AFM
Unit 6 Day 1 Notes– Central Tendency; Spread; Displaying Data

Name Key
Date ______________________

Measures of Central Tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.

- **mean** – the *average* of a set of data. The sum of a set of data divided by the number of data. (Do not round your answer unless directed to do so.)

- **median** – the *middle* value, or the average of the middle two values, when the data is arranged in numerical order.

- **mode** – The value (number) that appears the *most.*
  It is possible to have more than one mode, and it is possible to have no mode. If there is no mode-write "no mode", do not write zero (0).

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.

- When is it best to use the mean? no outliers ; symmetrical data

- When is it best to use the median? outliers ; skewed data

- When is it best to use the mode? categorical data

Other definitions

- **range** – the difference between the highest (max.) and lowest (min.) values in a data set

- **interquartile range** – $Q_3 - Q_1$ $1, \boxed{2}, 3, \overset{\text{median}}{\boxed{4}}, 5, \boxed{6}, 7$
  (Lower Quartile) Q_1 Q_3 (Upper Quartile)

- **five-number summary** - For a set of data, the minimum, first quartile (Q_1) , median, third quartile (Q_3), and maximum. Note: A boxplot is a visual display of the five-number summary. Q_2

- **random sample** – A subset of a statistical population in which each member of the subset has an equal probability of being chosen. A simple **random sample** is meant to be an unbiased representation of a group.

Quantitative vs Categorical Data

The data we will be working with in this unit is called **Univariate Data**. Univariate data involves a single variable. It does not deal with causes or relationships and its main purpose is to describe.

- **Quantitative Data** - Quantitative data are numeric. They represent a measurable quantity. For example, when we speak of the population of a city, we are talking about the number of people in the city - a measurable attribute of the city. Therefore, population would be an example of quantitative data.

- **Categorical Data** - take on values that are names or labels. The color of a ball (e.g., red, green, blue) or the breed of a dog (e.g., collie, shepherd, terrier) would be examples of categorical data.

Unit 6 Day 1 HW(1)

Determine whether the following variables are categorical (C) or quantitative (Q)

C 1. Brand of vehicle purchased by a customer

Q 2. Price of a CD

C 3. Type of M&Ms preferred by students (peanut, plain)

C+Q 4. Phone number of each student

Q 5. Height of a 1-year old child

C 6. Term paper status (turned in on time or turned in late)

C 7. Gender of the next baby born at a particular hospital.

Q 8. Amount of fluid (oz) dispensed by a machine used to fill bottles with soda

Q 9. Thickness of the gelatin coating on a Vitamin C capsule

C 10. Brand of computer purchased by a customer

C 11. State that a person is born in

Q 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

~~3~~ ~~4~~ ~~7~~ ~~8~~ $\boxed{9}$ ~~9~~ ~~10~~ ~~11~~ ~~12~~ $\boxed{12 \quad 13}$ ~~15~~ ~~16~~ ~~18~~ ~~19~~ $\boxed{20}$ ~~21~~ ~~21~~ ~~22~~ ~~33~~

Use the above data to find the following measures.

a. Mean: 14.15 lb b. Median: 12.5 c. Mode: 9, 12, 21 d. Range: 33-3 = 30

e. Five-number summary:

Min.	Q₁	Median	Q₃	Max
3	9	12.5	20	33

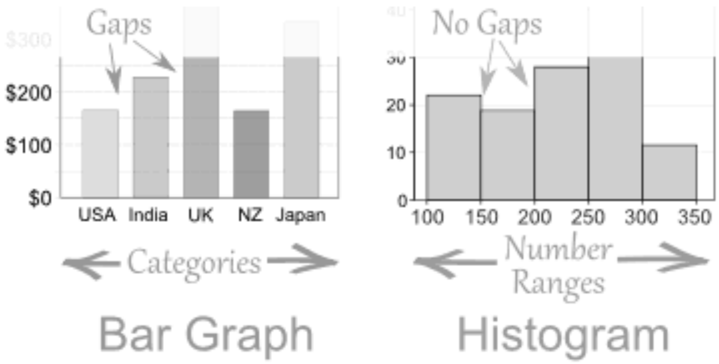
~~What would make the data in Owen's study unfair, or biased?~~

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

Displaying Data

- **Histograms** ✱ DO FREQUENCY TABLE 1ST

A **histogram** is a graphical representation of a one-variable 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.



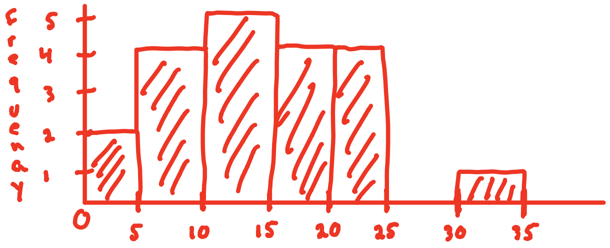
In a histogram, the height of each bin represents the frequency, or the number that falls in that interval. The width of each bin represents an interval, in this case each interval 50.

Let's make a histogram to represent the data that Owen collected:

Frequency Table

5 is in this bin

0 - 5	2
5 - 10	4
10 - 15	5
15 - 20	4
20 - 25	4
25 - 30	0
30 - 35	1

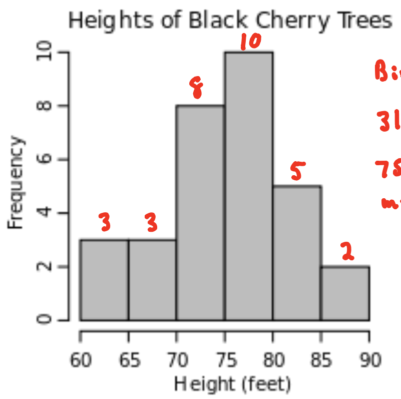


- Pros: Good way to display large data set; can see shape of data + distribution
- Cons: Can't see individual data values

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.

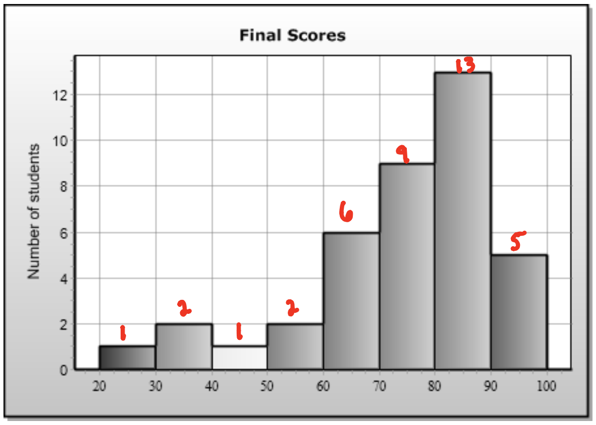


Bin Width: 5

31 data values

75 - 80 is where median will occur

b.



Bin Width: 10

39 data values

80 - 90

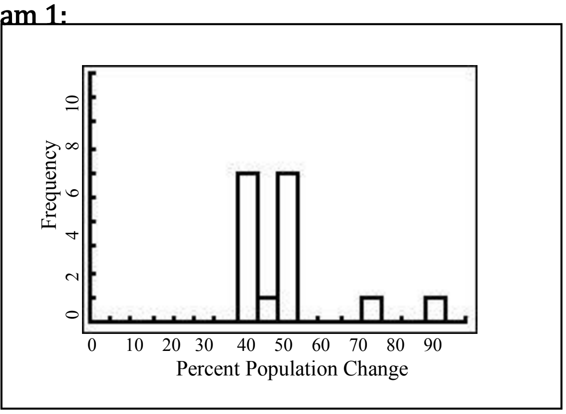
The **percentile rank** of a data value in a large distribution gives the percentage of data values that are below the given value. For example, if you are in the 95th percentile on your PSAT, you have done better than 95% of the other students your age that took that test.

Example B

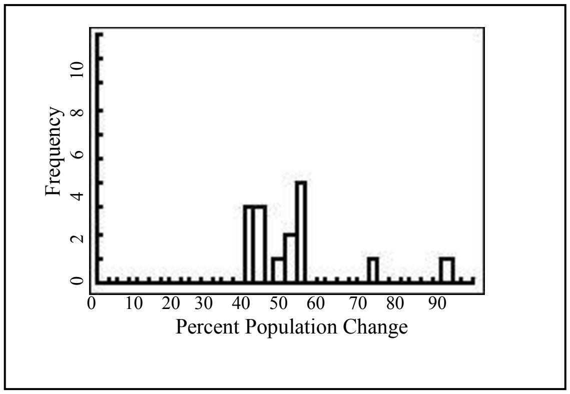
The following histograms were both constructed with the data below.

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

Histogram 1:



Histogram 2:



a. What is the range of the data?

83.3 − 35.1 = 48.2

b. What is the bin width of each graph?

#1 : 5

#2 : 2.5

~~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~~ 15 metropolitan areas?

Add frequencies of bins

e. Why are the columns shorter in Graph B?

Bin width changed

- **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.

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

Weight of Backpack	Frequency
0 – 5	2
5 – 10	4
10 – 15	5
15 – 20	4
20 – 25	4
25 – 30	0
30 – 35	1

- Pros: Good way to display large data set; can see shape of data + distribution
- Cons: Can't see individual data values

○ **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:

1 9

↑ ↑

tens ones

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

302

304

310

312

stem	leaf
30	2, 4
31	0, 2

Key: 2 | 1 = 21

↑ could mean 2.1 w/out 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 into a back to back stem and leaf plot separating Juniors and Seniors.

Juniors		Seniors
8, 7, 4	0	3, 9, 9
8, 6, 0	1	1, 2, 2, 3, 5, 9
1	2	0, 1, 2
3	3	

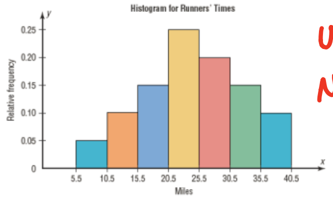
- Pros: similar shape as histogram; can see data values
- Cons tedious to create for large data sets

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

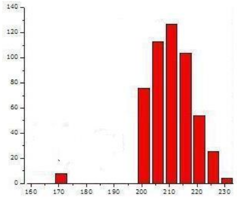
Use as many of these vocabulary words to describe the following displays

1.



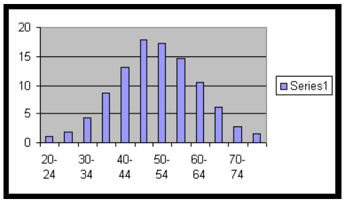
Uni-Modal
Normal

3.



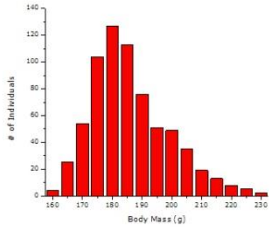
gaps
outliers
Uni-modal
Short tail

5.



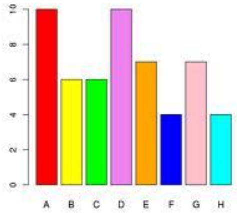
Uni-Modal
Normal
Symmetric

2.



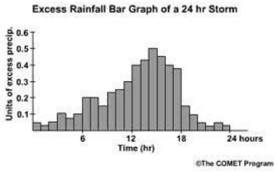
Skewed right
Long tail
Uni-modal

4.



Bi-Modal

6.



Uni-Modal
Long tail
Skewed left

Mean, Median, Mode…Which One?

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

Unit 6 Day 1 HW(2): 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 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.

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

Unit 6 Day 1 HW(3): 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

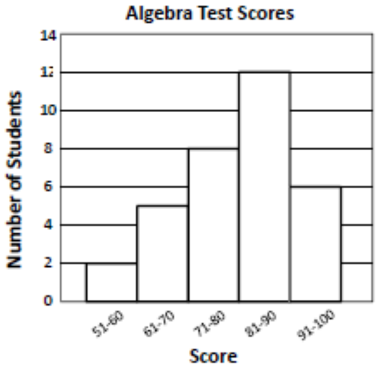
Interval	Frequency
80-90	
90-100	
100-110	
110-120	
120-130	

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.

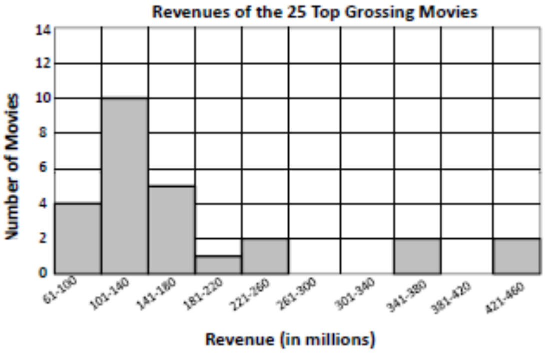


a. How many total students took the test?

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?

Unit 6 Day 1 HW(4): 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.