Section 10-4 Measures of Central Tendency + Variation (Spread)

Once data are collected, two important aspects of the data are center and spread.

Numbers that represent what is "average" or "typical" are called measures of central tendency since they are usually located near the center of a distribution.

The 3 most common measures of central tendency are:

1. Mean 2. Median 3. Mode

1. Mean

This is the most common measure of central tendency.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

sum of data points

\# data points

notation for mean "x bar"

- We sometimes use the words "mean" and "average" interchangeably. Average means typical, so the mean is one measure of average.

- The mean is the balance point for the data and is affected by extreme values (outliers).

2. Median

The median is the middle of the data set when it is ordered (usually from least to greatest.)

- If the number of data points is odd, the median is the middle number. You can find the middle number term by dividing n by 2 + rounding up.

- If the number of data points is even, the median is the mean of the middle 2 numbers. The middle 2 terms are the $\frac{n}{2}^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ terms.

- The median is not affected by outliers. This is why the median is often used to describe the average salary or average home price.

3. $\boxed{\text{Mode}}$

The mode is the easiest measure of average (central tendency) since it is simply the data point (s) that occur the most.

- A data set can have more than one mode. It can also have no mode at all (this occurs when all data points occur the same number of times.)

- The mode is most useful when you have categorical data (as opposed to numerical data.)

- You have to be careful when using the mode to represent the "average" or "typical" data value bc the data point that occurs the most may be an extreme value.

ex 1 Choose the best measure of central tendency.

a) the average home price in a city
median

b) the typical dog breed preferred by people aged 65+
mode bc categorical data

c) the average age of a university president
mean

d) the average nurse makes $80,000 per year
median

ex 2 Find the mean, median, and mode.

11, 6, 4, 0, 2, 1, 12, 0, 0, 2, 10

1. $\boxed{\text{Mean}}$: add them all up ÷ # data points

$48 \div 11 \approx \boxed{4.36}$

2. $\boxed{\text{Median}}$: put in order then find the middle

$0, 0, 0, 1, 2, \boxed{2}, 4, 6, 10, 11, 12$

middle #   Median = 2

* remember if the # data pts is odd, the median is the n ÷ 2 (rounded up) term.

So, $11 \div 2 = 5.5 \rightarrow$ round up 6^{th} term is the median

3. $\boxed{\text{Mode}}$: data point that occurs the most

$\boxed{\text{Mode} = \boxed{0}}$

ex 3 Find the mean, median, and mode. Which one best represents the data?

Home prices in Mathville (in thousands)

~~145~~	~~182~~	~~189~~	~~425~~
~~216~~	~~155~~	~~185~~	~~171~~
~~207~~	~~172~~	~~425~~	~~168~~
~~115~~	~~163~~	~~173~~	~~150~~

1. Mean: sum ÷ # data pts $= \frac{3241}{16} \approx \boxed{\$202.6}$

2. Median: put in order then find the middle

115	163	173	207
145	168	182	216
150	171	185	425
155	172	189	425

Best represents data

Mean of 172 and 173 $\frac{172+173}{2} = \boxed{\$172.5}$

16 ÷ 2 = 8, so, 8th and 9th terms give you median

3. Mode: # occurs most $\boxed{\$425}$

ex 4 Find the mode for each data set.

a) 15, 25, 32, 6, 25, 4, 15 $\boxed{\text{15 and 25}}$
bimodal

b) 5, 10, 15, 20, 25, 30 $\boxed{\text{no mode}}$

Since each # occurs the same # of times

ex 5 Is it possible to have a data set where the mean, median, and mode are all the same number? (without the data set being that number repeated)

20, 30, 30, 30, 40

+10 +10

Mean, Median + Mode are all 30.

Measures of Spread (Variation)

Knowing the average (center) of a data set is not enough to get an accurate picture of the data. To see how scattered (spread) the data is we use methods to measure the spread of data.

Range

The range is the quickest and easiest measure of variation.

Range = highest value − lowest value

The range is not helpful when there are extreme values in the data.

The range isn't enough to measure variability.

Interquartile Range (IQR)

The median splits a set of data in half. The median of the lower half of the data is called the lower quartile (Q_1). The median of the upper half of the data is called the upper quartile (Q_3).

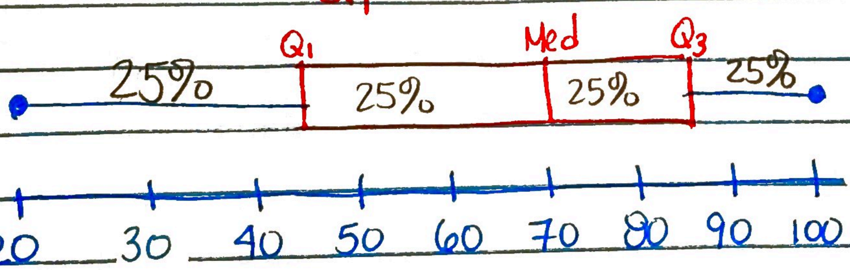
The $IQR = Q_3 - Q_1$. This tells us the spread of the middle 50% of the data values.

We can visualize spread by drawing a box plot (or box and whisker plot).

ex 6 Draw a box plot for the following.

[20, 25, 40, 50, 50, 60], 70, [75, 80, 80, 90, 100, 100]

$\frac{40+50}{2} = 45$ Q_1 Median $\frac{80+90}{2} = 85$ Q_3



If a data point is more than 1.5 times the IQR above the Q_3 (UQ) or below the Q_1 (LQ) the value is said to be an <u>outlier</u>.

<u>ex 7</u> Are there any outliers in the data in ex 6?

$Q_1 = 45$
$Q_3 = 85$

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 85 - 45 \\ &= 40 \end{aligned}$$

above $Q_3 + 1.5 \times IQR = 85 + 1.5(40) = \overset{\text{above}}{145}$

below $Q_1 - 1.5 \times IQR = 45 - 1.5(40) = \overset{\text{below}}{-15}$

No scores above 145 or below −15 so there are no outliers.

ex 8 Are there any outliers in the following data set?

Med $= (62+62)/2 = 62$

[40, 50, 60, 62], [62, 68, 70, 100]

$Q_1 = \frac{50+60}{2} = 55$ $\qquad$ $Q_3 = \frac{68+70}{2} = 69$

Median $= 62$ $\qquad$ $IQR = Q_3 - Q_1$

$Q_1 = 55$ $\qquad\qquad\quad = 69 - 55$

$Q_3 = 69$ $\qquad\qquad\quad = 14$

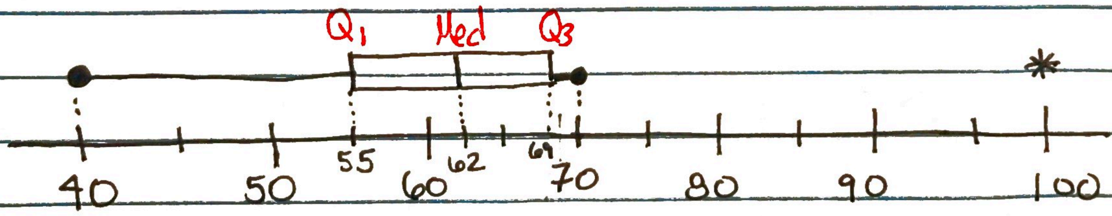
Are any data points:

above $Q_3 + 1.5(IQR) = 69 + 1.5(14) = 90$?

yes! 100 is above 90, so 100 is an outlier

below $Q_1 - 1.5(IQR) = 55 - 1.5(14) = 34$?

no. There are no scores below 34



Standard Deviation

As the name implies, the standard deviation is the "standard" or average of how far each data point deviates from the mean.

$$\sigma = \sqrt{\frac{\overset{\text{sum}}{\sum}(x - \underset{\text{mean}}{\bar{x}})^2}{\underset{\#\text{data points}}{n}}}$$

Sigma (lowercase)

$(\text{stand dev.})^2 = \text{variance}$

* We will use a calculator for this rather than going through all of the computations involved.

ex 9 Use the statistical capabilities of the calculator to find $\bar{x}$ and σ for each data set.

Set A: 18, 19, 20, 20, 26, 28, 30 $\bar{x} = 23$ $\sigma = 4.504$

Set B: 0, 1, 10, 20, 20, 50, 60 $\bar{x} = 23$ $\sigma = 21.706$

For Calculator

* To put data points in a list:

Stat
1. Edit enter

L_1
data pts

* To calculate mean, med, σ, Q_1, Q_3, ...

Quit (2nd mode)

Stat → Calc
1. 1-var stats enter twice

ignore the stuff b/w these →

$\bar{x}$	mean
σx	standard deviation
n	# data points
minx	min data value
Q_1	lower quartile
Med	Median
Q_3	upper quartile
max x	max data value

* We will not cover the Mean Absolute Deviation (MAD) or variance.

Percentiles

A percentile (or percentile rank) of a data point indicates the percent of data values in a set that are below that particular value.

Percentiles are commonly used in pediatrics and standardized tests.

A percentile is a measure of relative standing. It is not the same as a normal percentage.

If your child's birth weight is in the 90^{th} percentile, this means 90% of other baby's birth weights are below that of your baby.

If your ACT score is in the 85^{th} percentile, does this mean you scored an 85% on the ACT?

No!

It means your score is higher than 85% of other people who took that same ACT.

A percentile rank tells you nothing about your score - it only tells you how you compare to others.

ex 10 Elena ranked 100th out of her graduating class of 600 and Maria ranked 150th out of her graduating class of 750. Find each student's percentile rank.
↳ % below

Elena: # students below her 600-100=500

So, $\frac{500}{600} = 0.8\underbrace{33}3 = 83.3\%$ below her
below, total
So, 83rd percentile

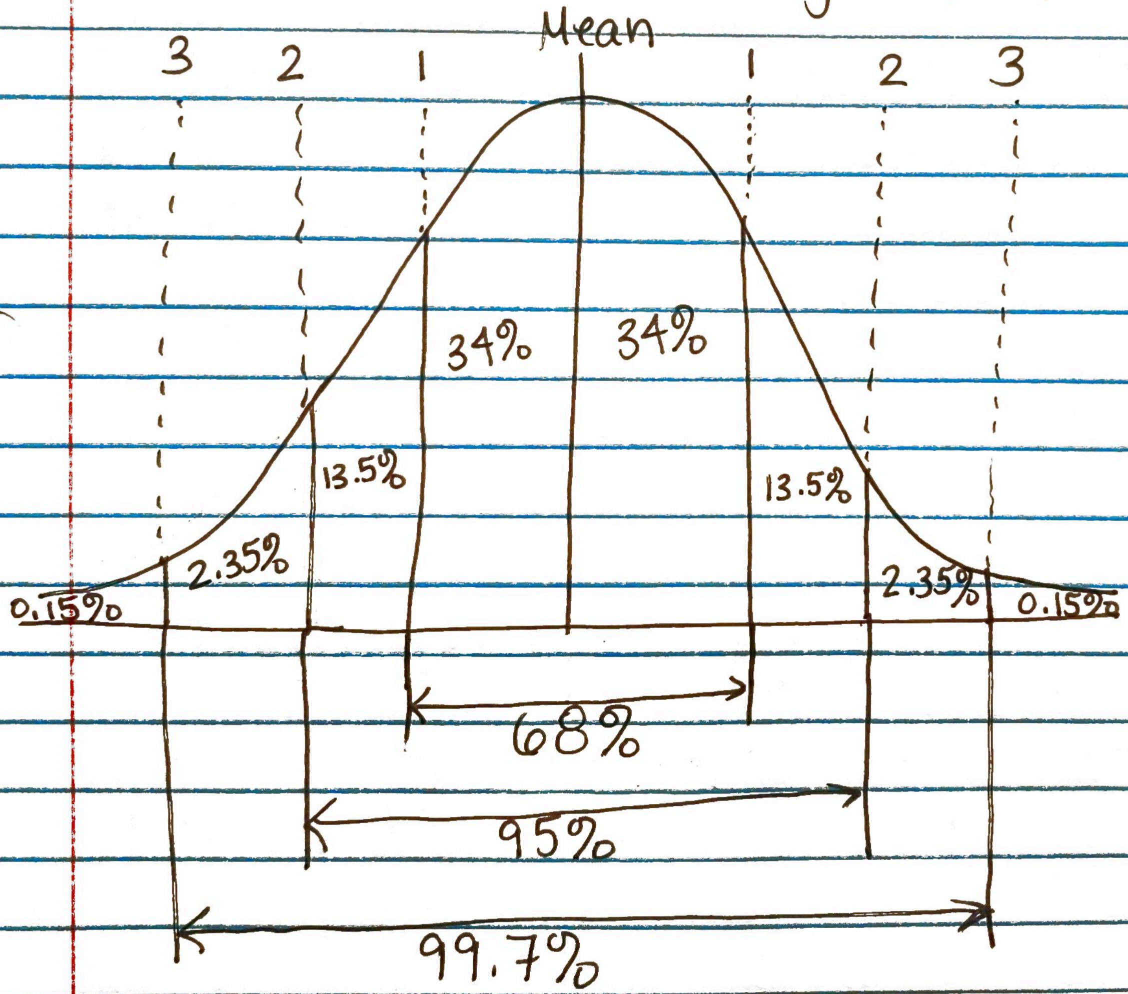
Maria: # students below her 750-150=600

So, $\frac{600}{750} = 0.\underbrace{8} = 80\%$ below her
below, total
So, 80th percentile

Elena: 83rd percentile
Maria: 80th percentile

Normal Distribution

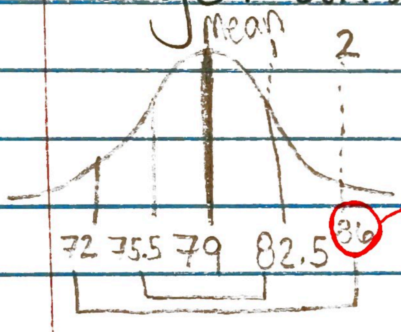
A normal distribution is a continuous, symmetric, bell-shaped curve. We call it "normal" since it occurs so frequently with data like: heights, IQ, life span, wait times + hours per night people sleep.



For Normal Distributions:

68% of data lie within 1 st. dev of the mean.
95% of data lie within 2 st. dev of the mean.
99.7% of data lie within 3 st. dev of the mean.

<u>ex 11</u> On an exam, the mean is 79 and the standard deviation is 3.5. If a grade of A is given to students who score 2 standard dev above the mean (or higher), what is the lowest score needed to get an A?



86 or higher gets an A

total!

<u>ex 12</u> The heights of 8585 men are normally distributed with a mean of 68 in and a standard deviation of 3 inches.

5'8"

a) What percent of men are bw 5'5" and 5'11"?

68%

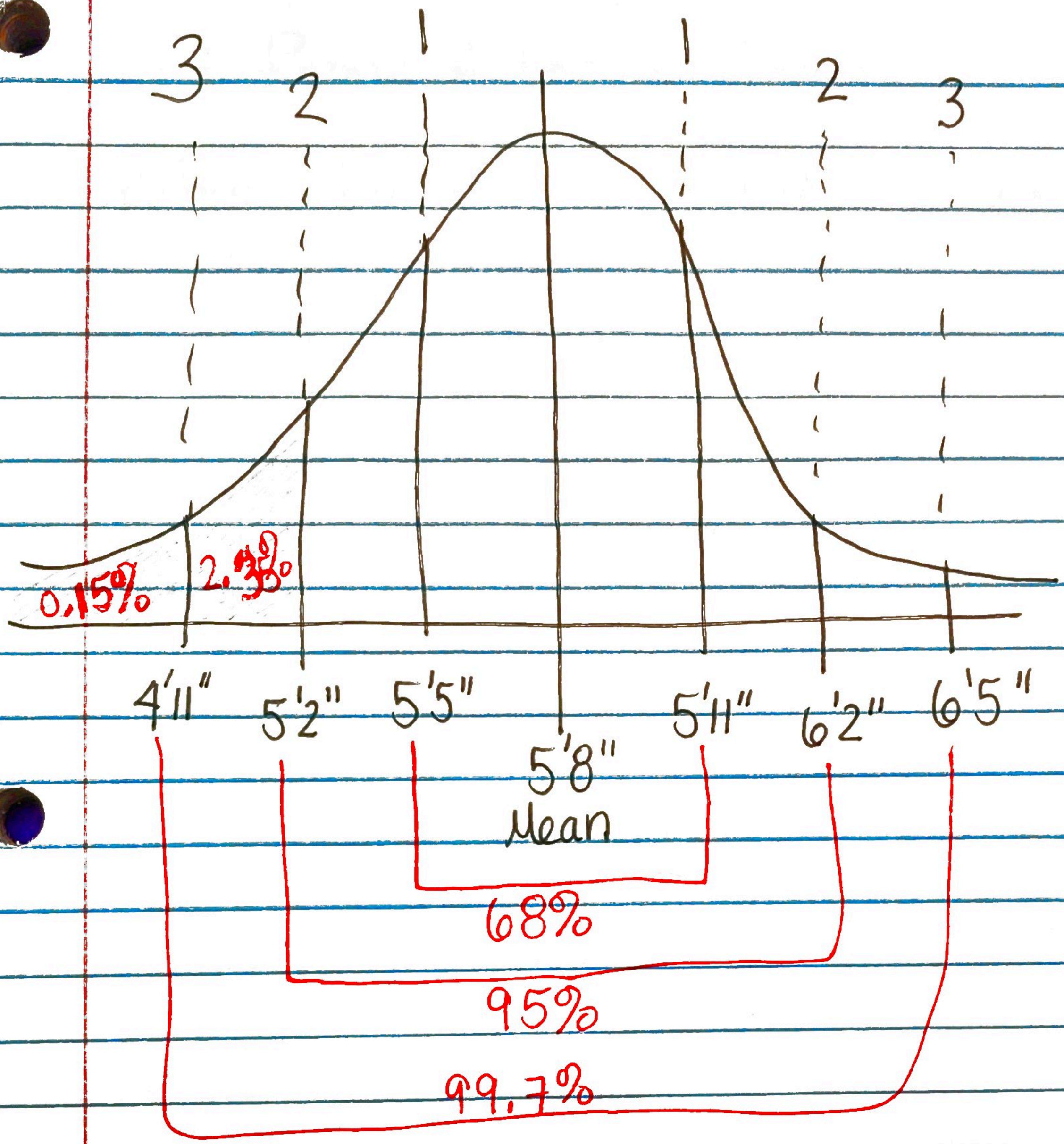
b) How many men are bw 5'2" and 6'2"?

95% of 8585 = (0.95)(8585) = 8156

c) How many men are less than 5'2"?

0.15% + 2.35% = 2.5% of 8585

(0.025)(8585) = 215



end 10-4

Section 10-4
Math 206