AFM - Unit 1, Statistics

Name ______________________

Notes 1.1, Central Tendency, Box Plots

Date ____________ Period __________

## Lesson 2.1, Measures of Central Tendency and Box Plots (AA text pp. 76 - 80)

- Measures of central tendency, also referred as measures of center, refer to different types of averages.
- The most common measures of central tendency are mean, median, and mode.

### MEAN

- The symbol for the mean is $\bar{x}$, which is read as x-bar.
- Another symbol for the mean is $\mu$, which is read as mew.

### MEDIAN

→ Numbers must be in order from least to greatest

- Median refers to the middle value of a set of data once it has been ordered from least to greatest. The median of a set of data with an even number of values is not a number in the data set. {4,7,9,15,24,40}

$\frac{9+15}{2} = \frac{24}{2} = 12$ is the median

{4,7,9,15,24} → 9 is median

### MODE

- Mode refers to the number that appears most frequently in a set of data. Data sets with two modes are said to be bi-modal. Sets have no mode when each item of the set has equal frequency

### Ex. 1: Salary Data

Find the mean, median, and mode of the salaries for the corporate employees listed below. Which measure of central tendency appears to most accurately represent the set of data?

| | |
|---|---|
| Allen: | $40,000 |
| Baker: | 42,000 |
| Chase: | 59,000 |
| Deitz: | 60,000 |
| Eckerd: | 62,000 |
| Francis: | 65,000 |

mean : $54,667

median : $ 59,500

mode : $ None → since no # repeats

* put in calculator as a list set.

How do extreme values (outliers) affect the measures of central tendency?

- Mean - changes significantly, because
- Median -
- Mode - does <u>NOT</u> change

Ex. 2: Backpack Weights

Owen is a member of the student council and wants to present data about backpack safety to the school board. He collects data on the weights of backpacks of 30 randomly chosen students. How much does the typical backpack weigh at Owen's school? 9 to 10 lbs.

Q1 Q2 Q3

{ 3, 4, 4, 4, 6, 7, 7, 7, 7, 7, 8, 8, 9, 9, 9, | 9, 9, 10, 10, 10, 10, 10, 10, 13, 15, 15, 16, 17, 20, 33 }

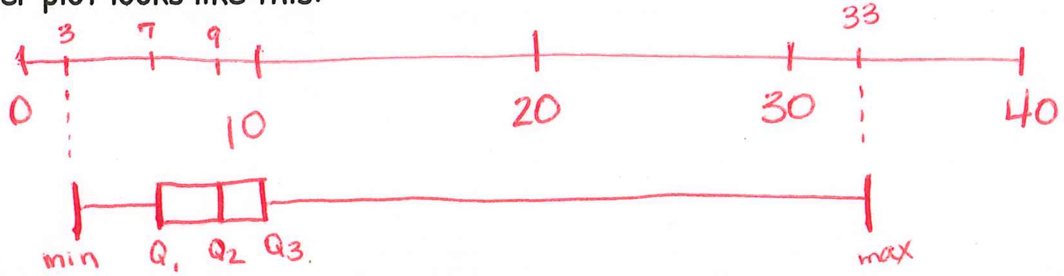
min max

mean : 10.2 lbs. mode: 10 lbs.

median: 9

## Box and Whisker Plots and the Five Number Summary

Next we will look at Box and Whisker Plots (aka Box Plots). They are used to summarize a data set and to visually illustrate the variability (spread) of the data. A Box and Whisker plot looks like this:



The five parts of a Box and Whisker plot for a particular data set correspond to the Five Number Summary for that data set. The five numbers in the Five Number Summary are the minimum, 1st quartile, 2nd quartile (median), 3rd quartile, and maximum.

1st: Arrange the data in order and find the median. This separates the data into 2 groups.

2nd: Find the median of the 1st half and 2nd half of the data set.

Now your data set is divided into four groups, and each of these four groups is called a quartiles. There are 3 points called quartile points, ($Q_1$, $Q_2$, and $Q_3$) that denote the breaks in the data for each quartile.

- $Q_1$ is the median of the first half of the data set
- $Q_2$ is the median of the entire data set
- $Q_3$ is the median of the second half of the data set
- The difference between Q1 and Q3 (i.e., Q3-Q1) is called the inter quartile range
- The difference between the maximum and minimum values is called the range

Box-and-Whiskers plots…

- can be drawn vertically or horizontally
- consists of a rectangular box with the ends, or edges, located at the first and third quartiles
- the segments extending from the ends of the box are called Whiskers
- the whiskers stop at the minimum and maximum values of a data set unless it contains outliers.

Outliers

- Outliers are extreme values
- The technical definition of an outlier is a data point that is more than 1.5 of the interquartile range beyond the upper or lower quartiles. That is, any number less than $Q_1$ - 1.5(IQR) or greater than $Q_3$ + 1.5(IQR) is considered an outlier.
- Outliers are exteme values represented by single points on a box plot.
- If outliers exist, each whisker is extended to the last value of the data set that is not an outlier.

Examples:

MATH III – Unit 1, Statistics

Name ____________________

Notes 8-2, Dot Plots and Distributions

Date ________ Period _______

A data set is a set of related numbers, often called data points

A distribution is the way the numbers in a data set are distributed

A dot plot is a way of representing a distribution in graphical form

Example 1) The heights (in inches) of each member of the girls' and boys' basketball teams at Holbrook High School are shown below.

Boys' team: 68, 69, 69, 73, 73, 74, 74, 74, 74, 76, 77, 79

Girls' team: 65, 69, 69, 70, 70, 71, 71, 71, 71, 72, 72, 74

Sketch a dot plot for each data set.

The two dot plots show how the heights of the two groups of basketball players are distributed. How would you describe, in words, these two distributions?

|  | Boys | Girls |
|---|---|---|
| mean : | 73.3 | 70.4 |
| median : | 74 | 71 |
| mode : | 74 | 71 |

MATH III Unit 8, Statistics Name ______________________

Notes 8-3, Normal Distributions, Mean, S.D. Date ____________ Period ____________

## The Normal Distribution

When you draw a dot plot for some data sets, you get a distribution that has a particular shape. It looks like this:

This distribution shape is so common, and there are so many different data sets that produce it, that it is given a special name. It is called a normal distribution. (You may have also heard it called a bell-shaped curve.)

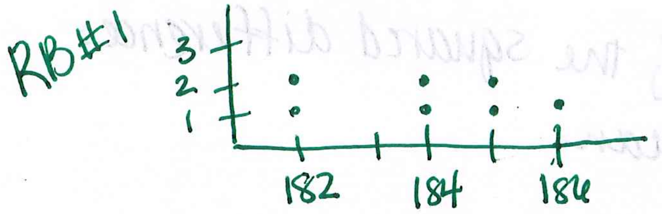
When you have a data set that is normally distributed, that means that if you were to draw a dot plot of the data set, it would have this characteristic "bell" shape.

For a normally distributed data set, there are two values that we can calculate that will tell us a GREAT DEAL about the data set.

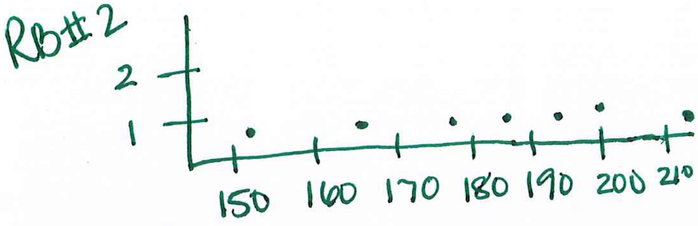
1. The value of the mean, which is a measure of central tendency
2. The value of the standard deviation (SD), which is a measure of spread or how spread out. (The greater the SD, the greater the spread of the data about the mean.)

Example 1: The Rubber Band Launch (P. 85-86 in Green AA text)
You want to find out how consistently rubber bands will travel when launched, so you use a ruler to launch two rubber bands seven times each. You generate the following data sets:

- Rubber band #1 distances (cm): {182, 186, 182, 184, 185, 184, 185}
- Rubber band #2 distances (cm): {152, 194, 166, 216, 200, 176, 184}



mean 184



mean 184

RB#1

| Data Point | Mean | Deviation from Mean | Squared Deviation From Mean |
|---|---|---|---|
| 182 | 184 | −2 | $(-2)^2 = 4$ |
| 186 | 184 | +2 | $(2)^2 = 4$ |
| 182 | 184 | −2 | 4 |
| 184 | 184 | 0 | 0 |
| 185 | 184 | +1 | 1 |
| 184 | 184 | 0 | 0 |
| 185 | 184 | +1 | 1 |
|  |  | total | 14 |

$$\text{Variance} = \frac{14}{7-1} = \frac{14}{6} = 2.33$$

← squared deviation

↑ # of data points

$$\text{Standard deviation} = \sqrt{\text{variance}} = \sqrt{2.33} = 1.53$$

*Variance: the average of the squared differences from the mean

Name ____________________

Notes 8-4, Normal Distributions and Percentages Date ____________ Period ________

# Normal Distributions and Percentages

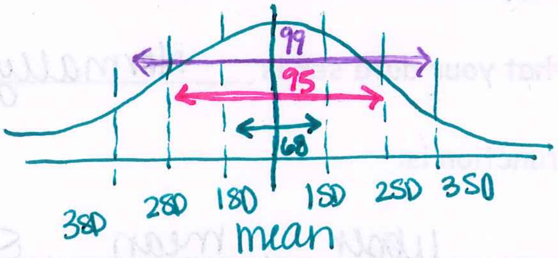
The Empirical Rule - (the 68-95-99 Rule)

In any large data set that is normally distributed:

Approx. 68% of the values will be within 1 standard deviation of the mean

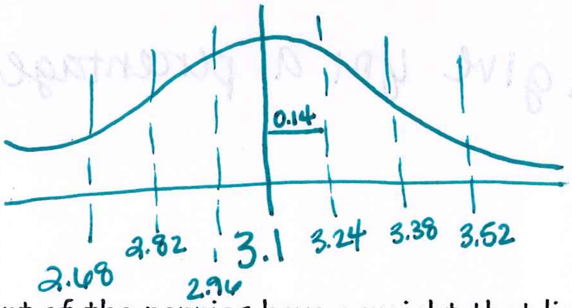
Approx. 95% of the values will be within 2 standard deviations of the mean

Approx. 99% of the values will be within 3 standard deviations of the mean



Ex. 1: A group of students weighs 500 US pennies. They find that the pennies have normally distributed weights with a mean of 3.1g and a standard deviation of 0.14g.

a) Sketch the normal curve for this distribution below. Label the mean and three standard deviations above and below the mean.



b.) What percent of the pennies have a weight that lies between:

2.96g and 3.24g (i.e., within one standard deviation of the mean)? 68.3%

2.82g and 3.38g (i.e., within two standard deviations of the mean)? 95.5%

2.68g and 3.52g (i.e., within three standard deviations of the mean)? 99.7%

c.) How many pennies have a weight that lies within

2.96g and 3.24g (i.e., within one standard deviation of the mean)? .68 × 500 = 342

2.82g and 3.38g (i.e., within two standard deviations of the mean)? .95 × 500 =

2.68g and 3.52g (i.e., within three standard deviations of the mean)? .99 × 500 =

What if I wanted to know the percentage of pennies that had a weight between 3g and 3.2g?

Calculator Function: **normalcdf()**

The TI83/TI84 calculators have a function called normalcdf() which will tell you:
the % of values that lie within a given interval

and all you have to give it is: the given interval
mean
standard deviation

(Note that normalcdf assumes that your data set is Normally Distributed.)

The format of the normalcdf() function is:

normalcdf(lower bound, upper bound, mean, standard deviation)

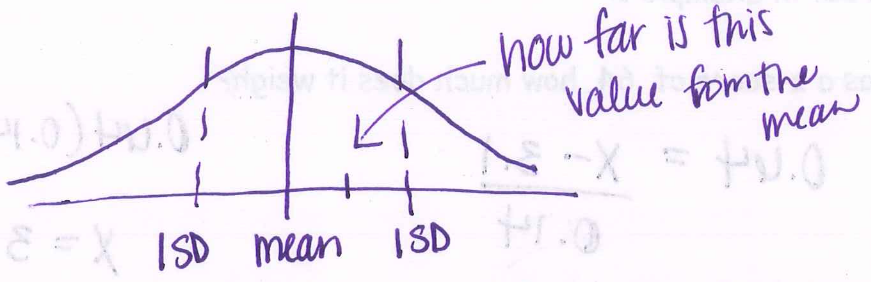
So if we wanted to know the percentage of pennies from our data set that had a weight between 3g and 3.2g, we would enter the following into our calculator:

normalcdf (3, 3.2, 3.1, .14)

this will give you a percentage

MATH III Notes, Unit 8, Statistics Name ____________

Notes 8-5, Z-scores Date ________ Period ________



The z-score of a data point: the number of standard deviations the point if from the mean

A z-score or z-value can be calculated for any point/value in a data set.

To calculate the z-value for a given data point:

$$Z = \frac{\text{deviation from mean}}{\text{standard deviation}} = \frac{\text{data point} - \text{mean}}{\text{S.D.}} = \frac{X-\bar{X}}{S_x}$$

Ex 1: A group of students weighs 500 US pennies.
They find that the pennies have normally distributed weights with a mean of 3.1g and a standard deviation of 0.14g

a) What is the z-score for a penny that weighs 3.24g?

$$Z = \frac{3.24 - 3.1}{0.14} = \frac{.14}{.14} = 1$$

b) What is the z-score for a penny that weighs 2.96g?

$$Z = \frac{2.96 - 3.1}{0.14} = \frac{-0.14}{0.14} = -1$$

c) What is the z-score for a penny that weighs 3.31g?

$$Z = \frac{3.31 - 3.1}{0.14} = \frac{0.21}{0.14} = 1.5$$

d) What is the z-score for a penny that weighs 2.89g?

$$Z = \frac{2.89 - 3.1}{0.14} = \frac{-0.21}{0.14} = -1.5$$

A positive z-score indicates the data point lies above the mean.

A negative z-score indicates the data point lies below the mean

Ex. 2: For the data set in Example 1:

a.) If a penny has a z-score of .64, how much does it weigh?

$0.64 = \frac{X-3.1}{0.14}$

$0.64(0.14) = X-3.1$

$X = 3.1896$

b.) If a penny has a z-score of -2.8, how much does it weigh?

$-2.8 = \frac{X-3.1}{0.14}$

$-2.8(0.14) = X-3.1$

$-0.392 = X-3.1$

$+3.1 \quad +3.1$

$2.708 = X$

$X = 2.708$

MATH III Notes, Unit 1, Statistics

Notes 1-6 Samples and Surveys

Name ____________

Date ____________ Period ____________

A population is all the members of a set.

A sample is part of a population.

If you determine a sample carefully, it can give a good estimate of the total population.

## Sampling Types and Methods

1. Convenience sample - select any members of the population who are conveniently and readily available.
2. Self-selected sample - select only members of the population who volunteer for the sample.
3. Systematic sample - order the population in some way, and then select from it at regular intervals.
4. random sample - all members of the population are equally likely to be chosen.

A bias is a systematic error introduced by the sampling method.

### Example 1 Analyzing Sampling Methods

A newspaper wants to find out what percent of the city population favors a property tax increase to raise money for local parks. What is the sampling method used for each situation? Does the sample have a bias? Explain.

A. A newspaper article on the tax increase invites readers to call the paper and express their opinions. self-selected

bias - some people who are against the tax might organize the tax ~~agaist~~ a campaigne to get friends + neighbors to call in

B. A reporter interviews people leaving the city's largest park.

convenience b/c ~~the phone listing is ordered alphabet.~~

bias if there is overreprenentation of park supports

b/c it is convenient for the reporter to be in one place

C. A survey service calls every 50th listing from the local phone book.

systematic - b/c the phone listing is ordered alphabetically
bias - if there is some link b/t ppl who are listed in phone book & pay property taxe.

## Study Methods

1. observational - measure or observe members of a sample in such a way that they are not affected by the study.

2. controlled experiment - divide the sample into two groups. You impose a treatment on one group but not on the other "control" group. Then you compare the effect on the treated group to the control group.

3. survey - ask every member of the sample a set of questions.

A poorly written survey question can introduce bias. It should avoid:

- combining two or more issues
- using double negatives
- overlapping answer choices
- words that cause strong reaction (loaded question)
- suggest that you want a particular answer (a leading question)

## Example 2 Analyzing Survey Questions

Is there any bias in the survey question? Explain.

A. Do you think farmers should use poison to control insects on crops?

loaded - using the term poison instead of pesticide & could cause strong reaction

B. Don't you agree that most childcare workers are underpaid?

leading question - suggest you want a certain answer that childcare workers are underpaid

C. Do you think teachers should communicate frequently with students and their parents about class grade?

Ask about two issues: teachers communicating w/ students & teachers communicating w/ parents

MATH III Notes, Unit 1, Statistics

Notes 1-7 Margin of Error

Name ____________________

Date ____________ Period ____________

## Margin of Error

Margin of Error is a value that tells the uncertainty in an estimate.

It is a measure of how close we believe the sample proportion is to the population proportion.

The formula used to predict MOE with 95% confidence ≈ $\frac{1}{\sqrt{n}}$.

(margin of Error)

The margin of error is roughly two standard deviations away from the mean.

Example 1

During the week of 08/10/2001, CNN conducted a poll asking 1000 Americans whether they approve of President Bush's performance as President. The approval rating was 57%. In the next poll conducted during the week of 09/21/2001, CNN conducted the same poll asking 100 Americans whether they approve of President Bush's performance as President. The approval rating was 90%.

1. Why the difference in ratings?

   The attacks of Sept. 11th

2. Find the MOE in the August poll.

   $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{1000}} = \frac{1}{31.622} = .032 = 3.2\%$

3. Find the MOE in the September poll.

   $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{100}} = \frac{1}{10} = .1$ or 10%

4. Explain why the MOE for the August poll is less than that in September.

   The total # of people asked was greater.