**Day 1 Notes**

**Univariate Data Notes**

1. Central Tendency: The measure of the middle or center of distribution (Central value, what single number represents a set of data)

Ex: Mean, median, mode

* 1. **Mean**: Average of data

Add all the data and divide by total number of data

Used in quantitative data

* 1. **Median**: Exactly midway in set of numbers

To find: Rank numbers low to high and find the middle

If there is an even number of data points, rank the numbers and find the two middle numbers. Add them together and divide by 2

* 1. **Mode**: The number that occurs most often (or numbers that occur most often)

Mostly used for categorical data

1. Types of Data:
   1. Univariate Data: 1 variable data (This is what we’re studying THIS unit)
   2. Quantitative data: numbers a quantity (ex. height, weight, age, etc)
   3. Categorical data: group names (categories)

* Each individual falls into a category (Ex: Freshmen, sophomore, junior, senior; male, female)

Concerned with how many or what percent fall into each category

* 1. Continuous Data: every value within the same range is possible
     + - Ex: height

1. Spread – how much the data values vary from each other, how “spread out”
   1. **Range**: The highest value minus the lowest value
   2. **Interquartile Range**: 3rd quartile – 1st quartile (will do later)
   3. **Standard deviation**: Will do in calculator, measures how far data varies from the MEAN.

**AFM Name \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Univariate Data Date \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Categorical vs. Quantitative Data**

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

**13.** Zip code of each student in this class

**14.** Actual weight of coffee in a one pound can

**15.** Length of a rattlesnake

**Histograms**

**Histograms** organize data into “bins”. A histogram looks like a bar graph but the bars are touching since the data is *continuous*. The height of a histogram is often called the “count” for each bin or the percentage of items in that bin. The “count” represents the frequency of each bin.

**Frequency Tables**: A chart used to show the amount of times an event occurs in a data set. A summary of a histogram.

**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.

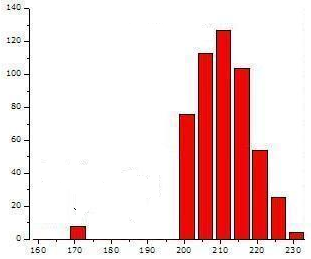
You can create a Stem and Leaf plot for separate sets of data. This is called a **“Back to Back” Stem and Leaf**

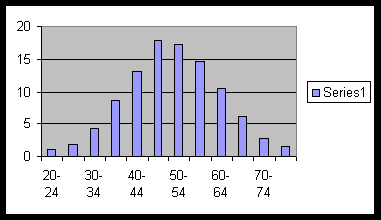
The following vocabulary words can be used to describe graphical displays

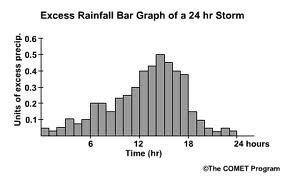
|  |  |  |  |
| --- | --- | --- | --- |
| **Uniform**  Each bin has approximately the same height | **Gaps**  Spaces between data points | **Uni-Modal**  One bin has the highest value | **Bi-Modal**  Two bins tie for the highest value |
| **Multi-Modal**  There are more than two ties for the highest 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 |

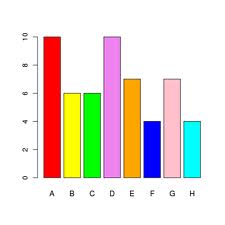
Use as many of these vocabulary words to describe the following displays

1. 
2. 







1. 

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Mean – Median – Mode – Range

In Exercise 1-4, order the data from least to greatest. 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

1. 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.

1. 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.
2. Number of students in 25 different 11th 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.
3. A baseball team scored the following number of runs in its games this season: 6, 2, 5, 9, 11, 4, 5, 8, 6, 7, 5. There is one more game in the season. If the team wants to end the season with an average of at least 6 runs per game, what is the least number of runs the team must score in the final game of the season?
4. 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 |

**Day 1 Classwork Cont’d**

**Histogram Worksheet**

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 |
| 80-89 |  |
| 90-99 |  |
| 100-109 |  |
| 110-119 |  |
| 120-129 |  |

1. 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 |
| 60-69 |  |
| 70-79 |  |
| 80-89 |  |
| 90-99 |  |

1. The following list is the age of Oscar-Winning Best Actors and Actresses.” Create a back top back stem and leaf plot. Use bins of 10 years.

Actors: 32 37 36 32 51 53 33 61 35 45 55 39 76 37 42 40 32 60 38 56 48 48 40 43 62 43 42 44 41 56 39 46 31 47 45 60  
  
Actresses: 50 44 35 80 26 28 41 21 61 38 49 33 74 30 33 41 31 35 41 42 37 26 34 34 35 26 61 60 34 24 30 37 31 27 39 34

4. Construct a frequency table given the following histogram of SAT scores. (Recall center intervals)



**AFM Unit 2 Day 1 Hwk** Central Tendency Name \_\_\_\_\_\_\_\_\_\_\_\_

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 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 |  |  |

|  |  |  |
| --- | --- | --- |
| 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 |  |  |

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

**Day 1 Homework Cont’d**

9. Construct a frequency table given the following histogram of SAT scores. (Recall center intervals)



10. Could you find the mean, median and mode if you just had the histogram from #2 and not all the actual scores? Explain.

11. Which central tendency must be a member of the set of data? Explain.

**Day 2 Notes**

**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 add up the total frequency.

Mode: This is obvious. Look at the histogram to see the bin with the highest frequency.

But can we be exact? No, because bins have a range of numbers so we cannot know exactly what the mode is.

Median: Divide the total frequency by two. This will help you find the middle number. Median will also be a range of numbers, since we cannot be exact.

If the frequency is odd, add 1 and then divide by two to find the middle number

Mean: Pick the center number out of each interval and multiply it by the frequency of that particular interval. Then divide by the total.

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

= mean

MED= median

**Day 2 Notes**

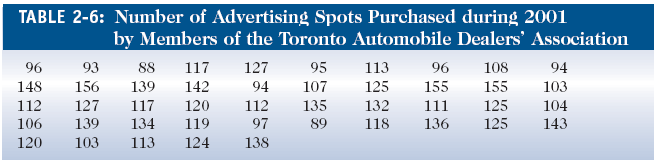
**SSHA SCORES** The Survey of Study Habits and Attitudes (SSHA) is a psychological test that evaluates college students’ motivation, study habits, and attitudes toward school. A private college gives the SSHA to a sample of 18 of its incoming first-year women students. Their scores were:

154 109 137 115 152 140 154 178 101

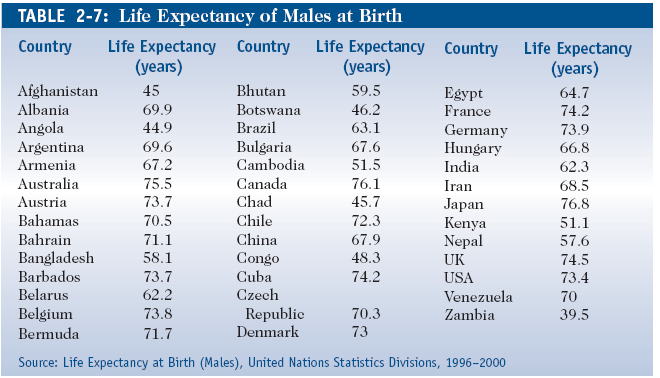
103 126 126 137 165 165 129 200 148

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. Then calculate it by hand. Compare the data. What do you notice?

**Numbers of Advertising Spots**



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



Refer to the data in Table 2-7 on life expectancy of males at birth in 40 countries. Construct a frequency table and histogram. What can you conclude from the graphs?

**Day 2 Notes**

**Five-Number Summary and Box and Whisker Plots**

When computing the five-number summary, you must have numbers in increasing order

**Five-Number Summary**

**Minimum (Min)**: The lowest number in the set

**Lower Quartile (Q1)**: The median of the data below the median

**Median (Med):** The middle number

**Upper Quartile (Q3):** The median of the data values above the median

**Maximum (Max):** The highest number in the set

**The Interquartile Range**: The difference between the first quartile and third quartile of a set of data. This is one way to describe the spread of a set of data.

For example, suppose we wish to compute the five number summary for the following observations and then find the IQR:

19 11 7 24 13 15 10 3 10 20

1. We order the observations.

3 7 10 10 11 13 15 19 20 24

2. The minimum and maximum are 3 and 24, respectively.

3. The median is because 11 and 13 are the two observations in the middle of the list.

4. The upper quartile range is every number in the set to the right of 12

13 15 19 20 24

The median of these observations is 19 because 19 is the middle number in this list. Thus, the upper quartile is 19.

5. Likewise the lower quartile is the median of the observations to the left of 12, i.e.

3 7 10 10 11

The first 10 in the list is the median of these observations, so the lower quartile is 10.

6. Interquartile Range: 12-10 = 2

**Day 2 Notes**

**Example**: The following set of data is the number of marbles fifteen different boys own. Compute a 5-number summary for the data.

18 27 34 52 54 59 61 68 78 82 85 87 91 93 100

Min: 18

Q1: 52

Med: 68

Q3: 87

Max: 100

IQR: 35

**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.

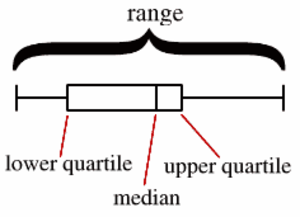
**Finding Outliers:**

* Compute the IQR and multiply it by 1.5
* Q1- (1.5\*IQR)= Outliers
* Q3+ (1.5\*IQR) = Outliers

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

**Day 2 Notes**

**Box and Whisker Plots**



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

* [Range](http://stattrek.com/Help/Glossary.aspx?Target=Range). If you are interested in the **spread** of *all* the data, it is represented on a boxplot by the horizontal distance between the smallest value and the largest value, including any outliers.
* [Interquartile range](http://stattrek.com/Help/Glossary.aspx?Target=Interquartile%20range) (IQR). The middle half of a data set falls within the interquartile range. In a boxplot, the interquartile range is represented by the width of the box (Q3 minus Q1).

And finally, boxplots often provide information about the shape of a data set. The examples below show some common patterns.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | |  |  |  |  | |  |  |  |  | | |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  |  |  |  |  |  |  |  |  | | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 |  | | | | | |  | |  |  |  |  | | --- | --- | --- | --- | |  |  |  |  | |  |  |  |  | | |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  |  |  |  |  |  |  |  |  | | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 |  | | | | | |  | |  |  |  |  | | --- | --- | --- | --- | |  |  |  |  | |  |  |  |  | | |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | |  |  |  |  |  |  |  |  |  | | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 |  | | | | | |
| Skewed right |  | Symmetric |  | Skewed left |

If most of the observations are concentrated on the low end of the scale, the distribution is skewed right; and vice versa.

If a distribution is symmetric, the observations will be evenly split at the median, as shown above in the middle figure.

**Day 2 Notes**

**How to construct a box-and-whisker plot**

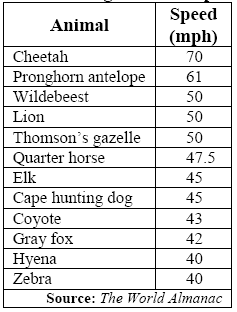
Construct a box-and-whisker plot for the following data: **The data:** Math test scores 80, 75, 90, 95, 65, 65, 80, 85, 70, 100

|  |  |
| --- | --- |
| Write the data in numerical order. Find the first quartile, the median, the third quartile, the minimum (smallest value) and the maximum (largest value).  These are referred to as a five statistical summary.  median (2nd quartile) = 80 first quartile = 70 third quartile = 90 minimum = 65 maximum = 100 | http://regentsprep.org/REgents/math/ALGEBRA/AD3/quartiles.gif |
| Place a circle beneath each of these values in relation to their location on an equally spaced number line. | http://regentsprep.org/REgents/math/ALGEBRA/AD3/quartiles2.gif |
| Draw a box with ends through the points for the first and third quartiles.  Then draw a vertical line through the box at the median point.  Now, draw the whiskers (or lines) from each end of the box to these minimum and maximum values. | http://regentsprep.org/REgents/math/ALGEBRA/AD3/quartiles3.gif |

**Day 2 Notes**

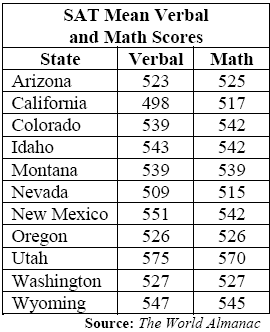
**Example 1 Draw a Box-and-Whisker Plot**

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



**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.



**Day 2 Notes**

**Box and Whisker Plots**

Given the data set  
 {85, 100, 97, 84, 73, 89, 73, 65, 50, 83, 79, 92, 78, 10},  
create a box and whisker plot to represent this data.

|  |  |
| --- | --- |
| **1.**   **CLEAR** out the graphs under **y =** (or turn them off). **2.** Enter the data into the calculator lists.          Choose **STAT, #1 EDIT** and type in entries.        (See [Basic Commands](http://mathbits.com/MathBits/TISection/Statistics1/BasicCommands.htm) for entering data.) | http://mathbits.com/MathBits/TISection/Statistics1/boxW1.gif |
| **3. Two icons for Box-and-Whisker Plots:**        Choose the **second icon** for beginning level work.                Press **2nd STATPLOT** and choose **#1 PLOT 1.**   You should see       the screen at the right.  Be sure the plot is ON, the second        box-and-whisker icon is highlighted, and that the list you will be       using is indicated next to **Xlist.**    **Freq: 1** means that each        piece of data will be counted one time. | http://mathbits.com/MathBits/TISection/Statistics1/boxpic2.jpg |
| ***What about that other icon?*** The first box-and-whisker icon is the modified box plot dealing with **outliers**.  This modified version will not plot points that are 1.5\*IQR beyond the quartiles. These points, called **outliers**, are plotted as individual points beyond the whisker in an attempt to give a more accurate picture of the dispersion of the data.  Notice the two plots displayed at the top of this page representing the same set of data. ***NOTE:***  IQR stands for the Interquartile Range which is Q3 – Q1. | http://mathbits.com/MathBits/TISection/Statistics1/boxpic3.jpg Used in more advanced statistics. |
| **4.**  **Seeing the graph:**       To see the box-and-whisker plot, press **ZOOM** and        **#9 ZoomStat.**   Press the **TRACE** key to see on-screen data       about the box-and-whisker plot.  The whiskers extend from the       minimum data point in the set to the first quartile, and from the       third quartile to the maximum point.   The box itself is defined       by Q1, the median and Q3.  The spider will jump from the       minimum value to Q1, to median, to Q3 and to the maximum value. | http://mathbits.com/MathBits/TISection/Statistics1/boxW2.gif |

You can create parallel box plots by turning on STATPLOT 2 (and/or 3) and choosing the appropriate list.

**Day 2: Homework**

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 |

1. 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.

1. Find the mean, median and mode wage.
2. How many employees earn more than the mean wage?\
3. 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.

1. Which class had the highest mean score?
2. Which class had the highest median score?
3. For which classes are the mean and median most different? Which is higher? Why?
4. Describe the shape of *each* graph.
5. Does there appear to be any gaps or outliers in any of the classes? If so, which ones? Explain.
6. Which class did better on the test overall? Explain.

**Day 2 Homework Cont’d**

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?

**Day 4 Notes**

**Variance and Standard Deviation**

Measures of dispersion describe how \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ the data are.

**Variance (ϭ2)**

Based Upon:

The sum of all\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ about the mean must equal \_\_\_\_\_\_\_\_\_

**Standard Deviation (ϭ):**

Measures the \_\_\_\_\_\_\_\_\_\_\_ of the distribution

*Example: Given the data set: 10, 8, 7, 5, 5*

1. Find the mean of the data set
2. Find the difference between each point in the set and the mean
3. Square the differences found in step 2
4. Take the average of the squares found in step 3

**Variance: Standard Deviation:**

How does the value of the standard deviation relate to the dispersion of the distribution? If we are comparing two populations, then the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ the standard deviation, the \_\_\_\_\_\_\_\_\_\_\_\_ dispersion the distribution has.

**Outliers:**  As previously discussed, the existence of outliers distorts the \_\_\_\_\_\_\_\_\_. The presence of outliers also affects the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_. Since these measures are often used for most statistical inferences, any conclusions drawn from a set of data that contains oultiers can be flawed.

**Day 4 Notes**

**Checking for outliers (same as with box-plots)**

Step 1: Determine the Q1 and Q3 of the set of data

Step 2: Compute the IQR (\_\_\_\_\_\_\_ - \_\_\_\_\_\_\_\_)

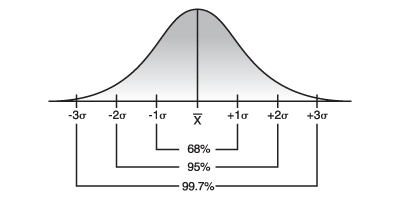
Step 3: Determine the fences. These fences serve as cutoff points for determining outliers.

Lower Fence= \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Upper Fence= \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Step 4: If a data value is less than or equal to the lower fence or greater than or equal to the upper fence, then it is considered an outlier

**Empirical Rule**

Given a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ of data, nearly all values lie within \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ standard deviations of the mean.

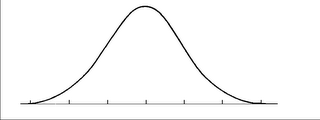
Within 1 ϭ of the mean:

Within 2 ϭ of the mean:

Within 3 ϭ of the mean:

**Example:**

Out of 50 students, the mean test score was 78% with a standard deviation of 3.

1. Draw the normal distribution curve
2. What percent scored below a 72?
3. What percent scored above an 81?
4. Out of 50 students, how many scored more than 81?

**Day 4 Classwork**

**Variance and Standard Deviation**

**1)** Why do we measure dispersion?

**2)** How do we calculate deviation and what does it mean?

**3)** How do we calculate variance?

**4)** How do we calculate standard deviation and what does it mean?

**5) Number of Students Deviation (Deviation)2**

8,603

8,605

5,538

6,541

4,980

5,420

4,367

4,466

4,505

5,048

**Sum**

Mean \_\_\_\_\_\_\_\_\_\_ Range \_\_\_\_\_\_\_\_\_ Variance\_\_\_\_\_\_\_\_\_\_ Standard Deviation\_\_\_\_\_\_\_\_\_\_

**6) Auto Sales (in millions) Deviation (Deviation)2**

11.2

11.9

12.0

12.8

13.4

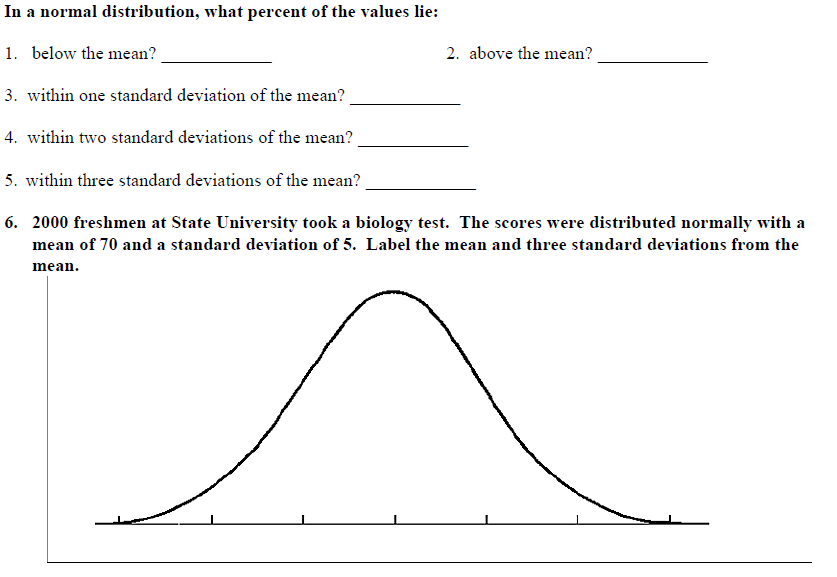
14.3

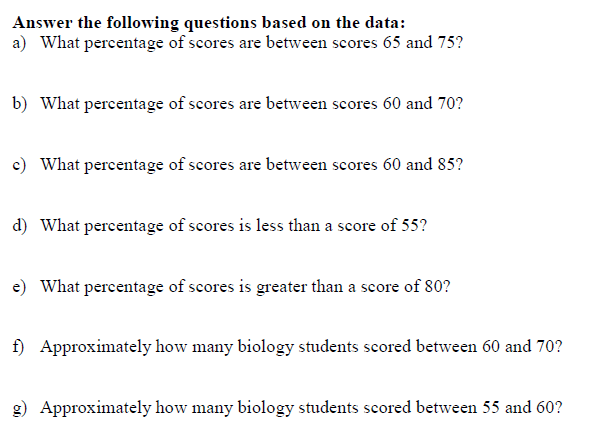
**Sum**

Mean \_\_\_\_\_\_\_\_\_\_ Range \_\_\_\_\_\_\_\_\_ Variance\_\_\_\_\_\_\_\_\_\_ Standard Deviation\_\_\_\_\_\_\_\_\_\_

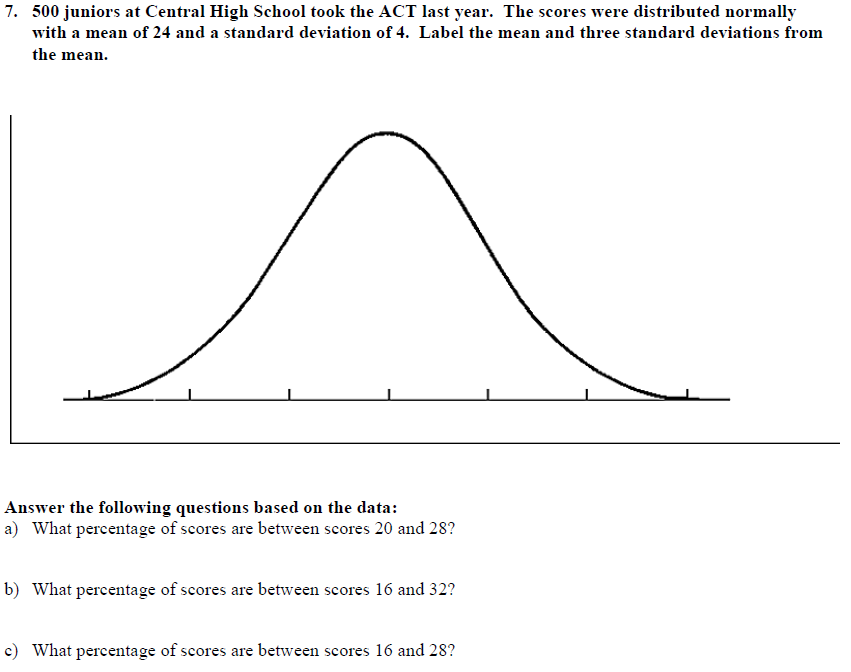
**Day 4 Classwork Cont’d**

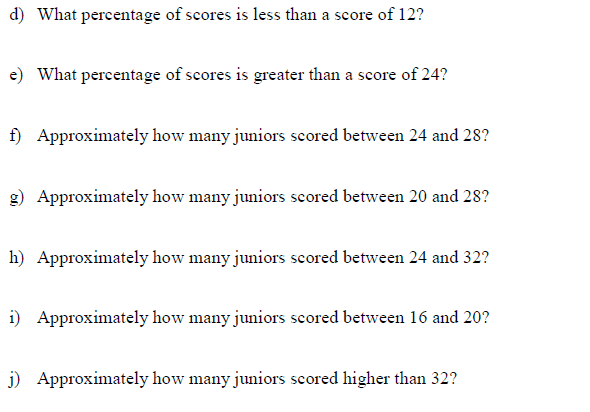
**Empirical Rule**



****

**Day 4 Classwork Cont’d**

****

****

**Day 4 Homework**

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

1. 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?

1. 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

1. 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.

1. 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 |

1. 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.
2. Make the curve to represent the normal distribution.
3. Of the parents surveyed, how many spent more than $220?
4. What percentage of the parents spent less than $184?

**Day 5 Notes**

**POPULATION**

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**

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?”

**Five good basic sampling techniques:**

***Simple random sampling, stratified sampling, systematic sampling, cluster sampling and convenience sampling.***

**A) Simple random sampling** (often abbreviated as random sampling):

Every possible sample has an equally likely chance of occurring.

The sample is always a subset of the population.

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.

Using a graphing calculator to obtain a simple random sample:

1. Press the MATH button. Highlight PRB menu and select 5: randInt(.

2. With randInt (on the Home screen enter 1, N where N is the population size. For example, if N=500, enter the following: randInt (1, 500)

Press ENTER to obtain the first individual in the sample, Continue pressing ENTER until the desired sample size is obtained.

**Day 5 Notes**

**B) Stratified sampling:**

A stratified sample is obtained by separating the population into nonoverlapping groups called *strata* and then obtaining a simple random sample from each stratum (or group). The individuals in each stratum should be homogeneous (or 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.]

**C) 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.]

**D)** **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.]

**E) Convenience Sampling:**

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.

**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

**Day 5 Homework** Samplings Name \_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

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

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

3. 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.

4. 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.

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

6. 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 9th grade students at the selected schools.

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

8. 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.

9. 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.

10. 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.

11. 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.

12. 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.

**What are some errors that you think might occur in the sampling situation discussed below:**

1. A senator explains a vote in favor of a new bill by stating that 60% of the mail received favored the bill.
2. A radio talk show host invites listeners to telephone the station and talk about their feelings on a proposed highway to be built in their county.
3. A newspaper reporter randomly stops people going in a grocery store and asks, “Does your family use newspaper coupons?”

**Day 6 Warm Up**

**Identify the POPULATION and SAMPLE**

1. In a class of 30 students, each student is asked if he or she has watched a

cartoon within the past month.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2. People in a live TV audience who had aisle seats were asked if they

traveled more than 10 miles to see the game show.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3. Survey every tenth child entering a park to find out how many rides they

plan to go on while visiting the park.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4. Survey all of your friends to see how many can play video games with

you.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5. Survey every thirtieth person at the exit door of the zoo to find out if the

zoo should increase its hours of operation.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Identify the sampling technique**

1. Mr. Crayton would like to know the average number of hours

that his students spend on homework each week.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2. A boy scout gives a survey to people as they walk outside of a movie

theater.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3. Pauly D has four passes to the movies. He can’t decide which one of his

roommates to take. He puts everyone’s name in a hat and selects four names

without looking.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4. Sammie interviews every 100th person in the phonebook on her duck phone.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5. MTV asks people to call in and vote on who their favorite Jersey Shore character is.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

6. Vinny interviews people as they walk out of a local shopping mall.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

7. Snooki wants to find out whether the girls at the Jersey Shore like the Shore

Store as much as boys. She asks interviews 50 boys and 50 girls at

the Shore.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Day 6 Classwork**

**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.

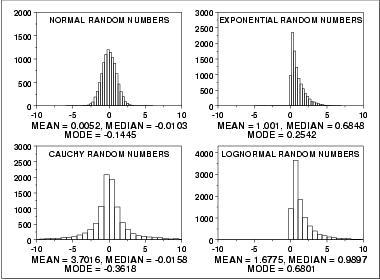
7. The researcher stands at a shopping mall and selects the first 75 shoppers as they walk by to fill out a survey.

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 distribution of each graph below.



**Day 6 Classwork**

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

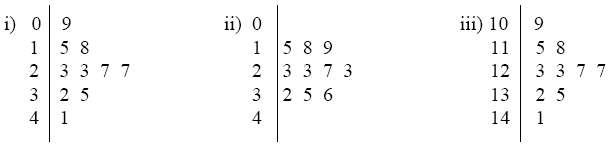
What percentage of the City Hall employees are between 31.8 and 68.4 years old?

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



**Day 6 Extra/Homework**

AFM Study Guide 2

I. 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 |  |

1. Create a stem and leaf display of these data.
2. Create a histogram of this data.
3. Find the 5 number summary.
4. State the IQR. What does this information tell you about the number of birds sighted?
5. Write a brief description of the distribution of the data.
6. 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?
7. Calculate the mean, median and mode. Which central tendency best represents the data? Explain.
8. 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)?
9. 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?
10. 68% of the data falls between \_\_\_\_\_\_\_\_\_\_, 95% of the data falls between \_\_\_\_\_\_\_\_\_\_, and 99% of the data falls between \_\_\_\_\_\_\_\_\_\_.
11. The percentage that spotted over 232 birds is?

II. 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.

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

II. Given the frequency table, find the following:

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

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

III. Explanation problems.

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

IV. KNOW ALL THE SAMPLING METHODS. BE ABLE TO GIVE EXAMPLES OF EACH! (Refer to your notes from blackboard and the homework problems)

Answer Key to Study Guide 2 Univariate Data

1.

|  |  |
| --- | --- |
| 15 | 23367 |
| 16 | 0022367 |
| 17 | 578 |
| 18 | 136 |
| 19 |  |
| 20 | 66 |
| 21 |  |
| 22 | 8 |

Key: 15/2 = 152 birds

2.



3. Min = 152 birds

Q1 = 158.5 birds

Med = 166 birds

Q3 = 182 birds

Max = 228 birds

4. IQR = 182 – 158.5 = 23.5 birds (This tells us that 50% of the counts falls between 158.5 and 182)

5. Spread = 76 birds

Gaps = 186 – 206 and 206 – 228

Shape = skewed right

Outliers = 228 bird count

6. Any count below 123.25 or above 217.25, yes count of 228 is an outlier. If we took the outlier out, the max would be lowered (because the outlier was the max value, so now the next highest number will be the max), the median, Q1, and Q3 would change because there would be an even number of data now, so we would take the average of the two center pieces of data.

7. Mean = 172.9  173 birds, Median = 166 birds, Mode: 153, 160, 162, 206 birds

The data is not continuous (due to the gaps in our data) and because we have an outlier, our data is skewed, therefore causing the median to be the best representation because it is unaffected by extreme values.

8. The mean goes down by 1 (changing the summation), the standard deviation stays the same (the data still varies the same amount from the mean), and the five number summary all goes down by 1 (because all values were decreased by 1, they are in the same location in the number line, just lowered by 1)

9. 2.65, 0.5

10. 153.2 – 192.6, 133.5 – 212.3, 113.8 – 232

11. 0.5%

12. Normal distribution curve with benchmarks, starting from left to right (30, 45, 60, 75, 90, 105, 120).

13. 475 students

14. 840 students.

15. Mean = 82.7

Median = 85

Mode = 85

Total students = 27

16. The mean goes up by 5 (again, summation changes), but the standard deviation stays the same (data varies the same still)

17. By using a frequency table, we don’t know the exact values, so in this case it appears nothing changes but if we knew the exact values, since their would be an even set of data now, it would change all but possibly the min/max.

18. The larger the standard deviation, the greater the variation among the data. This means that the data is not close to the mean, or the average of everything else.

19. Look at homework/quizzes!

20. Look at the sampling example