AP Statistics – Unit 1 – Exploring and Understanding Data

Chapter 2 -DATA – What is it?

Categorical Quantitative

Variable:

Statistic:

The 5 (6) W's

Example:

The State Department of Education requires local school districts to keep records of all students: Age, race/ethnicity, days absent, current grade level, standardized test scores in reading and math, and any disabilities or needs the students may have.

Identify as many of the W's as you can, classify the variables and name the units.

Chapter 3: Displaying and Describing Categorical Data

Example 1: Smoking and Education
200 adults shopping at a supermarket were asked about the highest level of education they had completed and whether or not they smoke cigarettes. Results are summarized in the table.

	Smoker	Non-Smoker	Total
High School	32	61	93
2 Year College	5	17	22
4+ Year College	13	72	85
Total	50	150	200

A. Identify the variables
Categorical:

Quantitative:

B. Find each percent:

i. What percent of shoppers were smokers with only a high school education?

ii. What percent of shoppers with only a high school education were smokers?

iii. What percent of smokers had only a high school educations?

C. Do these data suggest there is an association between smoking and education level? Give statistical evidence to support your conclusion.
(Are smoking and education level independent?)

D. Does this indicate that students who start smoking while in high school tend to give up the habit if they complete college? Explain

E. Create a segmented bar graph comparing education level among smokers and non-smokers. Label your graph clearly.

F. What other types of graphical displays could you use instead of a segmented bar graph?

Example 2:Weather forecasts:
The following table compares the daily forecast with a city's actual weather for a year:

	Actual Weather		
Forecast		Rain	No Rain
	Rain	27	63
	No rain	7	268

A. On what percent of days did it actually rain?

B. On what percent of days was rain predicted?

C) What percent of the time was the forecast correct?

D) Do you see evidence of an association between the type of weather and the ability of forecasters to make an accurate prediction? Explain, including an appropriate graph.

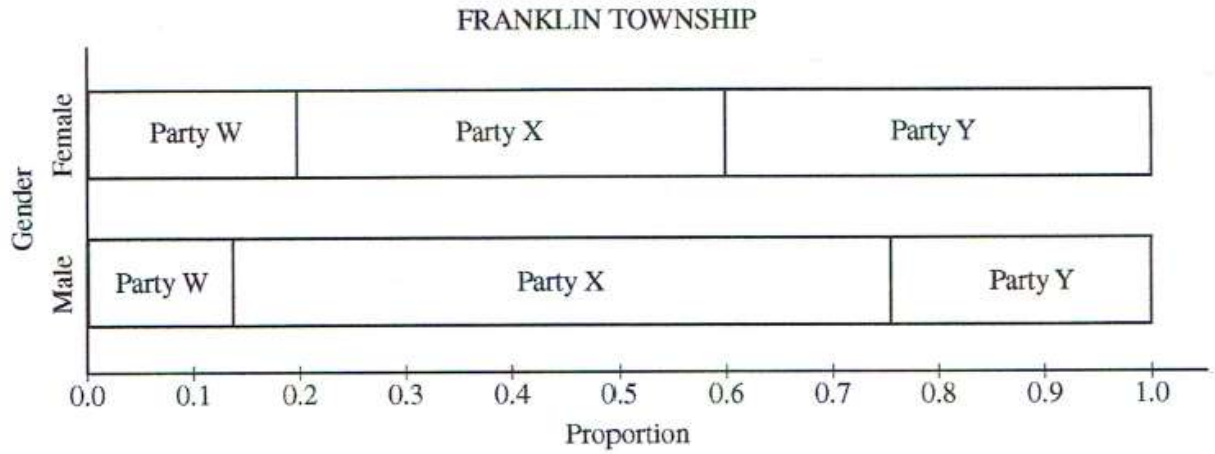
The table below shows the political party registration by gender of all 500 registered voters in Franklin Township

PARTY REGISTRATION–FRANKLIN TOWNSHIP

	Party W	Party X	Party Y	Total
Female	60	120	120	300
Male	28	124	48	200
Total	88	244	168	500

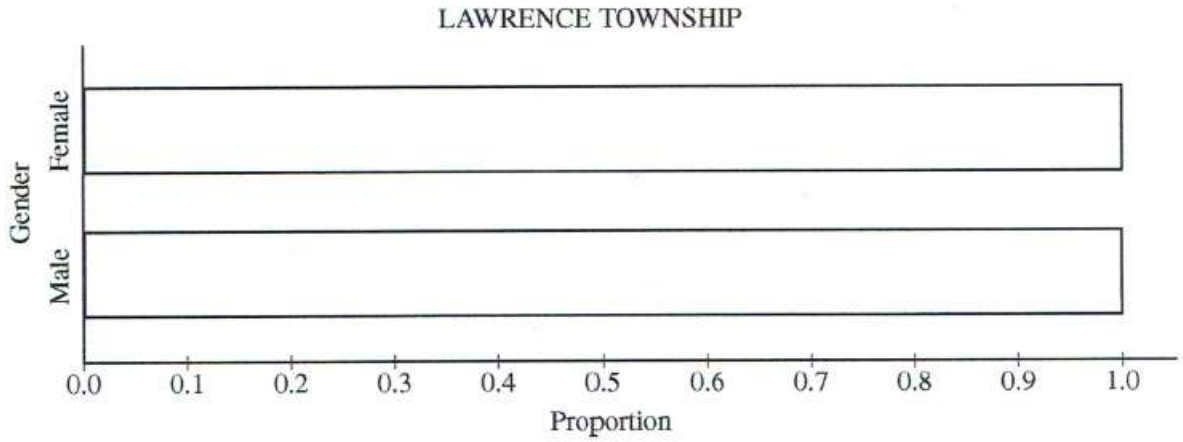
a) Given that a randomly selected registered voter is a male, what is the probability that he is a registered for Party Y?

b) One way to display the data in the table is to use a segmented bar graph. the following segmented bar graph, constructed from the data in the party registration – franklin township table, shows party registration distributions for males and females in Franklin township.



Is there evidence that gender and political party are independent? Explain.

c) In Lawrence Township, the proportion of all registered voters for Parties W, X, and Y are the same for Franklin Township, and party registration is independent of gender. Complete the graph below to show the distributions of party registration by gender in Lawrence Township.



Chapter 4 – Displaying Quantitative Data

Vocabulary:

mode	unimodal	bimodal	symmetric	skewed
outlier(s)	gaps	center	spread	variation

Example 1: In the Super Bowl, by how many points does the winning team outscore the losing team? here are the winning margins for the last 25 Super Bowl games:

8, 6, 14, 4, 35, 3, 4, 6, 14, 4, 3, 12, 11, 3, 3, 27, 3, 27, 7, 15, 7, 14, 10, 23, 17

a) Find the median

b) Find the Quartiles

c) Write a description based on the 5-number summary

d) Make a Stem-Leaf plot of the data:

Example 2:

A. Find the Mean and the Median:

Weight(kg)	45	50	55	60	65	70	75	80
Frequency	2	5	6	5	8	2	3	1

B. Estimate the Mean Weight

Weight (kg)	Frequency
$40 \le w < 50$	9
$50 \le w < 60$	15
$60 \le w < 70$	20
$70 \le w < 80$	11
$80 \le w < 90$	5

C. Determine which group contains the median

Weight (kg)	Frequency
$40 \le w < 50$	18
$50 \le w < 60$	14
$60 \le w < 70$	8
$70 \le w < 80$	4
$80 \le w < 90$	6

D. Draw a histogram, describe the distribution

Weight (kg)	Frequency
$40 \le w < 50$	12
$50 \le w < 60$	20
$60 \le w < 70$	18
$70 \le w < 80$	8
$80 \le w < 90$	6

Example 3: In March 2006, 16 gas stations in Grand Junction, CO, posted these prices for a gallon of regular gasoline:

2.22	2.21	2.45	2.24	2.27	2.28	2.27
2.23	2.26	2.46	2.29	2.32	2.36	2.38
2.33	2.27					

A. Use your graphing calculator to make a histogram of the data. draw a quick sketch

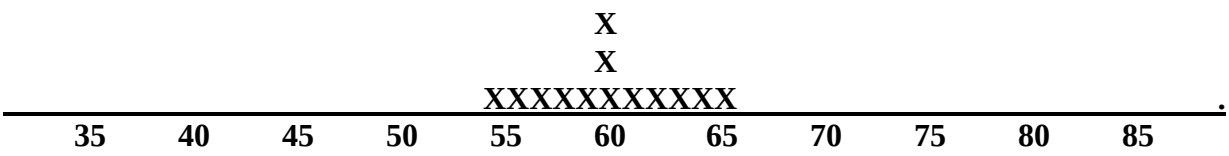
B. Describe the distribution (Shape, Center, Spread)

C. What unusual features do you see? (Gaps, clusters)

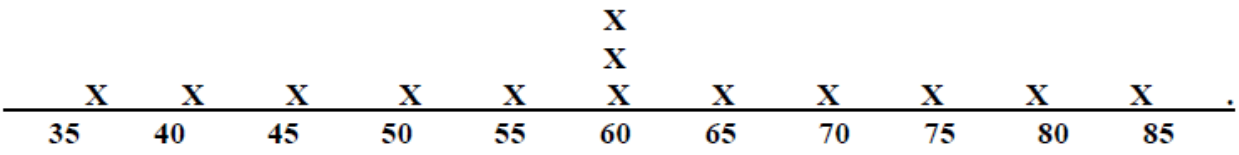
Example 4:

Find the mean. median, and mode of each distribution

List 1 : 55, 56, 57, 58, 59, 60, 60, 60, 61, 62, 63, 64, 65



List 2 : 35, 40, 45, 50, 55, 60, 60, 60, 65, 70, 75, 80, 85



Measures of Spread

Example 5

Find the Range, IQR, and Standard Deviation for the following:

A. 14, 13, 20, 22, 18, 19, 13

B. 4, 7, 7, 10

C.100, 140, 150, 160, 200

D. 10, 20, 35, 37, 50, 65, 150

Example 5: Would you expect distributions of these variables to be uniform, unimodal, or bimodal? Symmetric or skewed? What measure of center and spread would you use to describe the data? Explain.

A. Number of siblings of people in the class

B. Ages of people at a little league game

C. Pulse rate of college age males

D. Number of times each face of a dies shows in in 100 tosses.

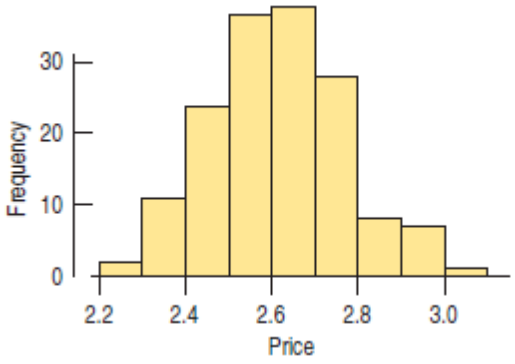
Measures of Center and Spread with Skewness:

Chapter 5 – Understanding and Comparing Distributions

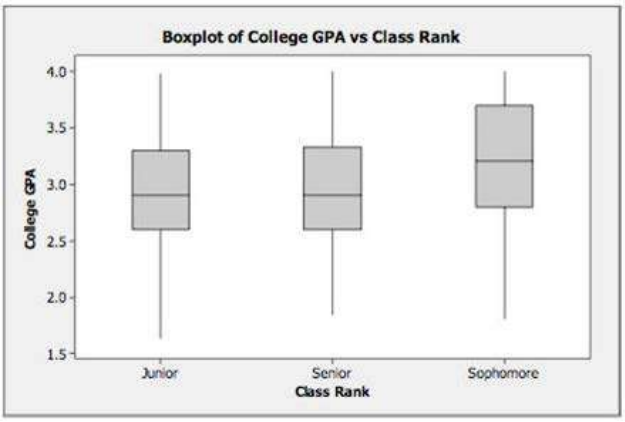
Describing Distributions:

CENTER	
Unusual Features	
SHAPE	
SPREAD	

Example 1: The histogram shows the distribution of the prices of plain pizza slices (in $) for 156 weeks in Dallas, TX. Describe the distribution



Example 2: The side-by-side boxplots show cumulative college GPAs for sophomores, juniors, and seniors taking an intro stats course.



A. Which class has the lowest cumulative college GPA? What is the approximate value of that GPA?

B. Which class has the highest median GPA and what is that GPA?

C. Which class has the largest range for GPA and what is it?

D. Which class has the most symmetric set of GPAs? The most skewed set of GPAs?

Example 3:

A survey conducted in the same college intro stats course asked students about the number of credit hours they were taking that quarter. The number of credit hours for a random sample of 16 students were as follows:

10, 10, 12, 14, 15, 15, 15, 15, 17, 17, 19, 20, 20, 20, 20, 22

A. Using your calculator create a histogram. draw a quick sketch.
Which statistic would you use to identify the center and spread of this distribution? Why?

B. Create a boxplot for this data.

C. Describe what the histogram shows that the boxplot doesn't and what the boxplot shows that the histogram doesn't.

D. Describe the distribution in a few sentences.

Example 4:
The body temperature of students is taken each time a student goes into the nurse's office. The five number summary for the temperatures (in F^o) of students on a particular day is:

Min	Q1	Median	Q3	Max
96.6	97.85	98.25	98.6	101.8

A. Would you expect the mean temperature of all students who have visited the nurse's office to be higher than or lower than the median? Explain.

B. Does this set of data contain any outliers?
- FENCE RULE FOR OUTLIERS

Example 5:
Do any of these data sets have outliers?

A. 7, 7, 2, 8, 5, 10, 9, 1, 10, 3

B. 6, 2, 6, 5, 3, 8, 5, 7, 15, 5

C.

Min	Q1	Median	Q3	Max
70	75	80	83	98

D.

Min	Q1	Median	Q3	Max
70	73	82	84	98

E.

Min	Q1	Median	Q3	Max
48	70	75	86	103

Example 6:
A student study of the effects of caffeine asked volunteers to take a memory test 2 hours after drinking soda. Some drank caffeine-free cola, some drank regular cola (with caffeine), and other drank a mixture of the two (getting a half-dose of caffeine.) here are the 5-number summaries for each group's scores (number of items recalled correctly) on the memory test:

	n	Min	Q1	Median	Q3	Max
no caffeine	15	16	20	21	24	26
Low caffeine	15	16	18	21	24	27
High Caffeine	15	12	17	19	22	24

A. Create parallel boxplots to display these results.

B. Write a few sentences comparing the performances of the three groups.

Chapter 6 – The Normal Model and Standard Deviation as a Ruler

Shifting and Rescaling Data:

The "Rules"
Let's take a simple data set: 2, 3, 4, 7, 9,

Find the measures of center:

Find the measures of spread:

a) Make a new data set by adding 6 to each score

Find the measures of center:

Find the measures of spread:

What happened?

b) Make a new data set by multiplying each original score by 4

Find the measures of center:

Find the measures of spread:

What happened?

What are the "rules" for rescaling data?

2. **Hotline.** A company's customer service hotline handles many calls relating to orders, refunds, and other issues. The company's records indicate that the median length of calls to the hotline is 4.4 minutes with an IQR of 2.3 minutes.
   a) If the company were to describe the duration of these calls in seconds instead of minutes, what would the median and IQR be?
   b) In an effort to speed up the customer service process, the company decides to streamline the series of push-button menus customers must navigate, cutting the time by 24 seconds. What will the median and IQR of the length of hotline calls become?

5. **SAT or ACT?** Each year thousands of high school students take either the SAT or the ACT, standardized tests used in the college admissions process. Combined SAT Math and Verbal scores go as high as 1600, while the maximum ACT composite score is 36. Since the two exams use very different scales, comparisons of performance are difficult. A convenient rule of thumb is $SAT = 40 \times ACT + 150$; that is, multiply an ACT score by 40 and add 150 points to estimate the equivalent SAT score. An admissions officer reported the following statistics about the ACT scores of 2355 students who applied to her college one year. Find the summaries of equivalent SAT scores.

Lowest score = 19 Mean = 27 Standard deviation = 3

Q3 = 30 Median = 28 IQR = 6

The Normal Model:

The Empirical Rule:

Examples: Adult American women follow a normal distribution with a mean height of 65 inches with a standard deviation of 3.5 inches. $N \sim (65, 3.5)$

1) What percent of American women are over 72 inches tall?

2) What proportion of American women are less than 61.5 inches tall?

3) What proportion of American women are between 58 inches and 68.5 inches tall?

4) Women who are 58 inches tall are at what percentile?

5) What percent of American women are taller than 60 inches?

The Wessler Intelligence Test is Normally distributed with a mean of 100 points and a standard deviation of 15.

1) What percent of people score below 110 on the Wessler Intelligence test?

2) What proportion of people score between 75 and 125 on the Wessler Intelligence test?

3) What proportion of people score above 125 on the Wessler Intelligence test?

4) What score is the 85^{th} percentile on the Wessler Intelligence test?

5) What score on the Wessler Intelligence test is at the 60^{th} percentile?

The mean composite score on the ACT is 20.8 with a standard deviation of 4.8 for college bound seniors

1) What percent of college bound seniors had an ACT score of above 25?

2) A college only accepts applicants that are above the 80^{th} percentile, what ACT score do you need to be considered for this college?

3) The mean ACT score for incoming freshman at UCLA is 29, what percentage of college bound seniors scored at least a 29?

3) The mean ACT score for incoming freshman at Cal State Fullerton is 22, what percentage of college bound seniors are below Cal State Fullerton's mean ACT entrance score?

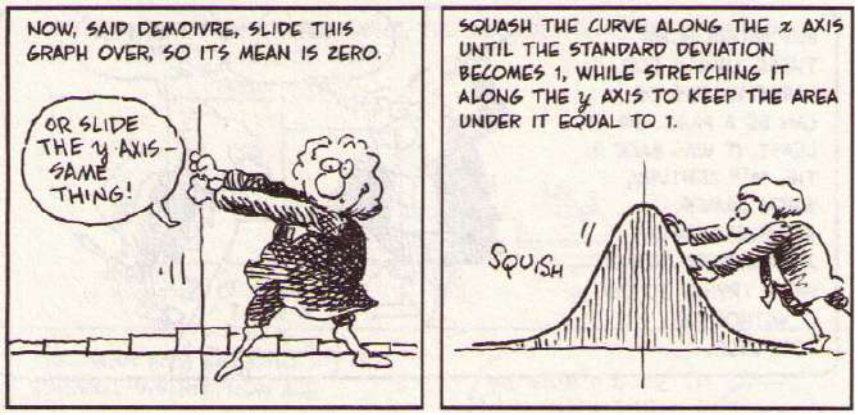
4) The mean SAT score for college bound seniors (without writing) was 1080 with a standard deviation of 194.
A college is looking at two students, Sandra and Camille, Sandra scored a 25 on the ACT and Camille scored a 1200 on the SAT. If otherwise their applications are equal, which student will have the higher admission ranking based on her test score?

Who is the best athlete?

	Event		
Competitor	**100 m Dash**	**Shot Put**	**Long Jump**
A	10.1 sec	66’	26’
B	9.9 sec	60’	27’
C	10.3 sec	63’	27’3”
Mean	10 sec	60’	26’
St Dev	0.2 sec	3’	6”

The z-score:

What does it measure?



What does standardization do?

1) The mean composite score on the ACT is 20.8 with a standard deviation of 4.8 for college bound seniors. The mean SAT score for college bound seniors (without writing) was 1080 with a standard deviation of 194. A college is looking at two students, Sandra and Camille, Sandra scored a 25 on the ACT and Camille scored a 1200 on the SAT. If otherwise their applications are equal, which student will have the higher admission ranking based on her test score?

More Examples:

1)

Stats test. Suppose your Statistics professor reports test grades as z-scores, and you got a score of 2.20 on an exam. Write a sentence explaining what that means.

2)

Stats test, part II. The mean score on the Stats exam was 75 points with a standard deviation of 5 points, and Gregor's z-score was -2. How many points did he score?

3)

Placement exams. An incoming freshman took her college's placement exams in French and mathematics. In French, she scored 82 and in math 86. The overall results on the French exam had a mean of 72 and a standard deviation of 8, while the mean math score was 68, with a standard deviation of 12. On which exam did she do better compared with the other freshmen?

4)

Cattle. The Virginia Cooperative Extension reports that the mean weight of yearling Angus steers is 1152 pounds. Suppose that weights of all such animals can be described by a Normal model with a standard deviation of 84 pounds.

a) How many standard deviations from the mean would a steer weighing 1000 pounds be?

b) Which would be more unusual, a steer weighing 1000 pounds or one weighing 1250 pounds?

5)

Cattle, part III. Suppose the auctioneer in Exercise 19 sold a herd of cattle whose minimum weight was 980 pounds, median was 1140 pounds, standard deviation 84 pounds, and IQR 102 pounds. They sold for 40 cents a pound, and the auctioneer took a $20 commission on each animal. Then, for example, a steer weighing 1100 pounds would net the owner $0.40\,(1100) - 20 = \$420$. Find the minimum, median, standard deviation, and IQR of the net sale prices.

6)

Small steer. In Exercise 17 we suggested the model $N(1152, 84)$ for weights in pounds of yearling Angus steers. What weight would you consider to be unusually low for such an animal? Explain.

7)

Normal cattle. Using $N(1152, 84)$, the Normal model for weights of Angus steers in Exercise 17, what percent of steers weigh

a) over 1250 pounds?

b) under 1200 pounds?

c) between 1000 and 1100 pounds?

8)

More cattle. Based on the model $N(1152, 84)$ describing Angus steer weights, what are the cutoff values for

a) the highest 10% of the weights?
b) the lowest 20% of the weights?
c) the middle 40% of the weights?

9)

Cattle, finis. Consider the Angus weights model $N(1152, 84)$ one last time.

a) What weight represents the 40th percentile?
b) What weight represents the 99th percentile?
c) What's the IQR of the weights of these Angus steers?

10)

Tires. A tire manufacturer believes that the treadlife of its snow tires can be described by a Normal model with a mean of 32,000 miles and standard deviation of 2500 miles.

a) If you buy a set of these tires, would it be reasonable for you to hope they'll last 40,000 miles? Explain.

b) Approximately what fraction of these tires can be expected to last less than 30,000 miles?

c) Approximately what fraction of these tires can be expected to last between 30,000 and 35,000 miles?

d) Estimate the IQR of the treadlives.

e) In planning a marketing strategy, a local tire dealer wants to offer a refund to any customer whose tires fail to last a certain number of miles. However, the dealer does not want to take too big a risk. If the dealer is willing to give refunds to no more than 1 of every 25 customers, for what mileage can he guarantee these tires to last?

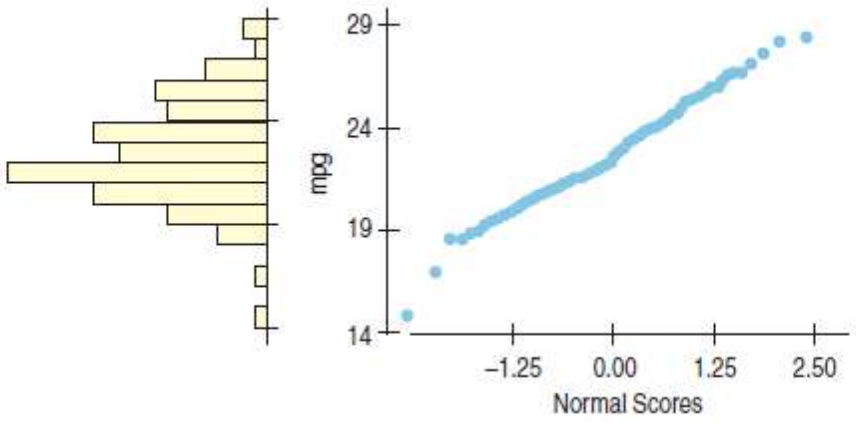
Are you Normal?

How can we check for Normality?

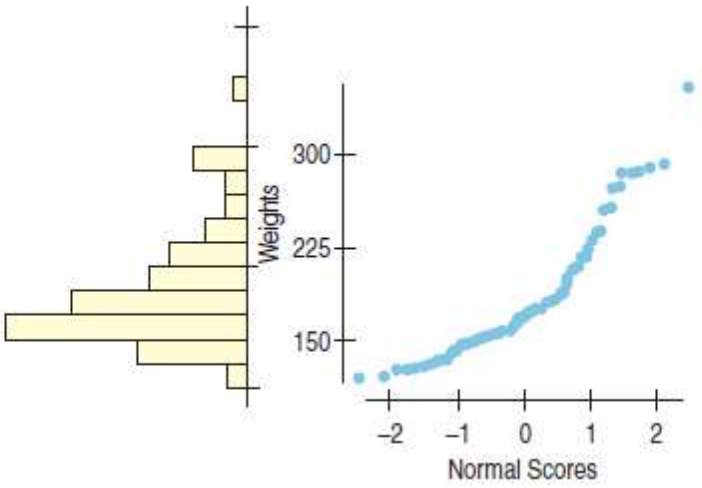
a) The Empirical Rule

b) Normal Probability Plot

Nearly Normal:



Skewed Data:



. **Cramming.** One Thursday, researchers gave students enrolled in a section of basic Spanish a set of 50 new vocabulary words to memorize. On Friday the students took a vocabulary test. When they returned to class the following Monday, they were retested—without advance warning. Both sets of test scores for the 28 students are shown below.

Fri	Mon	Fri	Mon
42	36	50	47
44	44	34	34
45	46	38	31
48	38	43	40
44	40	39	41
43	38	46	32
41	37	37	36
35	31	40	31
43	32	41	32
48	37	48	39
43	41	37	31
45	32	36	41
47	44		

Are the differences of the two scores Normal?