AFM – Unit 2, Statistics

Notes 2-1 Central Tendency and Box Plots

Name Key

Date ____________ Period ____________

Measures of Central Tendency

- 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 μ, which is read as mew.

MEDIAN

- 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 the mean of the two middle numbers.

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

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

- Mean - \$54,667
- Median - \$59,500
- Mode - none

Ex. 2: Backpack Weights

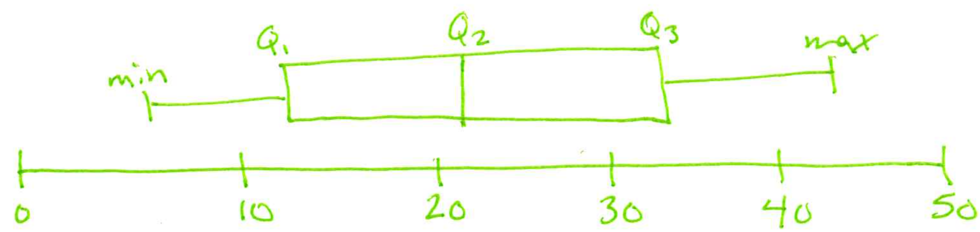
Owen is a member of the student council and wants to present data about backpack safety to the school board. He collects data on the weights of backpacks of 30 randomly chosen students. How much does the typical backpack weigh at Owen's school?

{ 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 }

mean 10.2 lbs median 9 lbs mode 10 lbs

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 1st quartile, 2nd quartile, 3rd quartile, minimum, 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 interquartile 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 medians, 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 outside 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 value 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.

AFM - Unit 2, Statistics

Notes 2-2 Dot Plots and Distributions

Name Key

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. (spread out)

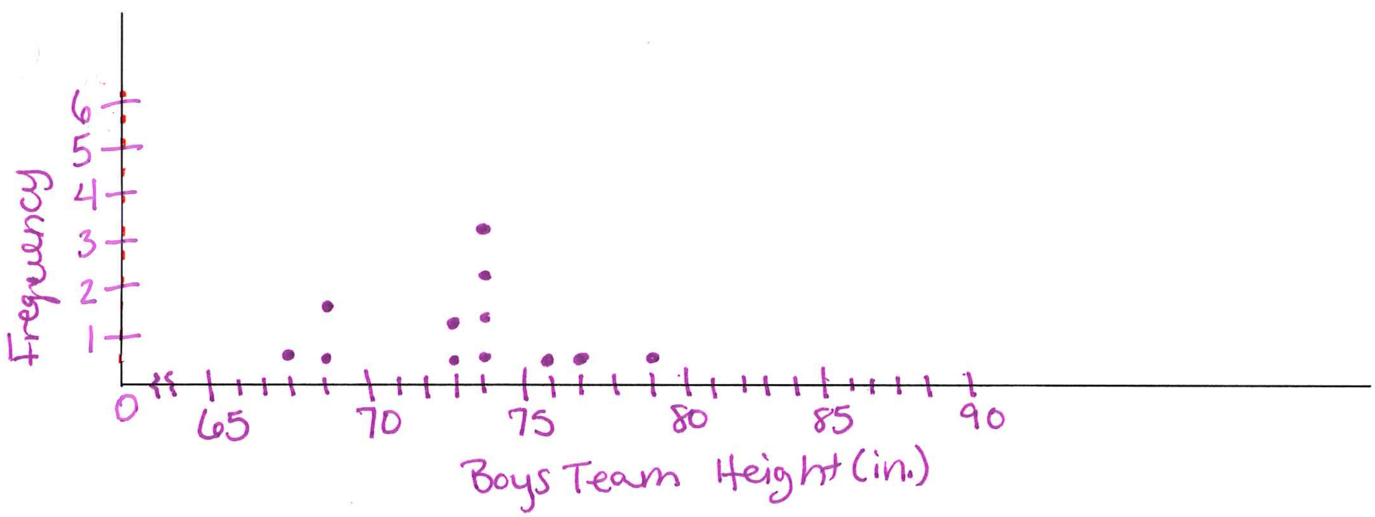
A dot plot is a way of representing a distribution in graphical form.

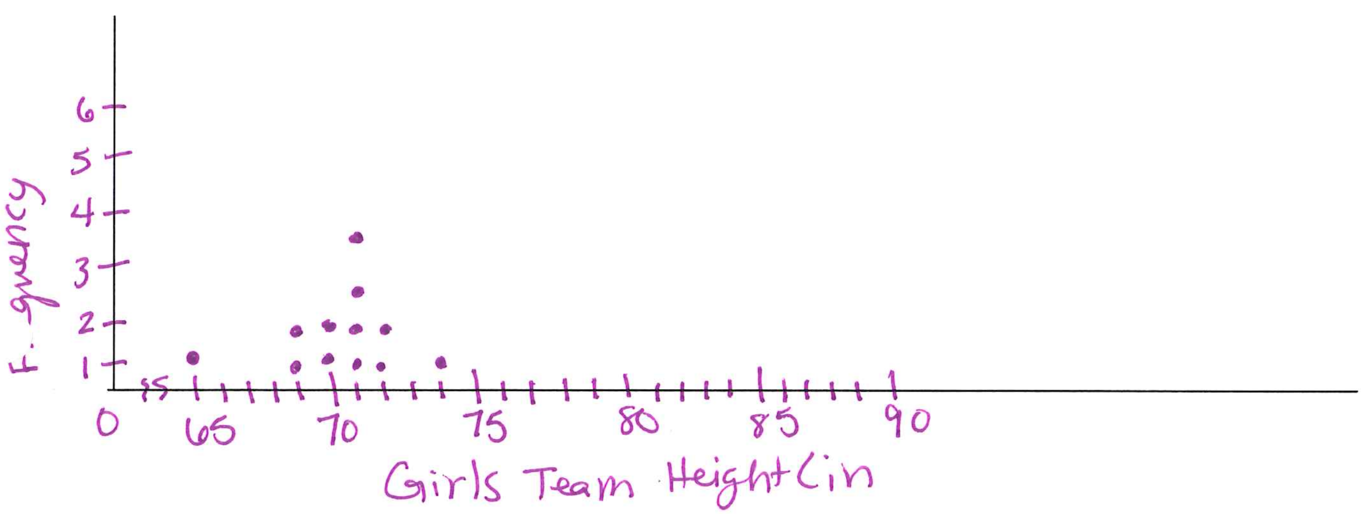
Example 1) The heights (in inches) of each member of the girls' and boys' basketball teams at Holbrook High School are shown below.

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?

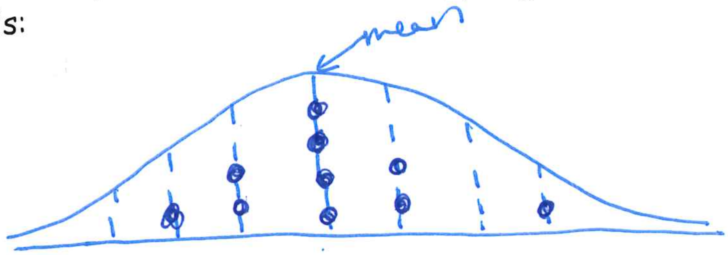
Boy's average height is 74.
Boy's hightare more spread out.
Girl's average height is 71.
Girls hight are closler together
Boys overall height is higher than the girls.

AFM Unit 2, Statistics — Name Key

Notes 2-3, 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:



Draw Dot Plot First

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-shape 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.

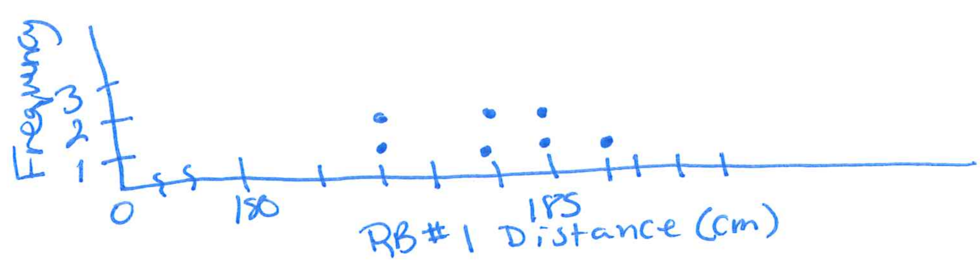
1. The value of the mean, which is a measure of central tendency
2. The value of the standard deviation (SD), which is a measure of spread or spread of the data. (The greater the SD, the greater the spread of the data about the mean.)

Example 1: The Rubber Band Launch (P. 85-86 in Green AA text)

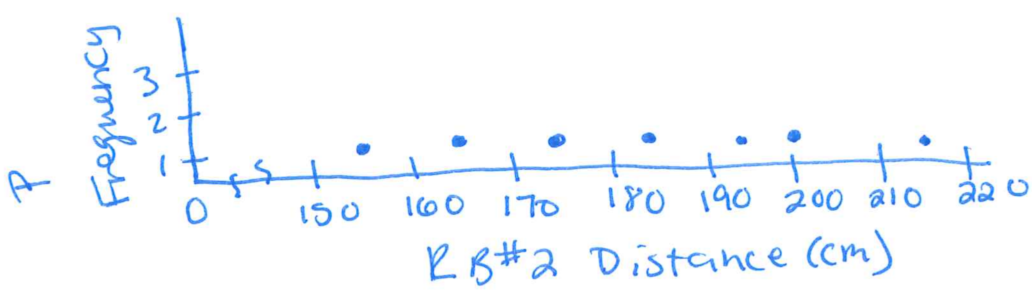
You want to find out how consistently rubber bands will travel when launched, so you use a ruler to launch two rubber bands seven times each. You generate the following data sets:

- 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: 184cm



mean: 184 cm



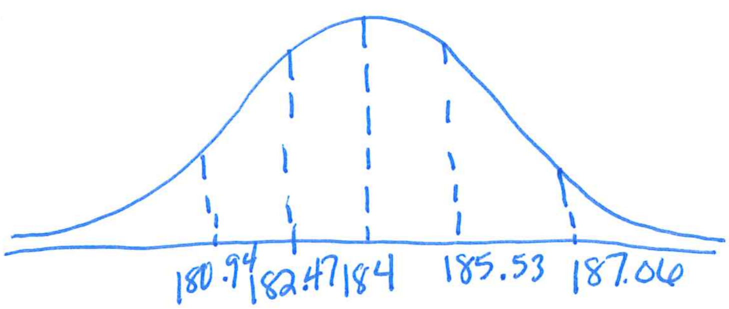
Data Point	Mean	Deviation from Mean	Squared Deviation From Mean
182	− 184	= $(-2)^2$	4
186	− 184	= $(+2)^2$	4
182	− 184	= $(-2)^2$	4
184	− 184	= $(0)^2$	0
185	− 184	= $(+1)^2$	1
184	− 184	= $(0)^2$	0
185	− 184	= $(+1)^2$	+ 1

14 ← Add up the squared deviation

variance = $\dfrac{\text{squared deviation total}}{\#\text{data points} - 1}$

$= \dfrac{14}{7-1} = \dfrac{14}{6} = 2.33$

standard deviation (SD) = $\sqrt{\text{variance}} = \sqrt{2.33} = 1.53$

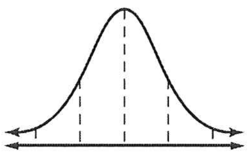
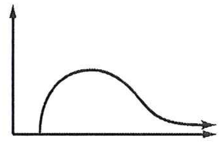
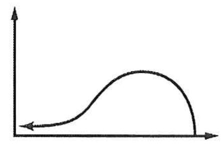


NAME Key DATE ____________ PERIOD _____

12-7 Study Guide and Intervention

The Normal Distribution

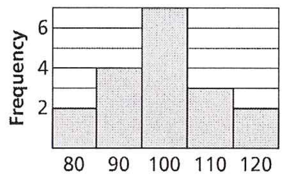
Normal and Skewed Distributions A **continuous probability** distribution is represented by a curve.

	Normal	Positively Skewed	Negatively Skewed
Types of Continuous Distributions			

Example **Determine whether the data below appear to be *positively skewed*, *negatively skewed*, or *normally distributed*.**
{100, 120, 110, 100, 110, 80, 100, 90, 100, 120, 100, 90, 110, 100, 90, 80, 100, 90}

Make a frequency table for the data.

Value	80	90	100	110	120
Frequency	2	4	7	3	2



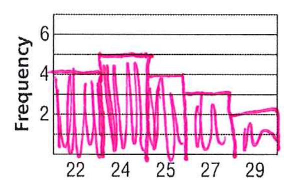
Then use the data to make a histogram.
Since the histogram is roughly symmetric, the data appear to be normally distributed.

Exercises

Determine whether the data in each table appear to be *positively skewed*, *negatively skewed*, or *normally distributed*. Make a histogram of the data.

1. {27, 24, 29, 25, 27, 22, 24, 25, 29, 24, 25, 22, 27, 24, 22, 25, 24, 22}

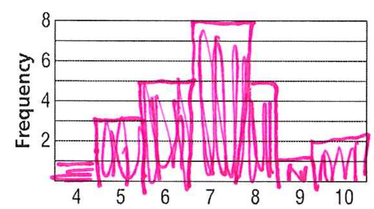
positively skewed



2.

Shoe Size	4	5	6	7	8	9	10
No. of Students	1	2	4	8	5	1	2

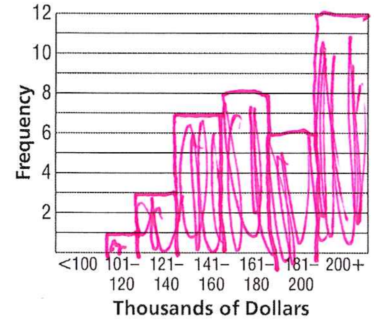
normally distributed



3.

Housing Price	No. of Houses Sold
less than $100,000	0
$100,00–$120,000	1
$121,00–$140,000	3
$141,00–$160,000	7
$161,00–$180,000	8
$181,00–$200,000	6
over $200,000	12

negatively skewed



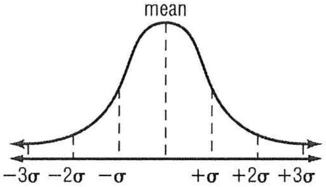
NAME Key DATE ______ PERIOD ______

12-7 Study Guide and Intervention *(continued)*

The Normal Distribution

Use Normal Distributions

Normal Distribution

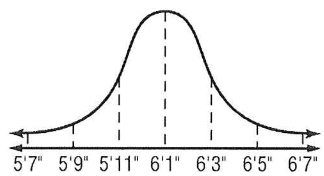


Normal distributions have these properties.
The graph is maximized at the mean.
The mean, median, and mode are about equal.
About 68% of the values are within one standard deviation of the mean.
About 95% of the values are within two standard deviations of the mean.
About 99% of the values are within three standard deviations of the mean.

Example **The heights of players in a basketball league are normally distributed with a mean of 6 feet 1 inch and a standard deviation of 2 inches.**

a. What is the probability that a player selected at random will be shorter than 5 feet 9 inches?

Draw a normal curve. Label the mean and the mean plus or minus multiples of the standard deviation.



The value of 5 feet 9 inches is 2 standard deviations below the mean, so approximately 2.5% of the players will be shorter than 5 feet 9 inches.

b. If there are 240 players in the league, about how many players are taller than 6 feet 3 inches?

The value of 6 feet 3 inches is one standard deviation above the mean. Approximately 16% of the players will be taller than this height.

$240 \times 0.16 \approx 38$

About 38 of the players are taller than 6 feet 3 inches.

Exercises

EGG PRODUCTION The number of eggs laid per year by a particular breed of chicken is normally distributed with a mean of 225 and a standard deviation of 10 eggs.

1. About what percent of the chickens will lay between 215 and 235 eggs per year?
68%

2. In a flock of 400 chickens, about how many would you expect to lay more than 245 eggs per year? 10 chickens

MANUFACTURING The diameter of bolts produced by a manufacturing plant is normally distributed with a mean of 18 mm and a standard deviation of 0.2 mm.

3. What percent of bolts coming off of the assembly line have a diameter greater than 18.4 mm? 2.5%

4. What percent have a diameter between 17.8 and 18.2 mm?
68.3%

AFM ~~MATH III~~ Notes, Unit 1, Statistics — Name ____________

Notes 2-4, Normal Distributions and Percentages — Date ________ Period ________

Normal Distributions and Percentages

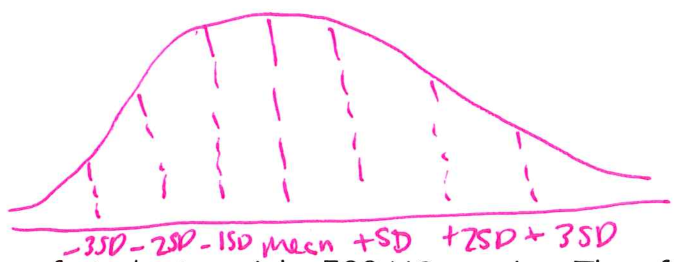
The Empirical Rule - States nearly all values lie within 3 standard deviations from the mean in normal distribution.

In any 68-95-99 data set that is normally distributed:

Approx. 68.3% of the values will be within 1 standard deviation of the mean

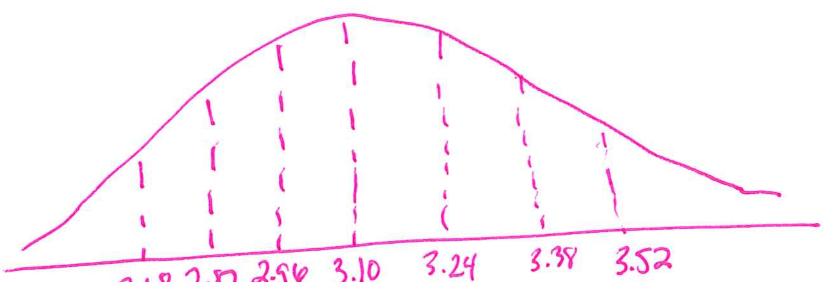
Approx. 95.5% of the values will be within 2 standard deviations of the mean

Approx. 99.7% 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)? 342 pennies

2.82g and 3.38g (i.e., within two standard deviations of the mean)? 478 pennies

2.68g and 3.52g (i.e., within three standard deviations of the mean)? 499 pennies

$$\frac{x}{500} = \frac{\%}{100}$$

What if I wanted to know the percentage of pennies that had a weight between 3g and 3.2g? 52.5%

Calculator Function: **normalcdf()**

The TI83/TI84 calculators have a function called normalcdf() which will tell you: the percentage of values that lie within a given interval
and all you have to give it is: interval (3.0 - 3.2)
mean 3.1
SD 0.14

(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, 0.14)

Percentile Ranks

A percentile is a measure that tells us what percent of the total frequency scored below that measure. A percentile rank is the percentage of scores that fall below a given score.

About Percentile Ranks:

- Percentile rank is a number between Ø and 1ØØ indicatibg the percent of cases falling at or below that score.
- Percentile ranks are usually written to the nearest whole percent percent: 74.5% = 75% = 75th percentile
- Scores are arranged in rank order from lowest to highest.
- There is no Ø percentile rank - the lowest score is at the first percentile.
- There is no 100th percentile rank - the highest score is at the 99th percentile.

Consider:

1. Karl takes the big Earth Science test and his teacher tells him that he scored at the 92nd percentile. What does it mean that he scored in the 92nd percentile?

   Karl did "as well or better than" 92% of students who took the test.

2. Sue takes the Chapter 4 math test. If Sue's score is the same as "the mean" score for the math test, she scored at the 50th percentile. What does this mean?

   Sue did "as well or better than" 50% of students who took the test.

Example 1: If Jason graduated 25th out of a class of 150 students, then 125 students were ranked below Jason. Jason's percentile rank would be:

$\frac{125}{150} = 0.8\bar{3} = 83^{rd}$ percentile

Example 2: The math test scores were: 50, 65, 70, 72, 72, 78, 80, 82, 84, 84, 85, 86, 88, 88, 90, 94, 96, 98, 98, 99. Find the percentile rank for a score of 84 on this test.

normalcdf (-5000, 84, 82.95, 12.18) = 0.534 = 53.4%

[illegible] 6, 88, 88, 90, 94, 96, 98, 98, 99. Find the percentile rank for a score of 86 on this test.

normalcdf (-5000, 86, 82.95, 12.18) = .599 = 59.9%

Example 4: Find the values at the 20th and 80th percentiles for each set of values.

a. 188 168 174 198 186 170 180 182 186 176

174, 188

b. 376 324 346 348 350 352 356 368 345 360

346, 368

Example 1: If Jason graduated 25th out of a class of 150 students, then 125 students were ranked below Jason. Jason's percentile rank would be:

$\frac{125}{150} = 0.8\bar{3} = 83^{rd}$ percentile

Example 2: The math test scores were: 50, 65, 70, 72, 72, 78, 80, 82, 84, 84, 85, 86, 88, 88, 90, 94, 96, 98, 98, 99. Find the percentile rank for a score of 84 on this test.

normalcdf(-5000, 84, 82.95, 12.18) = .534 = 53.4%

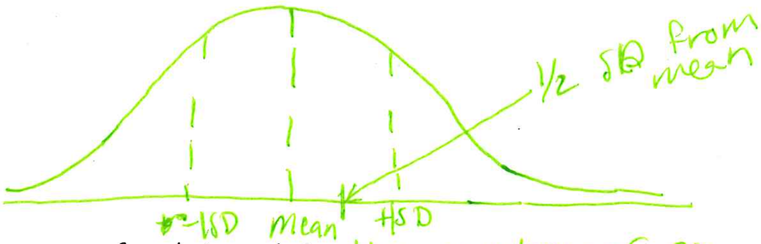
Example 3: The math test scores were: 50, 65, 70, 72, 72, 78, 80, 82, 84, 84, 85, 86, 88, 88, 90, 94, 96, 98, 98, 99. Find the percentile rank for a score of 86 on this test.

AFM, Unit 2 Statistics

Z-scores

Name Key

Date ______________ Period ______________



The z-score of a data point: the number of SDs the point is 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:

$$z = \frac{\text{deviation from mean}}{\text{standard deviation}} = \frac{\text{data point} - \text{mean}}{\text{standard deviation}} = \frac{X - \bar{X}}{S_X}$$

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?

$$z = \frac{2.96 - 3.1}{0.14} = \frac{-.14}{.14} = -1$$

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 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?

$$.64 = \frac{x-3.1}{0.14} \qquad 0.0896 = x - 3.1$$

$$+3.1 \qquad +3.1$$

$$x = 3.1896$$

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

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

$$-0.392 = x - 3.1$$

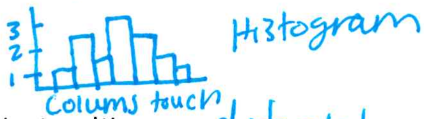
$$+3.1 \qquad +3.1$$

$$x = 2.708$$

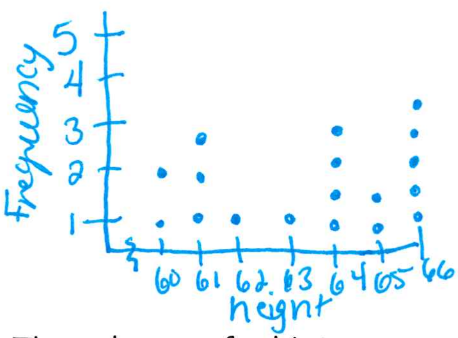
Name Key
Date ____________ Period ____________

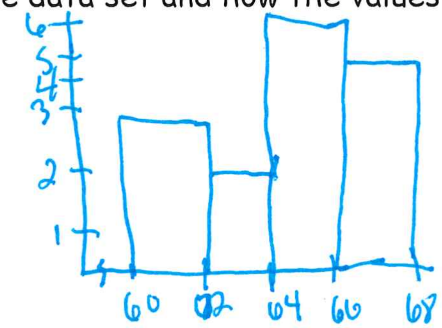
Lesson 2-6- Histograms and Percentile Ranks

Histogram - A way of displaying a distribution that use columns to show how the data points are distributed.



You can think of a histogram as a being like a dot plot, except that it doesn't show every single data point. For this reason, histograms are a good way to display information from very large data sets. Although you can't see individual data values, you can see the shape of the data set and how the values are distributed throughout the range.

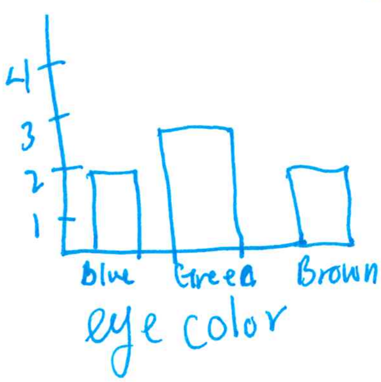


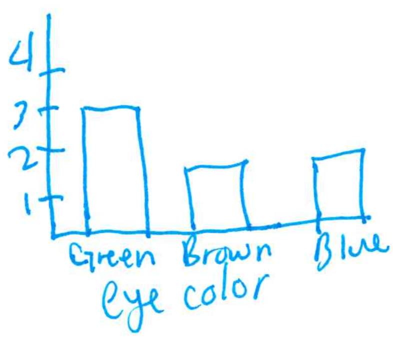


Intervals cannot change!

The columns of a histogram are called bins, which always have the same width. The height of the bins indicates how many data points fall within a given interval.

Note that a histogram is NOT the same as a bar graph. The bars of a bar graph indicate how many data points are in a particular category.





* The order of the categories can change!

All of the bins of a histogram should have the same width.
The bin width may change depending on how much detail you want your histogram to show.

Percentile Rank - For a given distribution, the percentile rank tells the percentage of data values that lie within a given value.

AFM Notes Statistics

Notes 1-7 Samples and Surveys

Name Key

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 - who calls the newspaper. People who call may over represent or underrepresented some views.

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

Convenience sample, since it is convenient for the reporter to stand in one place. May over represent park supporters.

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

Systematic Sample. If there is some link between people who are listed (or not listed) in a phone book and people who pay property tax

Study Methods

1. Observational Study - 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.
4. Simulation

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 reactions (loaded question)
- suggesting that you want a particular answer (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 Question

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

Leading Question

C. Do you think teachers should communicate frequently with students and their parents about class grade?   2 Issues