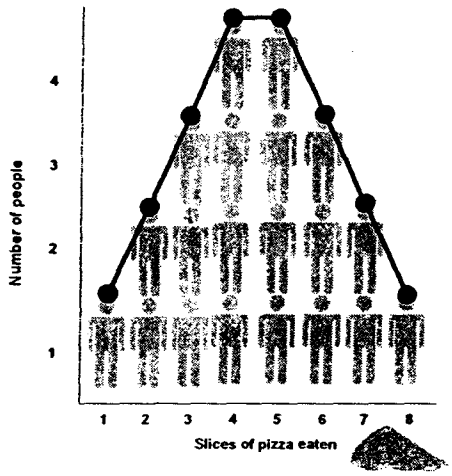
8

30

Example:
The Hungry Softball Team

Graph

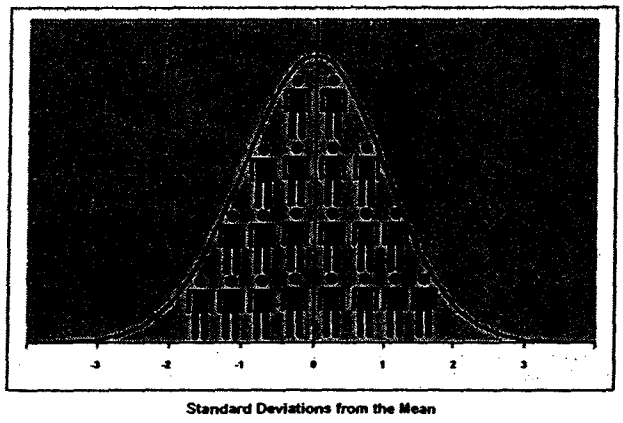
A line shows how you could draw a simple graph using the tops of the heads of each group of players.



Example:
The Hungry Softball Team

Graph

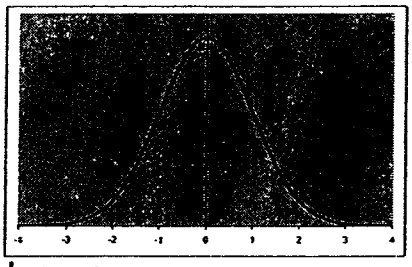
This graph is a simplification of how you could graph pizza slices eaten into a bell curve.



10

31

Bell Curves: What are they?



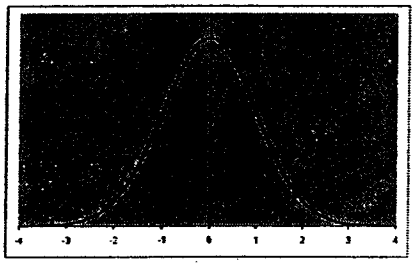
Basic Properties

- 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 at its mean.
- Its unique properties make it very useful in making statistical calculations.

11

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



12

32

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.

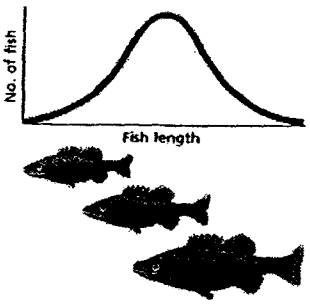


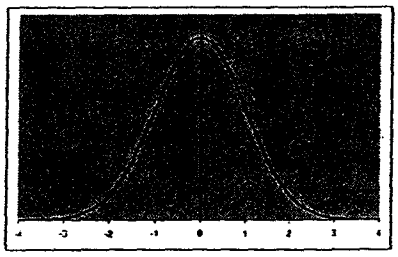
Figure 16-2 A bell curve illustrates how most members of a population are grouped in an average range for a given trait while only a few are at the extreme ends of the range.

13

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

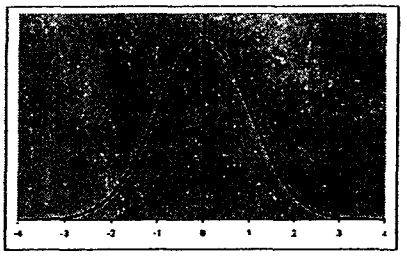
14

33

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



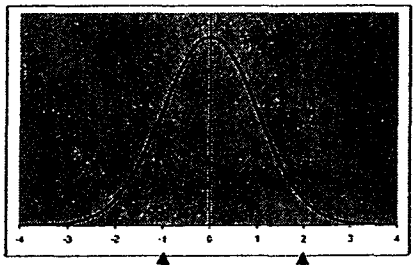
Variance = Sum of (each observation − mean)²
/ # of all observations

15

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

16

34

Bell Curves: What are they?

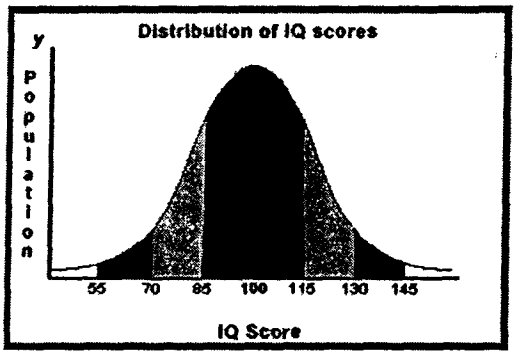
Advanced Properties

- 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.

17

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.



18

Z Scores: What are they?

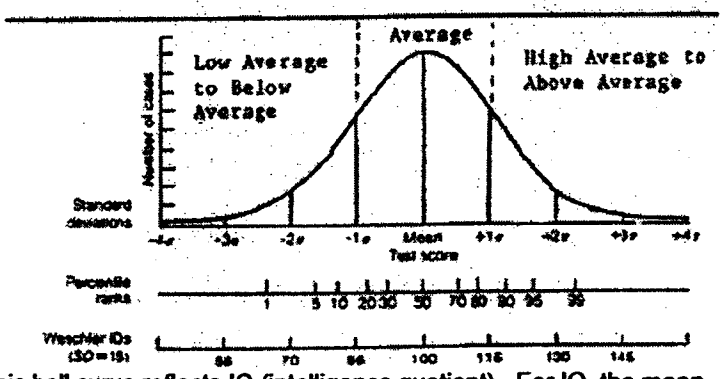
Mean and Standard Deviation

- 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.
  - 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.

19

Z Scores: How to Calculate Them

$$\text{Z score} = \frac{\text{Mean} - \text{observation value}}{\text{Standard deviation}}$$



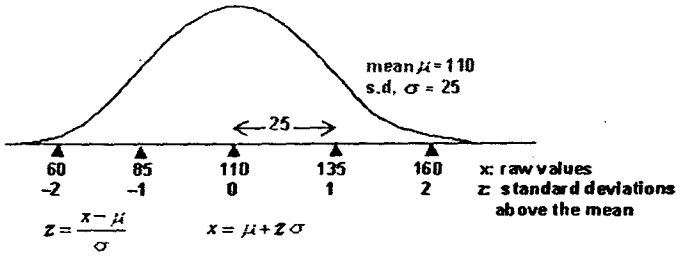
This bell curve reflects IQ (intelligence quotient). For IQ, the mean equals 100 and the standard deviation equals 15.

20

36

"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.



21

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%.

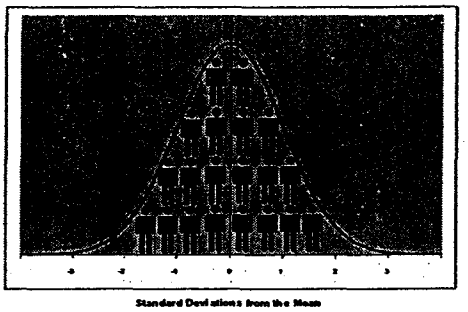
22

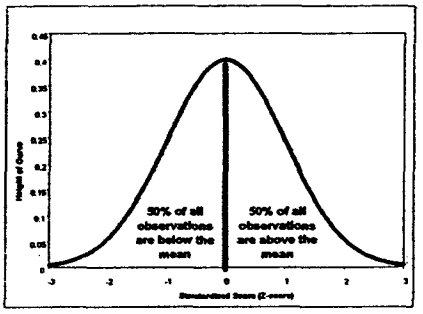
37

Proportion Example: The Hungry Softball Team Redux

Definition of Proportion

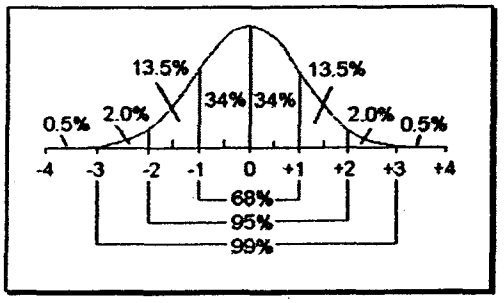
Think of proportion as counting the number of observations that would fit under a certain part of the curve.





23

Proportion: Properties of All Normal Distributions

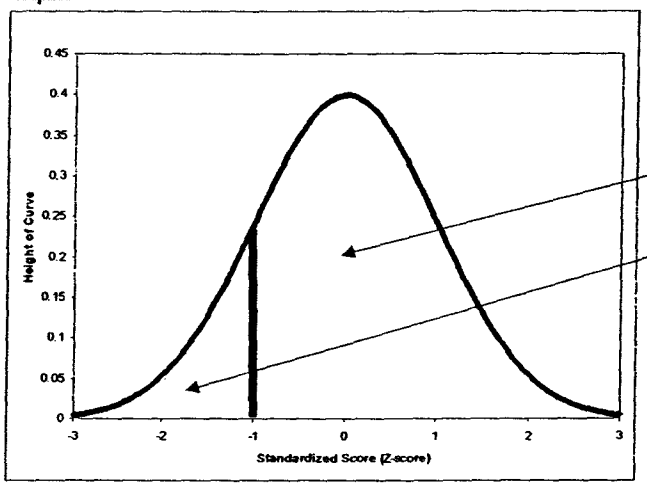


- 68% of observations fall within 1 standard deviation of the mean (34% on either side)
- 95% of observations fall within 2 standard deviations of the mean (47.5% on either side)
- 99% of observations fall within 3 standard deviations of the means (49.5% on either side)

24

38

Proportion: Example



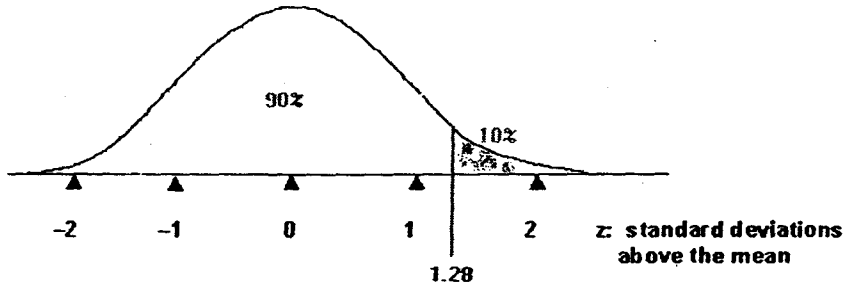
84% of observations fall to the right of the Z score of -1.

16% fall to the left of it.

25

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.

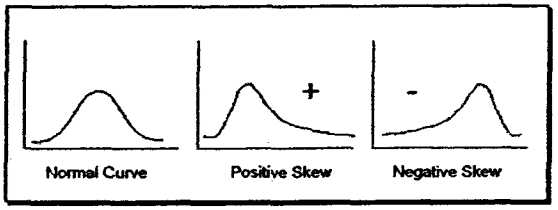


26

39

Skew in Bell Curves

- 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]



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

27

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.

28

40

Correlation: What is the relationship between two variables?

Direction

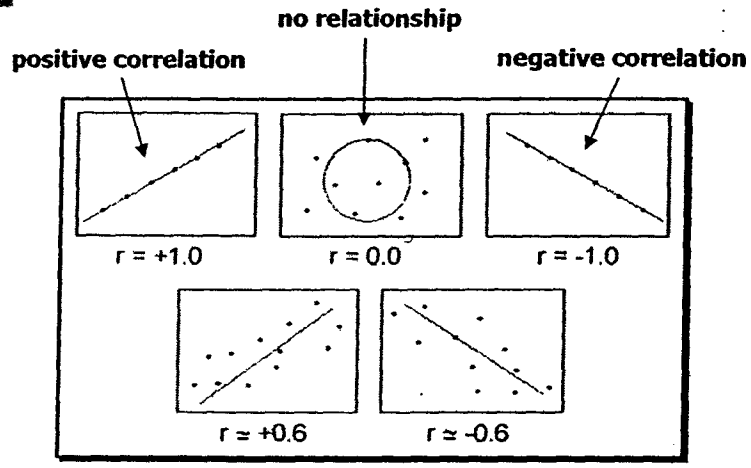
- Positive correlation: When you see an **increase** in one variable, you also tend to see an **increase** in the other variable.
  - Example: Income and SAT scores. As income rises, so, too, do SAT scores tend to rise for students.
- Negative correlation: When you see an **increase** in one variable, 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.

29

Types of Correlation



30

41

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
- 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.

31

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

32

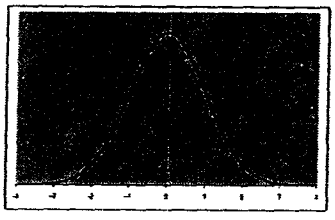
42

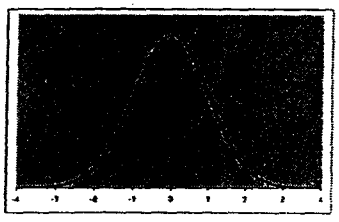
Sources/Additional Resources

- 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

33

Notes

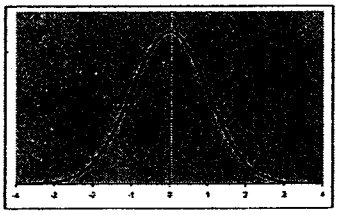


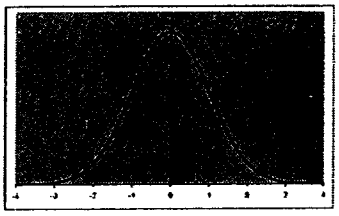


34

43

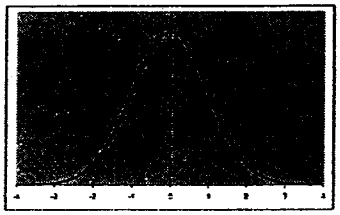
Notes

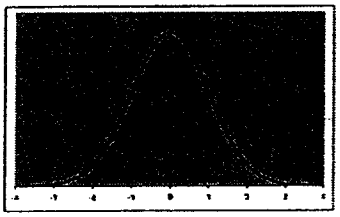




35

Notes





36

44

Module 2

LEARNING ACTIVITY I

THE BASIC CONCEPTS OF STATISTICS

The basic concept of statistics and science is that of the variable. Very simply, a variable is a thing, characteristic, or phenomenon that can take different values; it can vary. For example, weight as a variable can be 10 lbs. or 10,000 lbs., sex has two values (usually), male and female. In contrast to the variable is the constant. A constant is a thing, characteristic, or phenomenon that has a fixed value. For example, pi (π) has a fixed value. We use the term observation to refer to any statement of value we make about a variable in an individual case. For example, on the variable of Empathy, John Jones scores 80 on the Empathy in Interviewing Instrument. The score of 80 is an observation.

When numbers or measurements are collected as a result of observations, we have data. The scores of all ADC clients on an alienation test would be data, or the head count of individuals preferring one type of cash payment over another would be another type of data.

The complete set of things (individual objects, etc.) that we wish to study or observe is known as a population or a universe. An example of population would be all the Public Welfare employees in the State of New York. A part or a subset of a population is known as a sample. So the Public Welfare workers in Chemung County are a sample of the Public Welfare workers in the State of New York.

Any characteristic of a population which we can measure is called a parameter. For example, the mean age of Public Welfare workers in New York (our population) would be a parameter. Parameters usually are represented by Greek letters. (The Greek letter μ (mu) is the symbol we use for the population mean). Any characteristic of a sample we term a statistic: usually italic letters are used to represent statistics (the letter X with a bar over the top, hence X bar is used to refer to the mean score of a sample.)

Suppose the State Commissioner of DSS is interested in finding out the I.Q. scores of unmarried pregnant teenage girls in New York State. In this statement we have one variable which is I.Q. and our population which is all of the unmarried pregnant teenage girls in New York State. To save money and time and to insure accuracy, this project is given to the department research unit. They choose by random means 10 counties in which they are going to test the I.Q. of all unmarried pregnant teenage girls. We now have a sample--the ten counties chosen at random are a subset of all the counties in New York State. They collect all the I.Q. scores of the individual sample members (observations) and transfer all of the observations onto IBM cards. Now we have data. They find that the average I.Q. of the sample group is 98 ($\bar{X}=98$). This we call a statistic (with a small "s"). From this statistic (the average I.Q. of the sample member)

they estimate that the population average will be 98 or $\mu=98$ (given three points either way). This estimation of the population value would be a parameter.

Now Try These:

Match up these terms with the following phrases:
A statistic, a parameter, data, variable, population, observation, sample.

_____ 1. Age of AFDC clients in NY State

_____ 2. Sample: Average age of AFDC client is 32

_____ 3. IBM deck of cards containing all of AFDC clients ages

_____ 4. Population: Average age of AFDC clients in NY State is 32

_____ 5. All AFDC clients in NY State

_____ 6. Age of AFDC client #46 in Albany County is 28

_____ 7. 18 counties chosen randomly to estimate average age of AFDC clients in NY State

Module 2

LEARNING ACTIVITY 2

IDENTIFYING STATISTICAL SYMBOLS AND OPERATIONS

The symbols we use to stand for variables and scores are usually X and Y (capitals). If X stands for height and Y for weight we may wish to see if there is any association. Or X may stand for frustration and Y for aggression and we may want to see if there is any casual relationship between the two (traditionally X stands for the independent variable and Y stands for the dependent variable). A subscript is a symbol, either a letter or number placed slightly below and to the right of the variable symbol: X_1 or X_2 would designate specific individuals. For example, if we were to correlate the height and weight of 5 grade school children, we would use X for height and Y for weight.

	VARIABLES			
	X	(Height)	Y	(Weight)
Individuals	X_1	73	Y_1	120
	X_2	64	Y_2	115
	X_3	70	Y_3	100
	X_4	71	Y_4	98
	X_5	72	Y_5	110

Another class of statistical symbols is called operators. We know these symbols from grade school and high school math. They tell us to add (+), subtract (-), multiply (X), or divide (÷), or they may tell us to take the square root of ($\surd$) or to square (2^2).

An operator that is peculiar to statistics is the Greek capital letter sigma (Σ) which means to sum up or to add up what follows the sign. For example, we could tell someone to add up the five scores on variable X as follows:

$$X_1 + X_2 + X_3 + X_4 + X_5$$

or we could shorten it to:

$$X_1 + X_2 + \ldots X_5$$

(the three dots mean "and so on")

or we could say:

$$\sum_{i=2}^{5} Xi$$

Here the sigma instructs you to add up everything that follows Xi starting with the case below the sigma (i=1) and ending up with the case specified above the sigma which is 5 (we usually use the subscript ith (i) to refer to an unspecified individual).

If we see this:

$$\sum_{i=3}^{7} Xi$$

It would instruct us to sum up the observations starting with the third observation and continuing until the seventh.

We might see this:

$$\sum_{i=1}^{N} Xi$$

which means to sum up all the observations from 1 through the Nth or last (capital N stands for the population number and the small n stands for the sample number).

Frequently we just see

$$\Sigma \bar{X}i \text{ or } \Sigma X$$

When we see these we know we are to sum up all the cases under consideration.

The symbols for constants are most commonly the numbers 1 2 3 4 ... Their value remains constant no matter what the problem may be.

The symbols for parameters will be Greek letters. The two we will encounter are μ and σ (mu and small sigma) μ= parameter mean and σ= population standard deviation (a value we will discuss later).

The symbols for statistics are usually letters. For example, $\bar{X}$ as stated before is a sample mean.

The last set of symbols we will discuss are called connectives. We use them to connect parts of equations.

=	Equal sign
<	Less than
>	More than
≤	Less than or equal to
≥	More than or equal to
≠	Not equal to

These symbols connect statistics or parameters with the operations you wish to make on your raw data.

A formula for the population mean:

$$\mu = \frac{\Sigma\ X1}{N}$$

is broken down as follows:

μ	(mu) is the parameter
=	Connects the parameter with the operations and says it is equal to the result
X_1	Stands for the unspecified scores
Σ	Tells us to sum up all the scores from first to last
—	Tells us to divide by
N	Is the symbol for the number of cases in the population

Now Try These:

1. Match the following statistical symbols with the correct definition.

X	(a)_____	ample mean
Σ	(b)_____	Variable or score symbol
μ	(c)_____	Summation sign
$\bar{X}$	(d)_____	Not equal to
$\neq$	(e)_____	Population mean

2. Write the notation for summing up 12 observations where the N=12.

50

MODULE 4

CENTRAL TENDENCY

Rationale:

One of the keystones of descriptive statistics is the calculation of the appropriate measure of central tendency. With this data the Staff Developer will have a single score that will give him or her a picture of where the average scores are in relationship to the spread of the distribution.

Competency/Terminal objective:

Given formula and data of the appropriate level the Staff Developer will be able to compute three measures of central tendency.

Enabling objective:

The Staff Developer given level of data and computational formula, will be able to:

1. match level of data and measure of central tendency
2. find the mode of a distribution
3. find the median score of a distribution
4. calculate the mean of a distribution.

MODULE 4

LEARNING ACTIVITY I

"MEASURES OF CENTRAL TENDENCY"

INTRODUCTION

Scores on tests have little meaning by themselves. A score of 23 by Joan Pitts on a training post-test tells us very little. Now the score would have more meaning if we knew where the typical trainee scored, then we could compare Joan's score with the average or typical trainee. The point we wish to make is that scores or figures to have any useful purpose must be related to some statistical criterion.

One such set of statistical criteria we call measures of central tendency. We typically group under this heading the mode, median, and mean.

The mode refers to the most frequent score in a distribution; the median refers to the middle score of the distribution; and the mean refers to the arithmetic average.

If we know what the measures of central tendency are for a group of scores we can do several things:

1. We can show where the "typical" score lies
2. We can compare other scores in terms of this typical score
3. We can compare scores on a pre and post basis
4. We can compare the mean achievement of two or more groups
5. We can compare the means of two or more groups on a pre/post basis.

The most powerful measure of central tendency is usually the mean, but in terms of the quality of our data, we can use the mean only when our data is of interval level.

The following table lists the appropriate level of measurement for each measure of central tendency.

Level of Measurement	Measure of Central Tendency
Nominal	Mode
Ordinal	Median (Mode)
Interval	Mean (Mode-Median)
Ratio	Mean (Mode-Median)

Appendix I lists the characteristics and use patterns for the measures of central tendency.

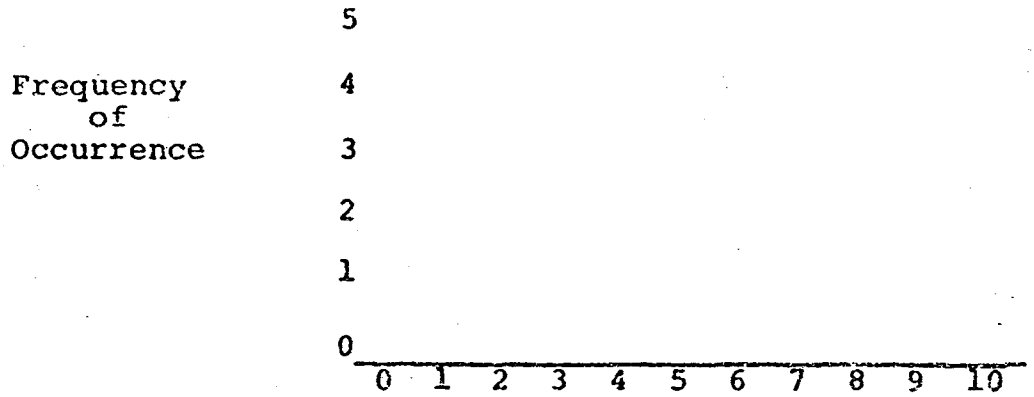
The Mode

Since we define mode as the score which occurs most often in a distribution of scores, try you hand at identifying the mode of the following distribution of scores

3 4 4 5 5 5 6 6 7 7 8 8 9

Mode_____1.

Now chart out the distribution on the matrix below



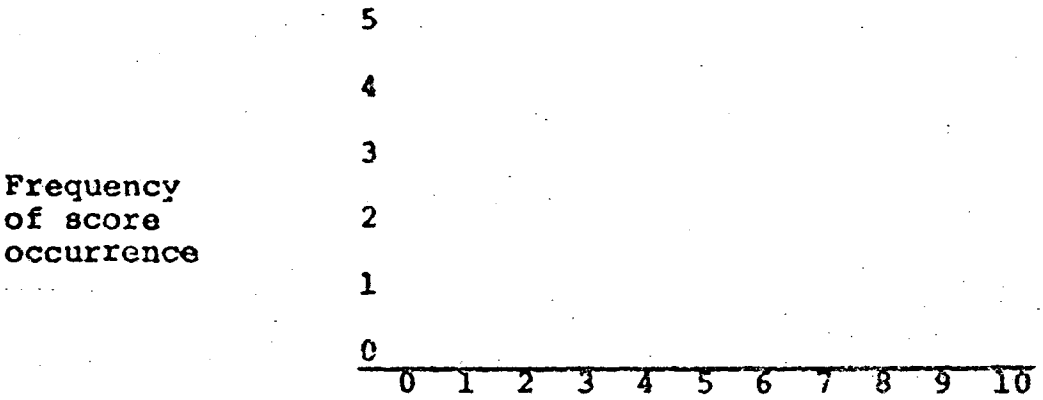
This distribution we term Unimodal.

Let's try our hand once more.

3 4 5 5 6 6 6 6 7 8 9 9 9 9 10

Mode_____2.

Chart out distribution on the matrix below



This distribution we term Bimodal.

When distributions of scores are bimodal it gives us hints at how well our material might be coming across. If the training population for this course was comprised of Ph.D.'s in Mathematics and Public Welfare Staff Development Coordinators, the distribution of scores on a pre-test involving correllation calculations might well be bimodal. We could then divide up the trainees into advanced and beginning levels. Of course Staff Development Coordinators would all be in the advanced group.

The Median

Since the median is the middle point of our distribution of scores, our first step is to arrange the scores in order. If there is an odd number of scores the median is the middle score.

4

8

12

16

18

24

26

Here 16 is our median because there are three scores above it and three scores below it.

Try to find the median for these scores.

6 3 8 20 5 9

Since we have an even number the median will fall between the 6 and the 8. The median is the point halfway between the two adjacent scores.

Here are some practice problems.

1. 1 7 5 3 2 Median = _____

2. 6 4 9 7 Median = _____

The computation of the median can get more complex especially when there is a repetition of the same score near the middle of the distribution. For our purposes we will not go further into this but the reader may want to look up the procedure in any of the statistics books listed in the references.

The Mean

The mean or arithmetic average can be put into formula terms as follows:

$$\text{Mean} = \frac{\text{Sum of scores}}{N}$$

Sometimes this appears as

$$\bar{X} = \frac{\xi X}{n} \quad \text{Sample Mean}$$

or

$$\mu = \frac{\xi X}{N} \quad \text{Population Mean}$$

Remember the Capital Sigma (ξ) means "to sum up" or "the sum of". In terms of operations it tells us to sum up the Xs or the Ys and then divide by the number of scores to get the mean.

Calculate the mean, mode, and median of the following scores:

X
23
21
18
17
15
14
14
12
9
7
ΣX =

Calculate the Mean_____

Median_____

Mode_____

Now let's use your calculator to find the mean of this distribution. Turn to page 1-20 of your handbook.

46
42
39
37
37
35
33
30
29
25
18
ΣX

$\bar{X}$ = _____

Appendix I

Measure of Central Tendency	Features	For Use
MODE	a) A nominal statistic, but can be used for all levels of data	a) When data is nominal or bimodal
	b) The most frequently occurring value	b) When quick knowledge of distribution is desired
	c) Some data distributions may have more than one mode	
	d) Poor estimate of population values	
	e) Incapable of mathematical manipulation	
MEDIAN	a) An ordinal statistic but can be used with interval and ratio level data	a) When ordinal data is used
	b) Can yield a rank or position average	b) When you wish to pinpoint the middle score
	c) Is not sensitive to extreme values	c) In cases where extreme scores would distort a mean
	d) Capable of few mathematical manipulation	
	e) More stable than mode but less stable than mean	
MEAN	a) An interval or ratio level statistic	a) With interval/ ratio level data (other use in "fear and trembling")
	b) An arithmetic average	b) When all scores should weigh equally
	c) Is affected by extreme score	c) When you intent to do further statistical computation
	d) Capable of many mathematical manipulations	
	e) It is the most stable of measures of central tendency	

MODULE 4

Answers to Module 4 - Learning Activity #1

Mode 5 1.

Mode 6 & 9 2.

Median 1. 3

Median 2. 6.5

Mean 15

Median 14.5

Mode 14

Mean 33.73

Module 5: Measures of Variability

Rationale:

The mean score (x) and the standard deviation are indispensable for summarizing distribution of evaluation scores. The mean gives us a one number summary of a distribution. It gives us the balance point;as such it is a type of average.

The standard deviation gives us a one number summary of how the scores are spread. It is useful because it is the most stable measure of variability (extreme scores do not bias it unduly) and it serves as a basis for further inferential statistical procedures.

It is necessary to compute a measure of variability if we wish to see how training scores are spread out around the mean. From this spread we can see how they do or do not approximate the normal curve.

Competency/Terminal Objective:

Given appropriate level of data and formula,the Staff Developer will be able to compute the appropriate measure of variability and relate this variability to the normal curve.

Enabling Objectives:

Given appropriate level of data and formula,the Staff Developer will be able to:

1. Compute the range scores
2. Compute the standard deviation
3. List the characteristics of the normal curve
4. Fit the concepts of standard deviation and the normal curve together.

Learning Activities:

Read attached pages:

1. Computing Measures of Variability
2. The Normal Curve or Gaussian Distribution

Module 5

LEARNING ACTIVITY I

COMPUTING MEASURES OF VARIABILITY

Besides knowing the central tendency of a distribution of training scores it is necessary to know how the scores are spread out or how they vary about the central value of a distribution. The purpose in computing these measures of dispersion is to:

1. first find the amount of spread;
2. secondly we can compare this spread with other distributions;
3. and we can compare the amount of spread with the same group on a pre and post basis.

The simplest measure of variability is the range, and it is simply the distance between the highest and lowest score.

Pre Test Training Scores		Post Scores	
X_1	95	X_1	98
X_2	80	X_2	95
X_3	70	X_3	90
X_4	50	X_4	92
X_5	40	X_5	93
X_6	35	X_6	90

We see the range on the pre-test scores is 60 points, from 95 to 35, while on the post the range is 8 points (98 to 90). We can see that training has cut down on the range of the scores.

The most valuable measure of variability, however, is the standard deviation (which is appropriate for interval level data, but this assumption is often violated). The standard deviation is based (like the mean) on all scores and is a must for calculating many inferential statistics. It is also a standard benchmark for use with the normal curve.

The formula for the standard deviation is

a) Population

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \quad \text{(Little x)} \qquad [\sigma = \text{small sigma}]$$

b) Sample

$$SD = \sqrt{\frac{\Sigma x^2}{N}} \quad \text{(Little x)}$$

When sample of less than 30 is used, we use N-1 in the denominator.

We calculate by setting our data up in a matrix.

Pre or Post Score

X	$x (X - \bar{X})$	x^2
23	(23-15) = 8	64
21	(21-15) = 6	36
18	(18-15) = 3	9
17	(17-15) = 2	4
15	(15-15) = 0	0
14	(14-15) = -1	1
14	(14-15) = -1	1
12	(12-15) = -3	9
9	(9-15) = -6	36
7	(7-15) = -8	64

$\Sigma X = 150$ $\qquad$ $\Sigma X^2 = 224$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{150}{10} = 15$$

The steps in computing are as follows:

1. Sum the X's in the first column which is 150
2. Calculate the mean score which is 15
3. Subract mean from all the individual large X scores i.e. 23 - 15 = 8. The 8 is the score for little x (which is a mean deviation score)
4. Square the little x's for the last column (little x squared = x^2)
5. Sum the squares of the little x's for a total of $\Sigma x^2 = 224$.

Now we have the data necessary to plug into our standard deviation formula

$$\sigma = \sqrt{\frac{\Sigma x^2}{N-1}}$$

We use N-1 because our N is less than thirty.

$$\sigma = \sqrt{\frac{224}{10-1}} = \sqrt{\frac{224}{9}} = \sqrt{24.889} = 4.988$$

Range = 23-7 = 16

Now Try These:

1. Now try your hand at finding the mean, standard deviation and range of this distribution by hand.

X	x	x^2
10		
8		
6		
4		
2		

Mean =_____

SD =_____

Range=_____

62

2. Now we will try to compute the same statistics with our calculator.

X	x	x^2
26		
25		
24		
21		
20		
18		
17		
14		
14		
12		
11		
10		
8		
4		

Mean =_____

SD =_____

Range=_____

Module 5

LEARNING ACTIVITY 2

THE NORMAL CURVE OR GAUSSIAN DISTRIBUTION

The normal curve is not a distribution of actual data but a theoretical distribution derived from mathematical equations (although the original curve was built on data concerning errors). See page 6-5 of Calculator Decision-Making Sourcebook.

The normal curve isone of the most frequently used distributions we have in terms of describing the distribution of scores for populations of at least one hundred. It is characterized by:

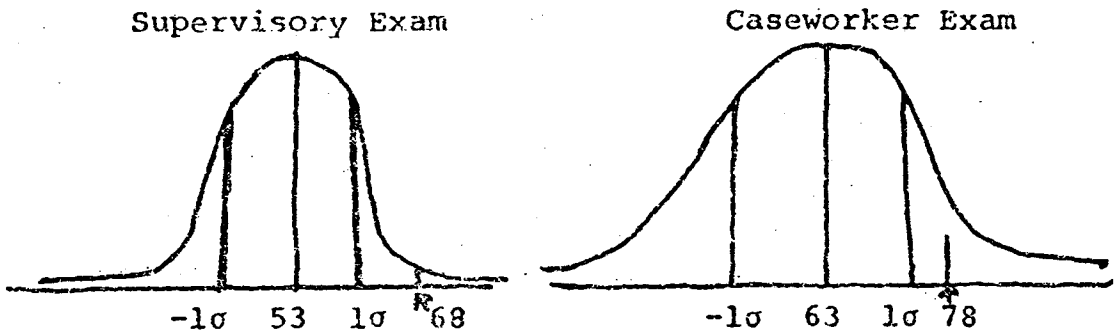
1. a smooth bell shape;
2. scores being distributed symmetrically about the mean;
3. the three measures of central tendency being identical;
4. the scores being distributed in a symmetrical pattern at various standard deviations from the mean.

See page 6-7 for a diagram of a normal curve. We can see from this figure that the standard deviation is an important concept for understanding the normal curve. If we make the assumption that our distribution is normally distributed then 68.26 percent of our cases will fall between plus or minus 1 standard deviation of the mean. In terms of mastery learning this concept is important for it states that scores on aptitude tests are normally distributed and highly correlated with achievement scores. So if we know our trainees' aptitude scores we can predict what the achievement scores will be. For mastery learning our efforts are to gear our training to move the achievement score mean over to the vicinity of the + 2 standard deviation.

The standard deviation is also a good estimate of our error. The more compactly the scores are distributed around the mean, the less our error will be in prediction. The larger our standard deviation the larger our error in prediction. We will compute a standard error in the module on correlation and regression.

The standard deviation helps us in describing individual differences in a group. Suppose you take two civil service tests, one for a caseworker position and one for a supervisory position in I.M. You receive a score of 78 on the caseworker exam and a score of 68 on the supervisory exam. On what exam did you do better? While the mean score on the supervisory exam was 53, the mean score on the caseworker exam was 63. This might indicate that it was easier to get a point on the caseworker exam than on the supervisory exam. In either case you are about fifteen points above the mean. Does that mean you did equally well on both tests? For that answer we need to calculate the standard deviation and we find that for the supervisory exam you are 2.6 standard deviation above the mean and for the caseworker exam

you are 1.2 standard deviation above the exam mean.



In conclusion then, relative to this, you did much better on the supervisory exam than you did on the caseworker exam.

Statistics – Spring 2014
Dr. Sybil M. DeVeaux

Your Name_____

1. Indicate whether each of the following statements typifies *descriptive* **or** *inferential* statistics:

    a) On the average, students in my statistics class are 20 years old.

    b) It was projected that the world's population will exceed 6 billion by the year 2000

    c) Four years is the most frequent term of office served by U. S. presidents.

    d) A recent poll indicates 74 percent of all Americans favor capital punishment.

    e) Children with no siblings tend to be more adult-oriented than children with one or more siblings.

2. Indicate whether the following observations are *qualitative* **or** *quantitative*:

    a) Height

    b) Religious affiliation

    c) Math aptitude score

    d) Years of education

    e) Military rank

    f) Favorite TV program

    g) Place of birth

    h) Grade point average

    i) Daily intake of calories

    j) Highest academic degree

3.

120	153	186	117	140	165	125	128	129	120	123
132	111	117	93	205	130	112	120	180	150	130
120	140	118	130	126	166	110	112	110	185	105
112	132	125	150	116	95	145	119	135	118	139
150	125	112	116	114	125	117	116	95	209	73
16	39	97	15	66	23	59	32	42	44	47
65	25	53	69	50	41	65	36	72	28	

Use the data above to describe the neighborhood drive to collect funds for underprivileged children. Be original.

a) What kind of data is presented?

b) Indicate the class in which each observation falls.

c) What is the shape of the curve?

d) Find the frequency distribution

e) What is your cumulative frequency?

HINT: *Stem & Leaf*

4. Draw a Standard Normal Curve and show the following:

a) The mean

b) Standard deviation

c) Upper and lower halves with values

d) The number of z-scores on either side

e) The proportions in columns B and C and the value

5. Compute the range, mean, median, and mode for the following:

2 17 5 2 28 7 2

6. Use the *Definition Formula* to compute the Standard Deviation:

   12 10 11 8 9 11 9

7. Use the *Computation Formula* to compute the Standard Deviation:

   9 11 12 10 8 11 9

8. Use the *Definition* **or** *Computation* Formula to find the standard deviation:

   1 3 7 2 0 4 7 3

9. Express each of the following scores as a z score:

   a) An IQ of 135, given a mean of 100 and a standard deviation of 15

   b) A verbal score of 470 on a Scholastic Assessment Test (SAT), given a mean of 500 and a standard deviation of 100.

   c) A daily production of 2100 units, given a mean of 2180 units, and a standard deviation of 50 units.

   d) A height of 68 inches, given a mean of 68, and a standard deviation of 3.

   e) A meter-reading error of -3 degrees, given a mean of 0 degrees, and a standard deviation of 2 degrees.

10. Use Table A to find the proportion of the total area identified with the following statements:

    a) Above a z score of 1.80
    b) Between the mean and a z score of -0.43
    c) Below a z score of -3.00
    d) Between the mean and a z score of 1.65
    e) Above a z score of 0.60
    f) Below a z score of -2.65
    g) Between a z score of 0 and -1.96

11. Employees of Corporation A earn annual salaries described below. Find the median salary and identify any outlier(s).

    $34,999 $134,999 $75,000 68,745 $86,745

    $62,888

12. The two main subdivisions of statistics are _____

    and _____.

13. Researchers use three types of data:

    a) _____
    b) _____
    c) _____

14. Draw and label the following curves:

    a) Standard normal curve
    b) Positively skewed curve
    c) Negatively skewed curve
    d) Bi-modal curve
    e) Multi-modal curve

15. To identify a particular normal curve, you must know the (a) **……** and (b)**……**… for that distribution. To convert a particular normal curve to the standard normal curve, you must convert original (c)…… into z-scores. A z-score indicates how many (d)…**……** .an (e)…… is above or below the mean of the distribution. Although there are infinite numbers of (f)……., there is only one (g) **……**. The standard normal curve

has a (h)……..of 0, and a (j) …… of 1. The total area under the standard normal curve equals (j)…….. When using the standard normal table, it is important to remember that for any z-score, the corresponding proportions in columns B and C always sum to (k) …….. or half, furthermore, the proportion in column B always specifies the proportion of area between the (l) …… and the z-score, while the proportion in column C always specifies the proportion of area (m) …… the z-score. Although any z-score can be either positive or negative, the proportions of area, specified in columns B and C, are never (n) ……..

6

Statistics – Spring 2015
Dr. Sybil M. DeVeaux

Your Name_____

1. Indicate whether each of the following statements typifies *descriptive* **or** *inferential* statistics:

   a) On the average, students in my statistics class are 20 years old. DESCRIPTIVE

   b) It was projected that the world's population will exceed 6 billion by the year 2000 INFERENTIAL

   c) Four years is the most frequent term of office served by U. S. presidents. DESCRIPTIVE

   d) A recent poll indicates 74 percent of all Americans favor capital punishment. INFERENTIAL

   e) Children with no siblings tend to be more adult-oriented than children with one or more siblings. INFERENTIAL

2. Indicate whether the following observations are *qualitative* **or** *quantitative*:

   a) Height - QUANTITATIVE

   b) Religious affiliation - QUALITATIVE

   c) Math aptitude score QUANTITATIVE

   d) Years of education QUANTITATIVE

   e) Military rank QUALITATIVE

   f) Favorite TV program QUALITATIVE

   g) Place of birth QUALITATIVE

   h) Grade point average QUANTITATIVE

   i) Daily intake of calories QUANTITATIVE

   j) Highest academic degree QUALITATIVE

3.

120	153	186	117	140	165	125	128	129	120	123
132	111	117	93	205	130	112	120	180	150	130
120	140	118	130	126	166	110	112	110	185	105
112	132	125	150	116	95	145	119	135	118	139
150	125	112	116	114	125	117	116	95	209	73
16	39	97	15	66	23	59	32	42	44	47
65	25	53	69	50	41	65	36	72	28	

Use the data above to describe the neighborhood drive to collect funds for underprivileged children. Be original.

a) What kind of data is presented? UNGROUPED QUANTITATIVE DATE

b) Indicate the class in which each observation falls. **0 – 200 (20S)**

c) What is the shape of the curve? NORMAL

d) Find the frequency distribution

e) What is your cumulative frequency? 76

HINT: *Stem & Leaf*

d) 0 – 19 = 2; 20 – 39 = 6; 40 – 59 = 7; 60 – 79 = 6; 80 – 99 = 4; 100 – 119 = 17; 120 – 139 = 20; 140 – 159 = 7; m160 – 179 = 2; 180 – 199 – 3; 200 – 219 = 2

4. Draw a Standard Normal Curve and show the following:

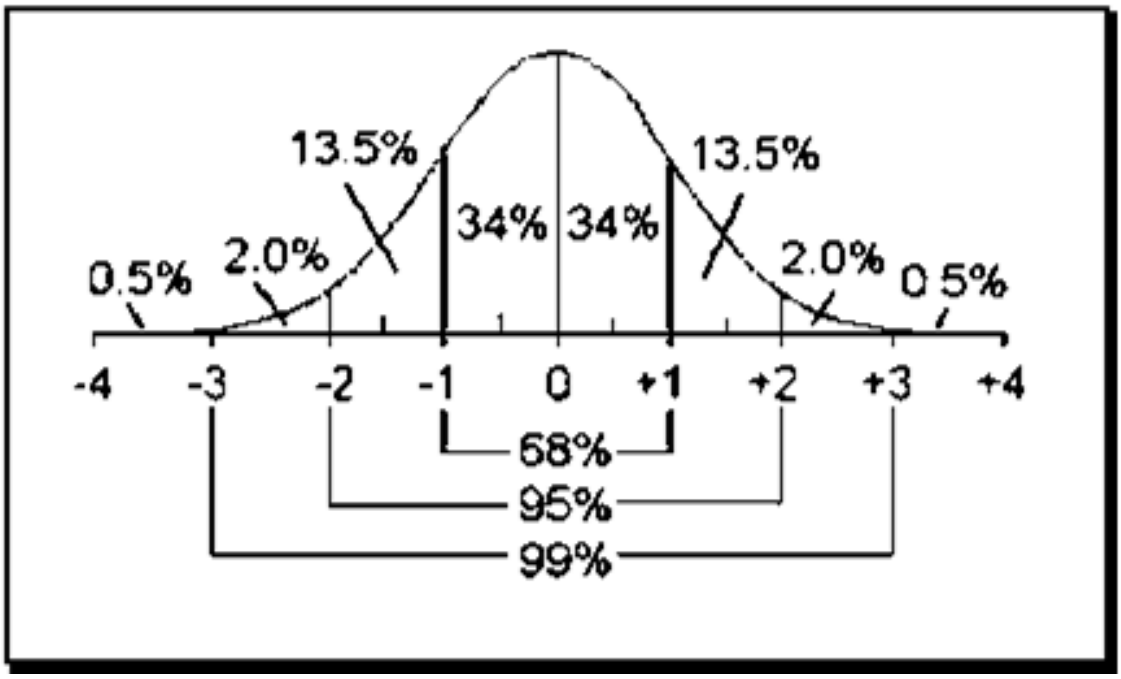
a) The mean = **0**

b) Standard deviation = **1**

c) Upper and lower halves with values = **r= .50; l= -.50**

d) The number of z-scores on either side

e) The proportions in columns B and C and the value



5. Compute the range, mean, median, and mode for the following:

2 17 5 2 28 7 2

Range = 26; Mean = 9; Median = 5; Mode = 2

6. Use the *Definition Formula* to compute the Standard Deviation:

12 10 11 8 9 11 9 = **1.3**

7. Use the *Computation Formula* to compute the Standard Deviation:

9 11 12 10 8 11 9 = **1.3**

8. Use the *Definition* **or** *Computation* Formula to find the standard deviation:

1 3 7 2 0 4 7 3 = **2.34**

9. Express each of the following scores as a z score:

   a) An IQ of 135, given a mean of 100 and a standard deviation of 15 = **2.33**

   b) A verbal score of 470 on a Scholastic Assessment Test (SAT), given a mean of 500 and a standard deviation of 100. **-3.3**

   c) A daily production of 2100 units, given a mean of 2180 units, and a standard deviation of 50 units. **-1.6**

   d) A height of 68 inches, given a mean of 68, and a standard deviation of 3. =**0**

   e) A meter-reading error of -3 degrees, given a mean of 0 degrees, and a standard deviation of 2 degrees. **=-1.5**

10. Use Table A to find the proportion of the total area identified with the following statements:

   a) Above a z score of 1.80 **=.0359**
   b) Between the mean and a z score of -0.43 **=-.1664**
   c) Below a z score of -3.00 **=-00005**
   d) Between the mean and a z score of 1.65 **= .4505**
   e) Above a z score of 0.60 **=.2743**
   f) Below a z score of -2.65 **=-.0040**
   g) Between a z score of 0 and -1.96 **=-.4750**

11. Employees of Corporation A earn annual salaries described below. Find the median salary and identify any outlier(s).

   $34,999 $134,999 $75,000 68,745 $86,745

   $62,888 **= $71,872.50**

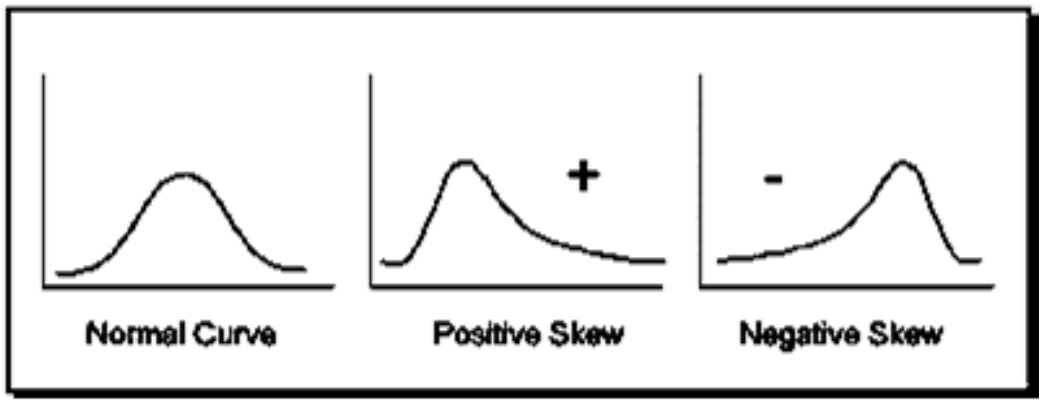
12. The two main subdivisions of statistics Descriptive and Inferential

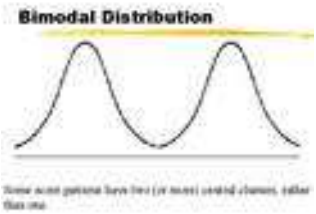
13. Researchers use three types of data:

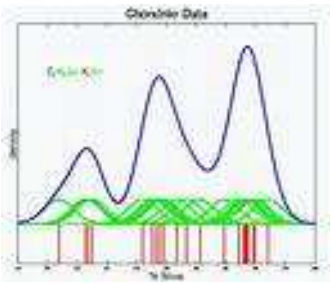
   a) QUALITATIVE_____
   b) QUANTITATIVE_____
   c) MIXED_____

14. Draw and label the following curves:

a) Standard normal curve
b) Positively skewed curve
c) Negatively skewed curve
d) Bi-modal curve
e) Multi-modal curve







STANDARD NORMAL CURVE

To identify a particular normal curve, you must know the (a) **mean** and (b) ...**standard deviation**... for that distribution. To convert a particular normal curve to the standard normal curve, you must convert original observations into (c) ...**z-score**. A z-score indicates how many (d) ...**standard deviations** an observation is (e) ...**above** or (f) ...**below** the mean of the distribution. Although there are infinite numbers of normal curves, there is (g) ...**one** standard normal curve. The standard normal curve has a (h)...**mean** of 0, and a (j) **standard deviation** of 1. The total area under the standard normal curve equals (j) ...**one**. When using the standard normal table, it is important to remember that for any z-score, the corresponding proportions in columns B and C always sum to (k) .**5000 or half**, furthermore, the proportion in column B always specifies the proportion of area between the (l) **mean** and the z-score, while the proportion in column C always specifies the proportion of area (m) **beyond** the z-score. Although any z-score can be either positive or negative, the proportions of area, specified in columns B and C, are never (n) **negative**.