Unit 6: Statistics

Objectives:

1.02b Summarize and analyze univariate data to solve problems. Apply statistical principles and methods in sample surveys.

1.02c Summarize and analyze univariate data to solve problems. Determine measures of central tendency and spread.

1.02d Summarize and analyze univariate data to solve problems. Recognize, define, and use the normal distribution curve.

1.02e Summarize and analyze univariate data to solve problems. Interpret graphical displays of univariate data.

1.02f Summarize and analyze univariate data to solve problems. Compare distributions of univariate data.

1.03c Use theoretical and experimental probability to model and solve problems. Create and use simulations for probability models.

DAY	TOPIC	ACTIVITY
1	Central Tendency Displays of data (Stem and Leaf Plots, Frequency Tables, Histograms)	
2	Central Tendency from a Histogram, Frequency Table, Distribution of Data, Calculator 5 Number Summary, Box and Whisker Plot, How to find outliers, Advantages/Disadvantages of Each Method	Pulse Rate Activity
3	Standard Deviation Variance	Ramp investigation
4	Distribution of Data, Normal Distribution Confidence Intervals Z-Score	**Quiz days 1-3**
5	Population Sample Sampling Methods Limitations/Bias	
6	Review!	
7	**TEST**	**TEST**

Measures of Central Tendency

- mean –
- median –
- mode –
- range –
- interquartile range –
- five-number summary -
- random sample –
- measures of central tendency –
- Types of Data:
    - Univariate Data
    - Quantitaive Data
    - Categorical Data

Practice:
Determine whether the following variables are categorical (C) or quantitative (Q)
1. Brand of vehicle purchased by a customer
2. Price of a CD
3. Type of M&Ms preferred by students (peanut, plain)
4. Phone number of each student
5. Height of a 1-year old child
6. Term paper status (turned in on time or turned in late)
7. Gender of the next baby born at a particular hospital.
8. Amount of fluid (oz) dispensed by a machine used to fill bottles with soda
9. Thickness of the gelatin coating on a Vitamin C capsule
10. Brand of computer purchased by a customer
11. State that a person is born in
12. Price of a textbook

Example

Owen is a member of the student council and wants to present data about backpack safety to the school board. He collects these data on the weights of backpacks of 20 randomly chosen students. How much does the typical backpack weigh at Owen's school?

Student	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Grade	Jr	Sr	Sr	Jr	Jr	Sr	Sr	Sr	Sr	Jr	Jr	Sr	Jr	Sr	Sr	Jr	Sr	Sr	Sr	Jr
Weight of Backpack (lb)	10	19	20	21	7	9	12	11	13	4	33	15	18	21	22	8	9	3	12	16

Use the above data to find the following measures.

a. Mean:

b. Median:

c. Mode:

d. Range:

e. Five-number summary:

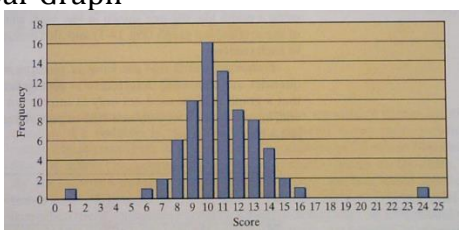
What would make the data in Owen's study unfair, or biased?

How could Owen insure that he had a good representation of the entire population?

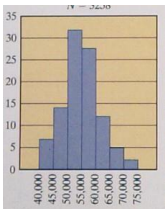
Histograms

A **histogram** is a graphical representation of a data set, with columns to show how the data are distributed across different intervals of values. The columns of a histogram are called **bins** and should not be confused with the bars of a bar graph. Bar graphs represent categories, while histograms measure data in certain intervals.

Bar Graph



Histogram



In a histogram, the height of each bin represents the frequency, or the number of students whose height falls in that interval. The width of each bin represents an interval, in this case each interval is 5,000. Boundary values fall in the bin to the right. So the first bin would be $40{,}000 \le x < 45{,}000$ and the next bin would be $45{,}000 \le x < 50{,}000$.

Let's make a histogram to represent the data that Owen collected (see top of page 3).

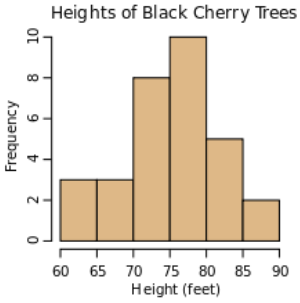
- Pros:

- Cons:

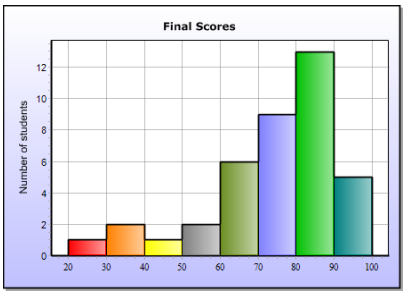
Example A

For each of the following histograms, give the bin width and the number of values in the data set. Then identify the bin that contains the median of the data.

a.



b.

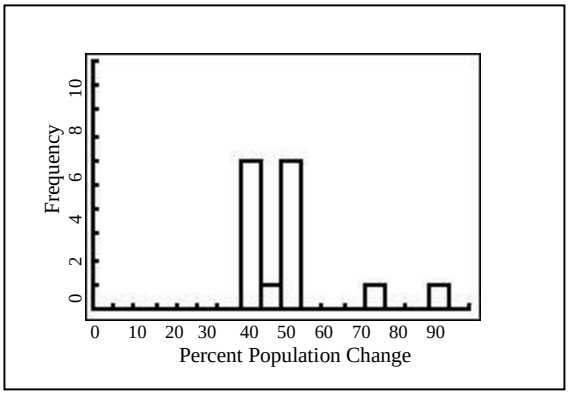


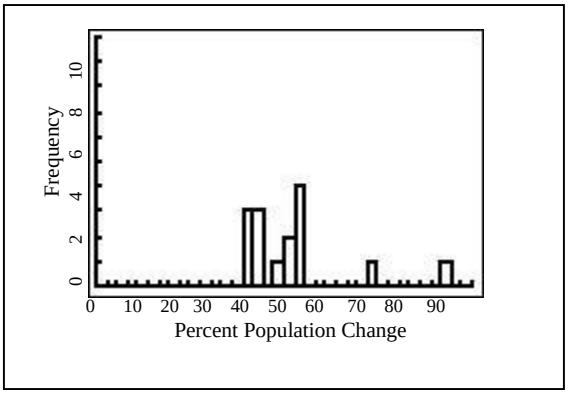
The **percentile rank** of a data value in a large distribution gives the percentage of data values that are below the given value. For example, if you are in the 95th percentile on your PSAT, you have done better than 95% of the other students your age that took that test.

Example B

The following histograms were both constructed with the data below.

Metropolitan Area	Percent Population Change (2000 – 1990)
Las Vegas, NV	83.3
Naples, FL	65.3
Yuma, AZ	49.7
McAllen, TX	48.5
Austin, TX	47.7
Fayetteville, AR	47.5
Boise City, ID	46.1
Phoenix, AZ	45.3
Laredo, TX	44.9
Provo, UT	39.8
Atlanta, GA	38.9
Raleigh, NC	38.9
Myrtle Beach, SC	38.9
Wilmington, NC	36.3
Fort Collins, CO	35.1





a. What is the range of the data?

b. What is the bin width of each graph?

c. Use the information in the table to create the same graphs on your calculator.

d. How can you know if the graph accounts for all 25 metropolitan areas?

e. Why are the columns shorter in Graph B?

Frequency Tables: A chart used to show the amount of times an event occurs in a data set. A summary of a histogram. Create a frequency table for Owen's data.

Handspan (cm)	Frequency

- Pros:
- Cons:

Stem and Leaf Plots are created much like histograms but they retain original data values. These plots have two parts:

Leaf: Represents the last digit of each number regardless of whether it falls before or after a decimal point.

Stem: Represents the other digits of each number. Stems should be in increasing order

It is important to ***ALWAYS have a key so viewers can read the plot.

Create a Stem and Leaf Plot for Owen's data (page 3)

Key:

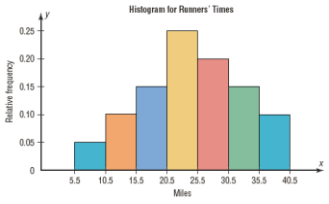
You can create a Stem and Leaf plot for separate sets of data. This is called a "Back to Back" Stem and Leaf Plot. Let's separate Owen's data (page 3) into a back to back stem and leaf plot separating Juniors and Seniors.

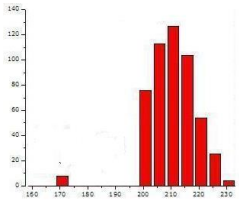
- Pros:
- Cons

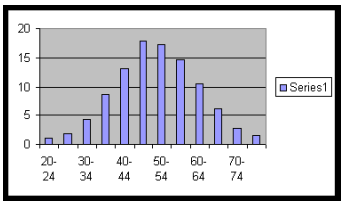
The following vocabulary words can be used to describe graphical displays

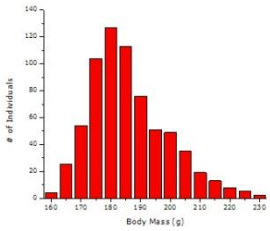
Uniform Each bin has approximately the same height	**Gaps** Spaces between data points	**Uni-Modal** One bin has the highest value	**Bi-Modal** Two bins tie for the highest value
Multi-Modal There are more than two ties for the highest bin	**Outliers** Extreme values that don't appear to belong with the rest of the data	**Symmetric** The two halves look like approximate mirror images	**Normal** Looks like a hill with the highest peak near the middle
Long Tails The edges slowly drop off	**Short Tails** The edges drop off quickly	**Skewed Left** The longer tail reaches to the left	**Skewed Right** The longer tail reaches to the right

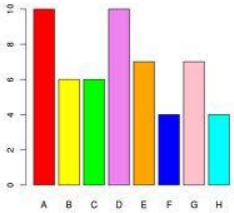
Use as many of these vocabulary words to describe the following displays

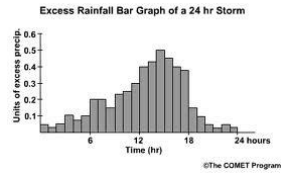
1. 

3. 

5. 

2. 

4. 

6. 

Mean, Median, Mode…Which One?

- **Skewed data or data with outliers: Median**
- **Continuous and Symmetrical: Mean**
- **Categorical (nominal) Data: Mode**

Practice: Mean – Median – Mode – Range

In Exercise 1-4, order the data from least to greatest using your graphing calculator. Then find the mean, median, mode and range of the data.

1. Number of inches of rain that fell on 14 towns in a 50 mile radius during a three day period: 8, 4, 7, 6, 5, 6, 7, 8, 9, 10, 11, 5, 4, 8

2. Cost of admission to a ballgame at 20 different stadiums: $4.25, $3.75, $5.00, $5.25, $4.00, $4.50, $5.00, $3.75, $5.25, $6.25, $5.75, $6.00, $5.50, $5.75, $6.25, $6.50, $7.00, $6.25, $6.50, $6.25.

3. Number of states 20 people have visited.: 5, 15, 2, 10, 30, 26, 2, 3, 20, 22, 14, 48, 18, 10, 8, 9, 12, 40, 15, 15.

4. Number of students in 25 different 11^{th} grade classes: 12, 17, 13, 5, 7, 20, 24, 18, 20, 21, 14, 18, 19, 8, 13, 25, 20, 21, 4, 10, 20, 21, 16, 14, 20.

5. The table shows the number of nations represented in the Summer Olympic Games from 1960 through 2004. Find the mean, median, mode and range of the data. Which do you think best represents the data? Explain.

Year	Nations
1960	83
1964	93
1968	112
1972	121
1976	92
1980	80
1984	140
1988	159
1992	169
1996	197
2000	199
2004	201

Practice: Histograms, Stem-and-Leaf and Frequency Tables

Create a frequency table, histogram, and stem and leaf plot using the given information. Then describe the graphs of the data.

1. Number of crimes committed in 1984

January	124	February	96	March	86
April	113	May	107	June	102
July	85	August	87	September	91
October	119	November	122	December	115

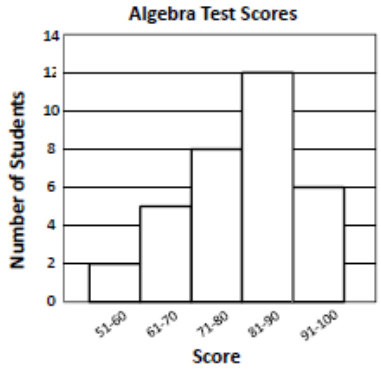
Interval	Frequency
80-90	
90-100	
100-110	
110-120	
120-130	

2. Test scores for a high school biology test

81, 77, 63, 92, 97, 68, 72, 88, 78, 96, 85, 70, 66, 95, 80, 99, 63, 58, 83, 93, 75, 89, 94, 92, 85, 76, 90, 87

Interval	Frequency
50-60	
60-70	
70-80	
80-90	
90-100	

1. The histogram below that shows data about scores on a history test.

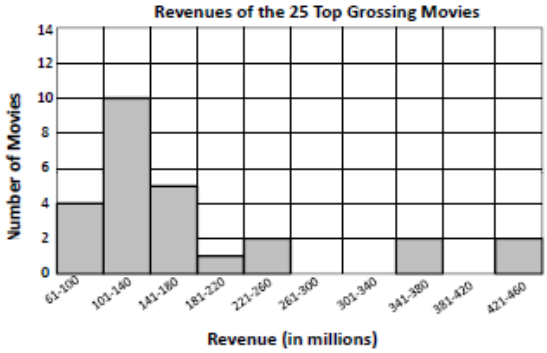


a. How many total students took the test?

b. How many students scored at least a 71 on the test?

c. Can you determine the highest grade from the histogram?

2. The histogram below shows data about movie revenues in a recent year.



a. How many movies grossed at least $141 million?

b. How many movies grossed between $61 million and $180 million?

c. Can you determine how many movies grossed between $121 and $140 million from the histogram?

Practice: Central Tendency

1. Which central tendency is most affected by extreme values?

2. Five workers on an assembly line have hourly wages of $8.00, $8.00, $8.50, $10.50, and $12.00. If the hourly wage of the highest paid worker is raised to $20 per hour, how are the mean, median and mode affected? Explain.

3. Is the mean of a group of numbers always, sometimes or never a number in the group? Explain.

4. Roger Maris's regular-season home run totals for his eleven year career are 14, 28, 16, 39, 61, 33, 23, 26, 13, 9, 5. Find the mean, median, and mode. How representative of the data is the mean? Explain.

5. A statistician was entering Roger Maris's data from #4 above into a spreadsheet. The statistician made a small error and instead of entering the 11th number as 5, she accidentally entered the number 50. Explain how this error will affect the median and mean of Roger Maris's data.

6. Suppose your mean on 4 math tests is 78. What score would raise the mean to 80?

7. The median height of the 21 players on a girls' soccer team is 5 ft 7 in. What is the greatest possible number of girls who are less than 5 ft 7 in? Suppose three girls are 5 ft 7 in tall. How would this change your answer to the first part of this question?

Please put your graphical displays and answers on another sheet of paper.

8. Below is the average number of runs scored in American League and National League stadiums for the first half of the 2001 season.

AMERICAN

11.1	10.8	10.3
10.3	10.1	10.0
9.5	9.4	9.3
9.2	9.2	9.0
8.3		

NATIONAL

14.0	11.6	10.4
10.3	10.2	9.5
9.5	9.5	9.5
9.1	8.8	8.4
8.3	8.2	8.1
7.9		

a) Create a back to back stem and leaf plot of this data. Be sure to label it and give it a key.
b) Create histograms for both groups. Be sure to label it!
c) Calculate the mean, median and mode for each league.
d) Write a brief summary comparing the average number of run scored per game in the two leagues.
e) Which central tendency best represents the American League data? Explain.
f) Which central tendency best represents the National League data? Explain.

Day 2: Central Tendency from a Histogram

To find the mean, median and mode from a histogram, you first need to know how many data points were used.

Use the frequency table and find the total frequency.

Type of Pet	Tally	Frequency
Dog	𝍸 𝍸 ||	12
Cat	𝍸 ||	7
Goldfish	𝍸 |	6
Budgie	|||	3
Hamster	||	2
Lizard	|	1
Snake	|	1
Rabbit	|||	3

Mode:

But can we be exact?

Median:

Mean:

Using a calculator to find mean, median and mode from a frequency table or histogram:

In the STAT list enter "category" into L1 and frequency into L2 (if the category is a range, take the midpoint)

Press STAT ▸ CALC ▸ #1 (1-var stats) ▸ L1, L2

$\bar{x}$ = mean

MED= median

Example 1:

Numbers of Advertising Spots

TABLE 2-6: Number of Advertising Spots Purchased during 2001 by Members of the Toronto Automobile Dealers' Association

96	93	88	117	127	95	113	96	108	94
148	156	139	142	94	107	125	155	155	103
112	127	117	120	112	135	132	111	125	104
106	139	134	119	97	89	118	136	125	143
120	103	113	124	138					

Create a histogram on your calculator. Then copy the graph. Then create a frequency table and stem and leaf plot for the data. Calculate the mean, median and mode from the frequency table/histogram

Example 2: The prices of seven race cars sold last week are listed in the table below.

Price per Race Car	Number of Race Cars
$126,000	1
$140,000	2
$180,000	1
$400,000	2
$819,000	1

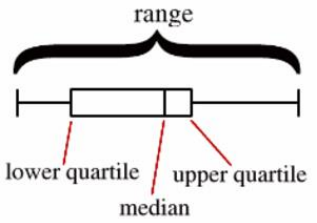
a) What is the mean value of these race cars, in dollars?

b) What is the median value of these race cars, in dollars?

c) State which of these measures of central tendency best represents the value of the seven race cars. Justify your answer.

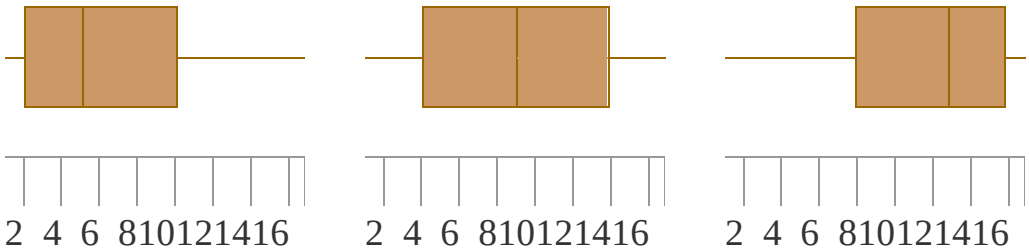
Box and Whisker Plots

A **box plot** (or **box-and-whisker plot**) provides a visual tool for analyzing information about a data set.



Boxplots display two common measures of the variability or spread in a data set; _____ and _____.

Boxplots often provide information about the shape of a data set. The examples below show some common patterns.



Outliers: A data point that is distinctly separate from the rest of the data. One definition of outlier is any data point more than 1.5 interquartile ranges (IQRs) below the first quartile or above the third quartile.

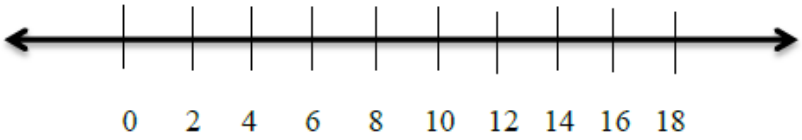
Finding Outliers:

- Compute the IQR and multiply it _____
-
-

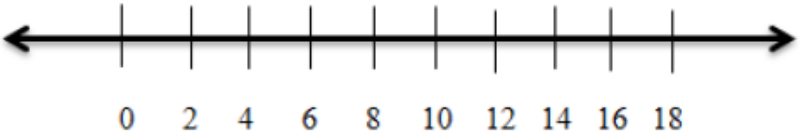
To denote an outlier on a box-and-whisker plot, use an asterisk (*)

How to Create a Box and Whisker Plot

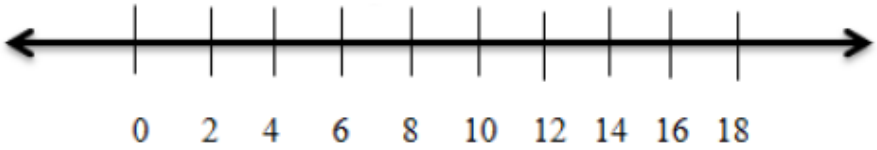
1. Before drawing a box plot, you must start with a set of data that you find the 5-number summary for. (Minimum, Q1, Median, Q3, and Maximum)
    a. Let's use this set: {5, 7, 9, 6, 13, 1}
    b. Always put the vales in order first! {1, 5, 6, 7, 9, 13}
    c. Now you can find the 5-number summary:
        i. Minimum _____
        ii. Q1 _____
        iii. Median _____
        iv. Q3 _____
        v. Maximum _____
2. Begin with a number line that will fit your data, and fill in enough numbers so others will know what scale you are using.



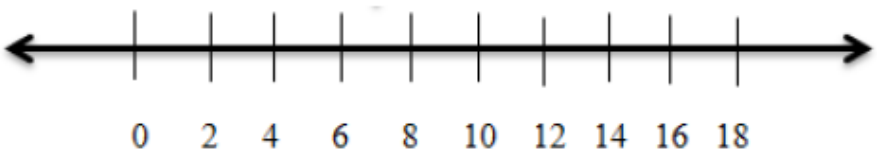
3. Draw 5 dots above the number line to represent the 5-number summary.



4. Next, draw a box around the first – third quartiles and separate using the median.



5. Connect the minimum and maximum to the box plot by drawing straight lines (the whiskers). This demonstrates the range of the data.



Example 1 Draw a Box-and-Whisker Plot

The following is a list of speeds of 12 of the fastest animals.

Animal	Speed (mph)
Cheetah	70
Pronghorn antelope	61
Wildebeest	50
Lion	50
Thomson's gazelle	50
Quarter horse	47.5
Elk	45
Cape hunting dog	45
Coyote	43
Gray fox	42
Hyena	40
Zebra	40
Source: *The World Almanac*	

Example 2 Draw Parallel Box-and-Whisker Plots

The following table shows the SAT mean verbal and math scores of college-bound seniors for several western states in 2001.

SAT Mean Verbal and Math Scores		
State	**Verbal**	**Math**
Arizona	523	525
California	498	517
Colorado	539	542
Idaho	543	542
Montana	539	539
Nevada	509	515
New Mexico	551	542
Oregon	526	526
Utah	575	570
Washington	527	527
Wyoming	547	545

Source: *The World Almanac*

Investigation: Pulse Rates

Pulse rate is often used as a measure of whether or not a person is in good physical condition. In this investigation you will practice making box plots, comparing box plots, and drawing conclusions about pulse rates using your displays.

Step 1
Find your pulse (either in your neck or on your wrist). After you hear, 'Go!', start counting the number of heart beats you feel until you hear, 'Stop!' (15 seconds). Multiply that number by 4 to get the number of heart beats per minute. Record your number below. This is your resting heart rate

Step 2
Write the data collected from the entire class below.

Step 3
For two minutes, we are going to dance/jog/jump/move around. Don't be too cool for school. You need to move around for the two full minutes.

Step 4
Find your pulse again. We are going to do Step 1 again. This is your active heart rate.

Step 5
Write the data collected from the entire class below.

Step 6
Calculate the five-number summary for the class' resting pulse rates.

Step 7
Calculate the five-number summary for the class' active pulse rates.

Step 8
Using the five-number summaries, construct a box plot to represent the data. Put both box plots on the same number line, one above the other. Make sure to label the axis with the appropriate intervals to contain your data.

Step 9
Draw two conclusions about pulse rates by comparing the two box plots. Be sure to compare not only centers, but also spreads and shapes.

Step 10
If you were a physician, how could you use the above data and conclusions to inform someone of their health?

Is the data we collected in class enough information to come to conclusions about anyone?

Day 2: Classwork

1. The chart below shows the frequency of runs batted in (RBI) by the American League batting leaders between 1907 and 1991.

RBI	Frequency
70-90	2
90-110	11
110-130	39
130-150	17
150-170	9
170-190	7

a) Find the mean, median and mode from the frequency table. (Be sure to use the middle number of the interval for calculations)

b) In 1991, the RBI champion in the American League was Cecil Fielder of the Detroit Tigers with 133 RBI's. Write a sentence to compare this number with the mean of the data.

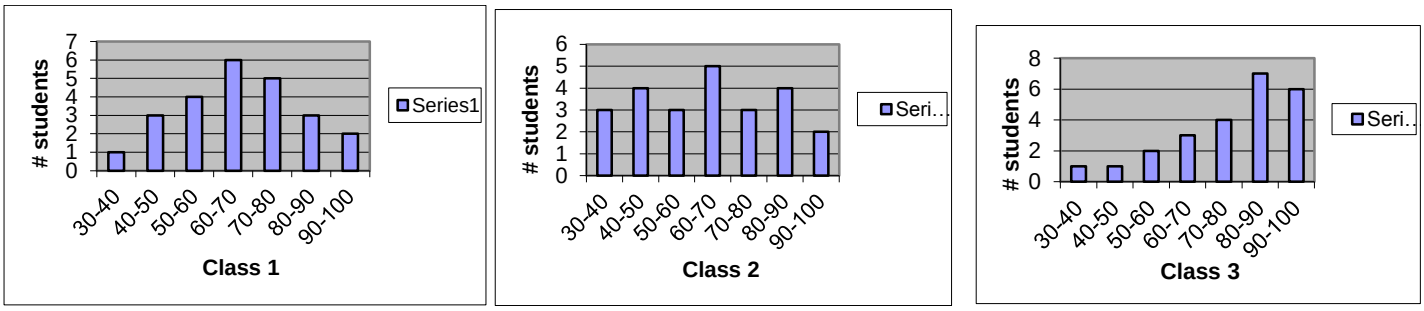
2. A small warehouse employs a supervisor at $1200 a week, an inventory manager at $700 a week, six stock boys at $400 a week, and four drivers at $500 a week.

a) Find the mean, median and mode wage.

b) How many employees earn more than the mean wage?

c) Which measure of center best describes a typical wage at this company, the mean or the median? Explain.

3. Three statistics classes all took the same test. Histograms of the scores for each class are shown below.



a) Which class had the highest mean score?

b) Which class had the highest median score?

c) For which classes are the mean and median most different? Which is higher? Why?

d) Describe the shape of *each* graph.

e) Does there appear to be any gaps or outliers in any of the classes? If so, which ones? Explain.

f) Which class did better on the test overall? Explain.

Refer to the precipitation data below that shows the normal monthly precipitation, in inches, from 1961 to 1990 for the following cities.

Month	Chicago	L.A.
Jan	1.53	2.40
Feb	1.36	2.51
March	2.69	1.98
April	3.64	0.72
May	3.32	0.14
June	3.78	0.03
July	3.66	0.01
Aug	4.22	0.15
Sept	3.82	0.31
Oct	2.41	0.34
Nov	2.92	1.76
Dec	2.47	1.66

4. Make a box – and – whisker plot for the amount of participation in each city. Use one number line to see the comparison between the two cities. Don't forget to label the five-number summary for each city. (You may want to use .5 as benchmarks on your number line, for ex: 0, 0.5, 1, 1.5, …)

5. How much precipitation would have had to fall in a given month to be considered an outlier in our given data? Find out for each city.

6. Describe the distribution of your data for each city (spread, gaps, outliers and shape of the graph)

7. Find the mean, median and mode for each city.

8. Compare the precipitation in each city, using their box – and –whisker plots and part b. Explain in complete sentences.

9. If next year, Chicago receives an extra inch of rain each month, how would that effect the five-number summary and the three measures of central tendency?

Standard Deviation and Variance Investigation

Step 1: Build a ramp from your books and notebooks so that a balled up piece of paper will roll off of your desk and onto the floor. Select the height and slope of your ramp and the distance from the edge of your desk, and determine any other factors that might affect your results.

Step 2: Ball up a piece of paper into a paper ball. Starting at the top of your ramp, let the paper ball go. One group member should be on the floor with the yard stick. Record the point where the ball hits the ground. Then record where the ball is when it stops rolling. Fill in the table below.

ROLL NUMBER	WHERE PAPER BALL HITS GROUND	WHERE PAPER BALL STOPS	DISTANCE PAPER BALL COVERED (Column 3 – Column 2)	DEVIATION	DEVIATION^2
1					
2					
3					
4					
5					
6					
7					
8					

Step 3: Calculate the mean distance for your trials.

Step 4: On average how much do your data values differ from the mean? Use deviation to calculate one value to represent this answer. (Deviation: value in column 4 – mean) Find the average of all of the deviations.

Step 5: What do you think caused the deviation from the mean?

Step 6: Square each of your deviations and record those values in the sixth column of the table.

Step 7: Find the **variance**. Add all of the values in the sixth column of the table and divide by one less than the total number of values.

Step 8: Find the **standard deviation**. To do that, take the square root of your variance.

Step 9: Using your calculator, find the standard deviation. Are they the same? If not, what could cause the difference?

Step 10: Write the formula for finding the standard deviation. (Hint: You just found the standard deviation. Put all of the steps together and find a way to write one formula to combine them all.)

Measures of Spread

- spread –
- deviation (or directed distance) –
- variance –
- standard deviation –
- range –
- interquartile range -
- outliers –

Measuring the variability of a data set allows a more complete description than just stating a measure of central tendency.

Variance

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

To do the calculations shown in the formula above:

1. Find the difference between each data point and the mean.
2. Square each of those values.
3. Add them all together.
4. Divide by 1 less than the number of total data points.

Standard Deviation

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

To do the calculations shown in the formula above:

1. Find the variance.
2. Take the square root of the variance

Example 1

14.1, 15.6, 17.9, 21.4, 31.7, 13.2, 15.8

a. Find the mean of the above data.

b. Find the deviation from the mean.

c. Find the variance.

d. Find the standard deviation.

Example 2

3.5, 6.7, 12.1, 43.5, 51.7, 23.1, 67.8, 31.2

a. Find the range of the above data.

b. Find the IQR of the above data.

c. Use the IQR to determine if there are any outliers in the data set.

Example 3

12, 14, 54, 35, 71, 1, 98, 22, 41

a. Find the outliers in the above data set.

Calculator Notes:

Under, STAT→Calc→1-Var-Satats

Comparing Cars

With all of the talk lately about fuel efficient cars, it's important to see car companies live up to their hype. Here are the top five fuel efficient cars. Each car was test driven 5 times on full tanks of gas to see how many miles per gallon they got. The results are as follows

Hyundai Elantra Predicted - 33	*Honda Civic* Predicted - 30	*Mazda 3* Predicted - 28	*Ford Focus* Predicted - 32	*Toyota Corolla* Predicted - 30
31.2	28.8	28.3	33.2	31.2
29.9	31.7	27.9	31.7	29.3
36.2	30.9	28.5	34.1	30.8
34.5	27.6	27.6	32.4	28.9
30.7	29.5	29.0	32.0	31.4

1.) Calculate the mean for each car (to the nearest hundredth)

2.) Calculate the variance for each car (to the nearest tenth)

3.) Calculate the standard deviation for each car (to the nearest tenth)

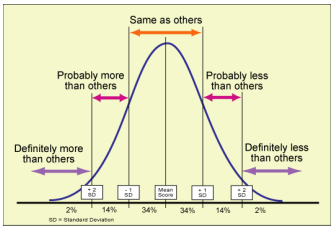
	Mean	Variance	Stnd. Dev.
Hyundai Elantra			
Honda Civic			
Mazda 3			
Ford Focus			
Toyota Corolla			

Now that you have the variance and deviation of each car, answer the following questions

4.) Which car company was the most consistent with miles per gallon? How did you know?

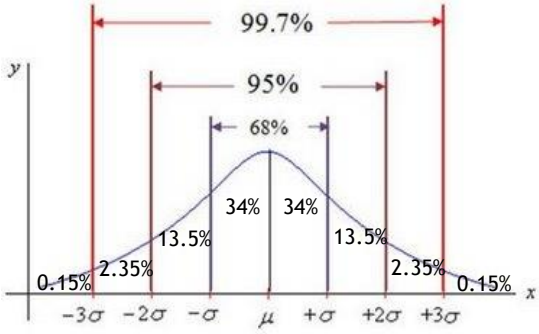
5.) Would you use the mean or the standard deviation if you were trying to determine if the car companies were close to their predicted miles per gallon? Explain your reasoning

Normal Distribution



Distributions for large populations often have a bell-shaped distribution. Heights, clothing sizes, and test scores are a few examples. In fact, the bell-shaped curve is so common that it is called a _____, and a bell-shaped distribution is called a _____

Most graphing calculators provide the normal distribution equation as a built-in function and you have to provide only the mean (μ - mu) and standard deviation (σ – lowercase sigma).



The standard **confidence intervals** are as labeled in the above picture. The mean (μ) is in the very center of a normal curve.

_____% of the data fall within one standard deviation (σ) to the right and left of the mean.

_____% of the data fall within two standard deviations of the mean.

_____% of the data fall within three standard deviations of the mean.

Example 1

A group of students weighs 500 U.S. pennies. They find that the pennies have normally distributed weights with a mean of 3.1g and a standard deviation of 0.14g.

a. Create a normal distribution curve for the data.

b. What is the percentage of pennies will weigh between 3.2 and 3.4 g?
c. What is the percentage of pennies will weigh more than 3.3 g?
d. How many pennies weigh less than 3g?
e. What is the probability that the weight of a penny will be within one standard deviation of the mean? Two standard deviations of the mean? Three standard deviations of the mean?

Example 2

A class has test scores that are normally distributed with a mean of 82 and a standard deviation of 5. Give the percentage of all data values that fall within each interval.

a. Within two standard deviations of the mean

b. Between the mean and two standard deviations above the mean

c. Below the mean

Z-Scores

The number of standard deviations from the mean are called _____. If the value is one standard deviation above the mean, it will have a z-value of 1. If a value is two standard deviations below the mean, it has a z-value of -2.

[z-score] [Finding Data Point Using z-score]

$z = \frac{x-\mu}{\sigma}$ $data\ point = \mu + \sigma(z)$

Example 3

The scores on a standardized test are normally distributed with a mean of 71 and a standard deviation of 3.8. Find the test scores for each of these z-values.

a. $z = 2$ b. $z = -3$

c. $z = 1.5$ d. $z = -2.8$

Find the z-score of the following test scores.

e. 80.5 f. 90

Example 4

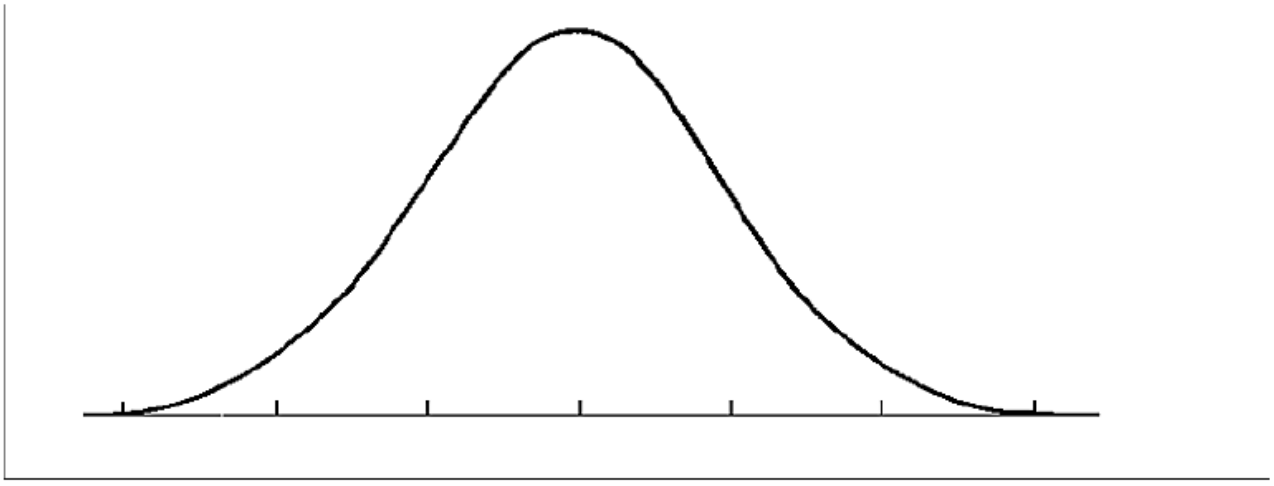
What is the probability that a test score from the above situation will be between 67.2 and 78.6?

Empirical Rule

In a normal distribution, what percent of the values lie:

1. below the mean? _____
2. above the mean? _____
3. within one standard deviation of the mean? _____
4. within two standard deviations of the mean? _____
5. within three standard deviations of the mean? _____

6. **2000 freshmen at State University took a biology test. The scores were distributed normally with a mean of 70 and a standard deviation of 5. Label the mean and three standard deviations from the mean.**



Answer the following questions based on the data:

a) What percentage of scores are between scores 65 and 75?

b) What percentage of scores are between scores 60 and 70?

c) What percentage of scores are between scores 60 and 85?

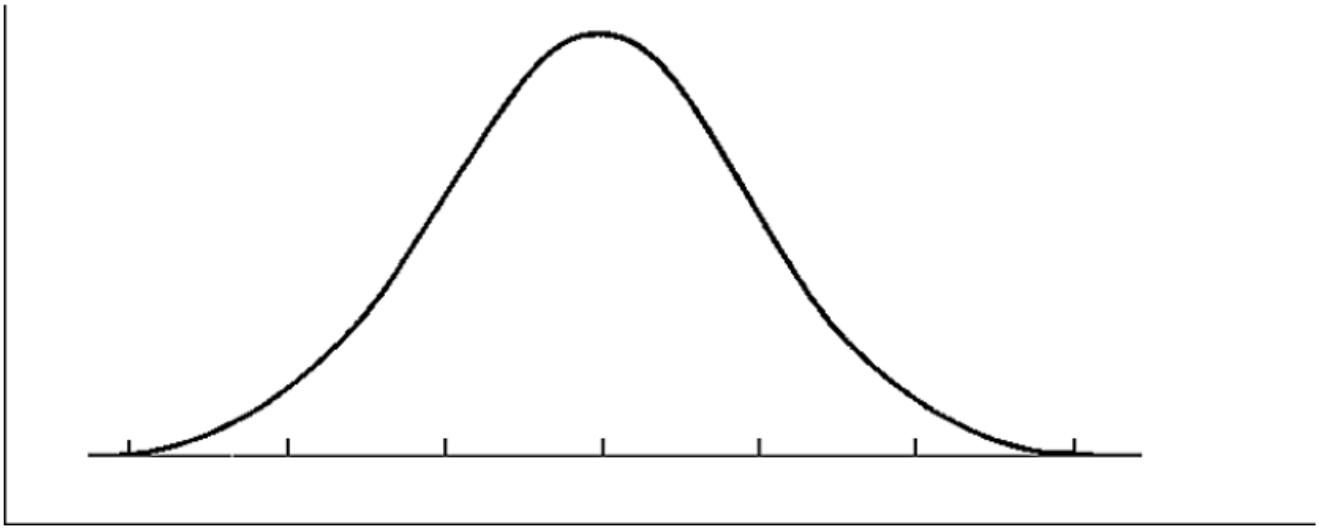
d) What percentage of scores is less than a score of 55?

e) What percentage of scores is greater than a score of 80?

f) Approximately how many biology students scored between 60 and 70?

g) Approximately how many biology students scored between 55 and 60?

7. **500 juniors at Central High School took the ACT last year. The scores were distributed normally with a mean of 24 and a standard deviation of 4. Label the mean and three standard deviations from the mean.**



Answer the following questions based on the data:

a) What percentage of scores are between scores 20 and 28?

b) What percentage of scores are between scores 16 and 32?

c) What percentage of scores are between scores 16 and 28?

d) What percentage of scores is less than a score of 12?

e) What percentage of scores is greater than a score of 24?

f) Approximately how many juniors scored between 24 and 28?

g) Approximately how many juniors scored between 20 and 28?

h) Approximately how many juniors scored between 24 and 32?

i) Approximately how many juniors scored between 16 and 20?

j) Approximately how many juniors scored higher than 32?

Standard Deviation

1. The sum of the deviations about the mean always equals _____.

2. Given the following numbers, find the standard deviation:
   43, 26, 92, 11, 8, 49, 52, 126, 86, 42, 63, 78, 91, 79, 86

3. a) Compute the standard deviation of the following test scores: 78, 78, 78, 78, 78.
   b) What can be said about a data set in which all the values are identical?

4. The mean and standard deviation of a data set are mean = 10 and standard deviation = 4. Given the following numbers in the set, they are *within* how many standard deviations from the mean?
   a) 14 b) 18 c) 20 d) 8 e) 10

5. You are filling out an application for college. The application requests either your ACT or SAT Math score. You scored 26 on the ACT composite and 650 on the SAT Math. On the ACT exam, the composite mean score is 21 with a standard deviation of 5, while the SAT Math has a mean score of 514 with a standard deviation of 113. Which test should you provide on the application? Explain your reasoning.

6. Suppose the frequency of each data item in the table below is doubled. What is the effect, if any, on the mean and standard deviation of the data?

Item of Data	3	6	7
Frequency	4	10	6

7. The cost of groceries when parents go shopping is normally distributed. Of the 26 different parents' visits this year, the mean amount of money they spent was $196 with a standard deviation of $12.
   a) Make the curve to represent the normal distribution.

   b) Of the parents surveyed, how many spent more than $220?
   c) What percentage of the parents spent less than $184?

Population and Sampling

- The entire set of individuals or objects in which we are interested in is the **population**.
- Subset of population is a **sample** and the number of objects in a sample is called a **sample size**.
- The process of selecting a sample that is representative of the total population is called **sampling**.

We need to ask ourselves:

1) How should the sample be collected?
2) How large is our sample size?
3) How reliable are our conclusions?

Example: Senior citizens are 20% of Littleton's voting population. In a poll of 100 citizens, half of whom were senior citizens, 30 senior citizens voted yes and 20 non-senior citizens voted yes. What is the population, sample and sample size?

Sampling Methods: The goal in sampling is to obtain individuals that will participate in a study so that accurate information about the population can be obtained. We want the sample to provide as much information as possible, but each additional piece of information has a price. So the question becomes "How can a researcher obtain accurate information about the population through the sample while minimizing the costs in terms of money, time, personnel, and so on?"

When doing a survey, it usually is not practical to get the opinion of every member of a population. You can get a fairly accurate picture of the opinion of a population by surveying a sample of the population. A sample is a smaller group that represents the whole population. There are five good ways to choose a sample: **Simple random sampling, stratified sampling, systematic sampling, cluster sampling and convenience sampling.**

1. Simple random sampling (often abbreviated as random sampling):
Every possible sample has an equally likely chance of occurring.
Example: Sophia has 4 tickets to a concert. Six of her friends, Yolando, Michael, Kevin, Terri, Annie, and Casey, have all expressed an interest in going to the concert. Sophia decides to *randomly* select three of the six. (There are 20 subsets of size 3, YMK, YMT, YMA, YMC, YKT, YKA, MKT, …, TAC)
How do we actually select the individuals in a simple random sample? Simple random sample is just like drawing the names out of a hat. We could write the six names on different sheets of paper and then select three from the hat. It's that easy! But, often the size of the population is so large that performing simple random sampling in this fashion is not practical. Each person could be assigned a number and then you can use a calculator or computer to randomly select the number you need in your sample. Each person is equally likely to be chosen.
THINK ABOUT:

2. Stratified sampling:
A stratified sample is obtained by separating the population into non-overlapping groups called strata and then obtaining a simple random sample from each stratum (or group). The individuals in each stratum should be similar in some way. An advantage of stratified sampling over simple random sampling is that it may allow fewer individuals to surveyed while obtaining the same (or more) information. This occurs because individuals within each subgroup have similar characteristics, so opinions within the group do not vary much from one person to the next.
In other words, a stratified sample is a simple random sample of different divisions of the population.
THINK ABOUT:

3. Systematic Sampling:
A systematic sample is obtained by selecting every nth individual from the population. The first person selected is a random number between 1 and n, and then survey every nth person after that random number. For example, you want to survey every 8th person. Randomly choose a number between 1 and 8, such as 5. This means you survey the 5th, 5+8 = 13th, 13+8 = 21st, 21+8 = 29th and so on.
In other words, systematic sampling is like selecting every 5th person out of a line.
THINK ABOUT:

4. Cluster Sampling:
A cluster sample is obtained by selecting all individuals within a randomly selected collection or group of individuals.
For example, a quality control engineer wants to verify that a certain machine is filling bottles with 16 ounces of liquid detergent. To obtain a sample of bottles from the machine, the engineer could use systematic sampling by sampling every nth bottle from the machine; however, it would be time consuming waiting next to the filling machine for the bottles to come off the line. Instead, suppose that as the bottles come off the line, they are placed into cartons of 12 bottles each. Then the engineer could randomly select a few cartons and measure the contents of all 12 bottles. This would be cluster sampling. It is good in this situation because it speeds up the data collection process.
In other words, imagine a mall parking lot. Each subsection of the lot could be a cluster - Section F-4, for example.
THINK ABOUT:

5. Convenience Sampling: (Sometimes called "Self-selection")
A convenience sample is a sample in which the individuals are easily obtained. The most popular convenience sample is one in which the individuals in the sample are self-selected (the individuals themselves decide to participate in a survey).
Examples: 1) a radio DJ asks his/her listeners to phone the station to submit their opinions 2) Dateline will present a story on a certain topic and ask its viewers to "tell us what you think" by going on-line to complete a questionnaire. CAUTION: Convenience sampling is generally not a good design because the individuals who decide to be in the sample generally have strong opinions about the topic. A typical individual in the population will not bother phoning or logging on to their computer to complete a survey. Therefore, convenience sampling has limitations or is biased.
THINK ABOUT:

Errors in sampling that cause bias:
1) nonresponse of individuals selected to be in the survey
2) inaccurate responses
3) poorly worded questions ("Do you oppose the reduction of estate taxes?" would be better if written as "Do you favor or oppose the reduction of estate taxes?") The question should be balanced.
4) bias in the selection of the individuals

Example: A news program reports on a proposed school dress code. The purpose of the program is to find out what percent of the population in its viewing area favors the dress code. Identify the type of sampling and any bias in each sampling method.

a) Viewers are invited to call the program and express their preferences.
b) A reporter interviews people on the street near the local high school.
c) During the program, 300 people are selected at random from the viewing area. Then each person is contacted.

Population and Sampling Practice

1. An important part of employee compensation is a benefits package, which might include health insurance, life insurance, child care, vacation days, retirement plan, parental leave, bonuses, etc. Suppose you want to conduct a survey of benefits packages available in private businesses in Hawaii. You want a sample size of 100. Some sampling techniques are described below. Categorize each technique as a simple random sample, stratified sample, systematic sample, cluster sample or convenience sample.

a. Assign each business in the Island Business Directory a number, and then use a random number table to select the businesses to be included in the sample.

b. Use postal ZIP codes to divide the state into regions. Pick a random sample of 10 ZIP code areas and then include all the businesses in each selected ZIP code area.

c. Send a team of five research assistants to Bishop Street in downtown Honolulu. Let each assistant select a block or building and interview an employee from each business found. Each researcher can have the rest of the day off after getting responses from 20 different businesses.

d. Use the Island Business Directory. Number all the businesses. Select a starting place at random, then use every 50th business listed until you have 100 businesses.

e. Group the businesses according to type: medical, shipping, retail, manufacturing, financial, construction, restaurant, hotel, tourism, other. Then select a random sample of 10 businesses from each business type.

2. Modern Managed Hospitals (MMH) is a nation for-profit chain of hospitals. Management wants to survey patients discharged this past year to obtain patient satisfaction profiles. They wish to use a sample of such patients. Categorize each technique as a simple random sample, stratified sample, systematic sample, cluster sample or convenience sample.

a. Obtain a list of patients discharged from all MMH facilities. Divide the patients according to length of hospital stay (2 days or less, 3 – 7 days, 8 – 14 days, more than 14 days). Draw simple random samples from each group.

b. Obtain a list of patients discharged from all MMH facilities. Number these patients, and then use a random number table to obtain the sample.

c. Randomly select some MMH facilities from each of five geographic regions, and then include all the patients on the discharge lists of the selected hospitals.

d. At the beginning of the year, instruct each MMH facility to survey every 500th patient discharged.

e. Instruct each MMH facility to survey 10 discharged patients this week and send in the results.

Determine the population and sample, if possible, then determine the sampling used. Are there any errors in the sampling that may cause bias? Explain.

3. An interviewer in a mall is told to survey every 5^{th} shopper, starting with the 2^{nd}.

4. A researcher randomly selects 5 of the 70 hospitals in a metropolitan area and then surveys all of the surgical doctors in each hospital.

5. A researcher segments the population of car owners into four groups: Ford, General Motors, Chrysler, and foreign. She obtains a random sample from each group and conducts a survey.

6. A list of students in elementary statistics is obtained in which the individuals are numbered 1 to 540. A professor randomly selects 30 of the students.

7. In order to estimate the percentage of defects in a recent manufacturing batch, a quality control manager at Intel selects every 8^{th} chip that comes off the assembly line starting with the 3^{rd} chip, until she obtains a sample of 140 chips.

8. In order to determine the average IQ of ninth-grade students, a school psychologist obtains a list of all high schools in the local school system. She randomly selects five of these schools and administers an IQ test to all 9^{th} grade students at the selected schools.

9. In an effort to determine customer satisfaction, United Airlines randomly selects 50 flights during a certain week and surveys all passengers on the flights.

10. In an effort to identify whether an advertising campaign has been effective, a marketing firm conducts a nation-wide poll by randomly selecting individuals from a list of known users of the product.

11. A school official divides the student population into four classes: freshman, sophomore, junior, senior. The official takes a random sample from each class and asks the members' opinions regarding student services.

12. A survey regarding download time on a certain web site is administered on the internet by a market research firm to anyone who would like to take it.

13. A lobby group has a list of the 100 senators of the United States. In order to determine the Senate's position regarding farm subsidies, they decide to talk with every seventh senator on the list starting with the third.

14. A manufacturing company would like to determine the approximate market share of a certain product. A representative of the company is asked to stand in front of a certain grocery store and ask the first 100 people who go into the store whether they use their product.

Unit 2: Univariate Data Review

Sampling: Determine which type of sampling method was used in each survey.

1. To get a sense of election outcomes, a political group chooses ten precincts to conduct a survey of voters in those areas.

2. A company is taking a survey of its employees and separates them into the following groups: Male/Part-Time, Male/Full-Time, Female/Part-Time, Female/Full-Time.

3. A researcher wants to randomly select certain classes, then interview every student in only those classes.

4. A group of students in a high school do a study about teacher attitudes. They interview teachers at the school.

5. A researcher wants to select ten students for a survey. Each student's name is placed in a hat and 10 names are selected.

6. A researcher wants to sample eight houses from a street of 120 houses. Every 15th house is beginning with house #11. The houses selected are 11, 26, 41, 56, 71, 86, 101, and 116.

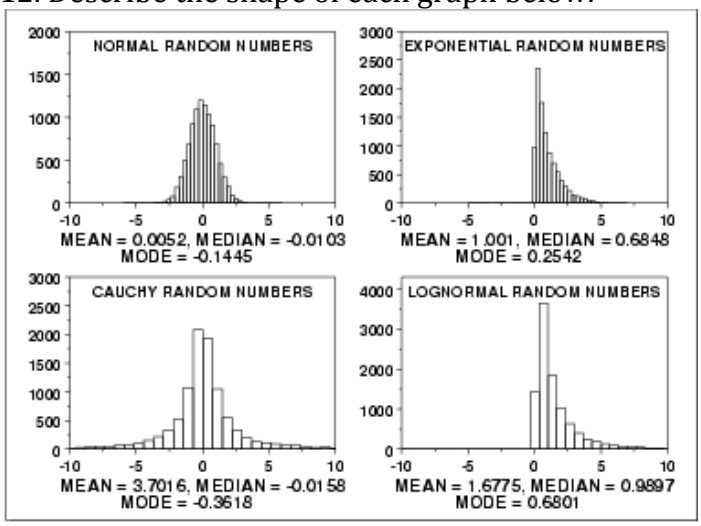
7. The researcher stands at a shopping mall and selects the first 75 shoppers as they walk by to fill out a survey.

8. To determine the average milk yield of each cow type in his herd, a farmer divides his herd into four sup-groups and takes samples from each group.

9. All senior's names are placed into a fishbowl and 5 names are drawn to complete a college survey.

10. A researcher selects 15 households from each zip code in the Houston area.

12. Describe the shape of each graph below.



13. The heights (in inches) of 30 adult males are listed below.
70 72 71 70 69 73 69 68 70 71
67 71 70 74 69 68 71 71 71 72
69 71 68 67 73 74 70 71 69 68
Construct a frequency table, stem-and-leaf, histogram, and compute the 5 number summary, and finds the standard deviation and variance. Then describe the distribution of the data.

14. SAT verbal scores are normally distributed with a mean of 489 and a standard deviation of 93. Use the Empirical Rule (also called 68-95-99.7 Rule) to determine what percentage of the scores lie:
a) between 303 and 582.
b) above 675?
c) If 3,500 students took the SAT verbal test, about how many received between 396 and 675 points?

15. The scores of the top ten finishers in a recent golf tournament:
71 67 67 72 76 72 73 68 72 72.
Suppose the players increase their games by 5 points. How will the measures of central tendency be changed?

16. Approximate the mean, median and mode of the grouped data:

Heights of Males	**Frequency**
63-65	3
66-68	6
69-71	7
72-74	4
75-77	3

17. A random sample of the age of employees in a City Hall:

Age	**frequency**
20-29	5
30-39	10
40-49	12
50-59	8
60-69	5

What percentage of the City Hall employees are between 31.8 and 68.4 years old?
If there are 120 employees in a City Hall, approximately how many of them are:
a) between 31.8 and 56.2 years old?
b) older than 68.4?

18. Which data set has a) highest mean and b) standard deviation

i)

0	9
1	5 8
2	3 3 7 7
3	2 5
4	1

ii)

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

iii)

10	9
11	5 8
12	3 3 7 7
13	2 5
14	1

19. The Laboratory of Ornithology holds an annual Christmas Bird Count, in which birdwatchers at various locations around the country see how many different species of birds they can spot. Here are some of the counts reported from sites in Texas during the 1999 event.

228	178	186	162	206	166	163	183	181	206	177
175	167	162	160	160	157	156	153	153	152	

a. Create a stem and leaf display of these data.
b. Create a histogram of this data.
c. Find the 5 number summary.
d. State the IQR. What does this information tell you about the number of birds sighted?
e. Write a brief description of the distribution of the data.
f. Considering the data collected, what count would be considered an outlier? Are there any outliers? If we took the outlier out, how would this affect our five number summary?
g. Calculate the mean, median and mode. Which central tendency best represents the data? Explain.
h. If each person said they counted one less than they had previously stated, how would this affect the mean, five number summary, and standard deviation (if at all)?
i. Calculate the standard deviation. 225 is w/in how many standard deviations from the mean? 163 is w/in how many standard deviations from the mean?
j. 68% of the data falls between _____, 95% of the data falls between _____, and 99% of the data falls between _____.
k. The percentage that spotted over 232 birds is?

20. A grading scale is set up for 1000 students' test scores. It is assumed that the scores are normally distributed with a mean score of 75 and a standard deviation of 15.

a. Construct a normal distribution curve.
b. How many students will have scores between 45 and 75?
c. If 60 is the lowest passing score, how many students are expected to pass the test?

21. Given the frequency table, find the following:

Score	#Students
60-70	2
70-80	8
80-90	11
90-100	6

a. The mean, median, mode, and total number of students in the class.
b. If the teacher decided to give everyone a 5 point curve, how would that affect the mean and standard deviation (if at all)?
c. If a student made up a test and made a 62, how would that affect the five number summary?

22. Explanation problems.

a. When comparing data, how would you know which collection had more variation among its data?
b. Look back over ALL the homework problems and your quizzes!

23. Susan's test scores in biology are shown below.

73, 84, 91, 68, 83

A. Which of the following measures would be the best for her report card?
   a. mean
   b. median
   c. range
   d. mode

B. Which of the following statements is true?
   a. The mean is greater than the median.
   b. The range is greater than 50.
   c. The median is between 80 and 90.
   d. The mode is 23

24. The average monthly precipitation (in inches) in Richmond,Virginia, for the months of January through April is listed below.

3.24 3.16 3.61 2.96

A. If the precipitation in May is 3.84 inches, which of the following measures will remain unchanged?
   a. mean
   b. median
   c. mode
   d. range

B. With May's precipitation added to the data set, which of the following statistical measures will change the most?
   a. mean
   b. median
   c. mode
   d. range

25. Justine calculated the median price for six new cars. The prices she used are listed below.

$19,580 $24,987 $26,594 $10,876 $12,235 $19,699

If she suddenly realizes that the price of the fourth car was supposed to be $11,876, what would the effect on the median be if she recalculated it?

A. The median price decreases by $1,000.
B. The median price increases by $166.66.
C. The median price increases by $1,000.
D. The median price increases by $0.