5 FOR 5

AP Stat **Lesson 1.3 – Describing with Quantitative Data** Name:

4 steps to solving problems: 1) State 2) Plan 3) Do 4) Conclude

How do you find the mean?

add all up / total #

What are the two notations for mean?

Sample mean $\bar{x}$ Population mean μ

Let's find the mean for the following stem – plot using the calculator.

Stem and Leaf Plot

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

Key: 4|1 is a runner who races a 41 minute 10k.

$\bar{x} = 69.5$

Calc steps:

$Stat \rightarrow 4 \rightarrow 2nd\ 1\ enter \rightarrow Stat \rightarrow edit \rightarrow type\ in\ each\ value \rightarrow Stat \rightarrow calc \rightarrow 1-var\ stat \rightarrow 2nd\ 1\ (L1\ in\ List) \rightarrow Freq\ List\ leave\ blank \rightarrow Enter \rightarrow Calculate \rightarrow x\ is\ the\ mean$

Calculate the mean again without the outlier of 95. What do you notice?

$\bar{x} = 67.54$ went down

What does the mean mean? Is the mean a resistant measure? → no, outliers move it

How large each observation would be if the total were split equally among all observations. "Fair share"

With the same numbers from above, find the median.

70

What is the median? Describe it. How do you find it?

middle # ½ way thru your #'s, the #'s have to be in ORDER

2 # → add / 2

Is the median a resistant measure? What does the median stand for in any context?

Yes → outliers don't move it.

Midpoint, ½ way

When a distribution is roughly symmetric, what do you know about the mean and median?

Close together

When the distribution is exactly symmetric, what do you know about the mean and the median?

equal – the same

In a skewed distribution, what do we know about the mean and the median? Draw a picture below.

left

mean < med

outliers pull mean down

Right

mean > med

outliers pull mean up

Let's measure **spread**/variability.

Given the following 18 quiz scores, find the following values.

20 55 60 60 65 65 70 70 70 75 80 80 80 85 85
85 90 100

Q1

65

Median left (lower) side data

Q3 85

median right (upper) side data

IQR Q3 − Q1

20

Middle 50%

Context

IQR measures middle 50% of quiz scores. The range of scores varies by 20.

Q1 → lower 25% of the data falls below 65

Q3 → upper 25% of the quiz score that falls above 85.

Outliers – we have a formula!

Lower fence: Q1 – 1.5(IQR)

Upper fence: Q3 + 1.5(IQR)

See if the above example has any outliers. Show the proof below.

$65 - 1.5(20)$ LF 35 — 20 outlier

$85 + 1.5(20)$ UF 115 — none

Now let's make a **box plot**. Below is the data for the Atlanta Falcons football team and the weight in pounds of the top 10 linemen on the team.

338 318 353 313 318 326 307 317 311 250

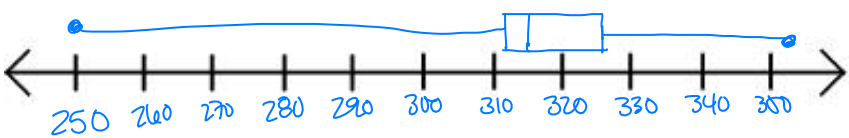
Find the 5 number summary first.

Min	Q1	Med	Q3	Max
250	311	317.5	326	353

Then go to stat plot and turn on the box and whisker plot without outliers. Copy it below and label correctly.

w/o outliers

Skew left



w/outliers

Now turn on the box and whisker plot with outliers. This is called a modified box and whisker plot. Copy it above and then jot down the 5 number summary. Did the numbers change at all from above?

Min	Q_1	Med	Q3	Max
307	311	317.5	326	338

Outliers → 250 & 353

Min & Max (Range) got smaller

What do you notice about both graphs? What changed from one graph to the next?

Min & Max got smaller, less varied from top to bottom

Q1, Med, Q3 → resistant to outliers

Standard deviation

The data provided is the number of pets owned by a group of 8 adults.

3 4 1 5 7 4 8 9

Find the mean of the data:

$\bar{x} = 5.125$

Write each value in the value column.

Take the value – mean in the 2nd column

Square that number to get the 3rd column

Add up the 3rd column and put that in the total. Count up the number of values, take 1 less than it and divide the total by it. That is the **variance** – average squared distance from the mean

Value	Distance from mean	(Distance from mean)^2
3	-2.125	4.516
4	-1.125	1.266
1	-4.125	17.016
5	-0.125	0.016
7	1.875	3.516
4	-1.125	1.266
8	2.875	8.266
9	3.875	15.016
	Total:	50.878
	Average (Distance from mean)^2:	$\frac{50.878}{7}$

pos

7.268 – variance

Variance – take the square root – **Standard deviation** – measures the ~~average distance of the values~~ from the mean.

$\sqrt{7.268} = 2.696$

^{2}mean=5 ^{8}

Now let's do it in the calc and see if we get the same thing. Type the values into L1 and follow the steps like we did before when we found 1 – variable stats. Scroll around. Look for the two symbols for standard deviation.

Population standard deviation ↙ less variable Sample standard deviation

$\sigma_x = 2.522$ $S_x = 2.696$

- Should only use std. deviation as a measure of spread when the mean is chosen as the center
- The sample std. deviation is always greater than or equal to 0.
- More spread out the values are, the ↑ std. dev – more variability
- The standard deviation is not resistant to outliers.

Which measure of center is best?

Median	IQR	Mean	Standard deviation
skewed (outliers)	skewed	symmetric	symmetric

Let's look at this data and see what we can figure out.

In 2017, AP Stat students asked "Who snaps more, male or females?" They asked a simple random sample of students from their school to record the number of snaps sent and received in a 2 week period.

Males	127	44	28	83	0	6	78	6	5	213	73	20	214	28	11	
Females	112	203	102	54	379	305	179	24	127	65	41	27	298	6	130	0

What conclusions should the students draw? Give appropriate evidence to support your claims.

Males $\bar{x}=62.4$ $s=71.4$ Min 0 Q1 6 Med 28 Q3 83 Max 214

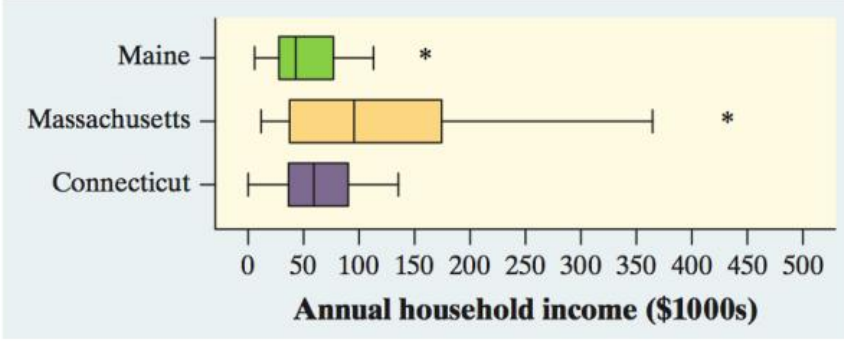
Females snap more

Females $\bar{x}=128.3$ $s=115.92$ Min 0 Q1 34 Med 107 Q3 191 Max 379

Both tend to have a slightly skewed right shape. We would compare the median and the IQR to show spread. On average, it appears females Snap more. Over 50% of the females snap more than 75% of the males.

Check Your Understanding:

The following boxplots show the total income of 40 randomly chosen households each from Connecticut, Maine, and Massachusetts, based on U.S. Census data from the American Community Survey. Compare the distributions of annual incomes in the three states.



The shape of the distribution of Connecticut is roughly symmetric while Maine incomes are slightly skewed right and Massachussettes has a stronger right skew.

The center is the highest for Massachussettes, then Connecticut then Maine.

The variability is the highest four Mass with Maine and Connec having similar spread.

Main and Mass have a higher outlier, where Connec appears to have no outlier for income.

Why do we divide by n-1 in Std. dev and variance?

* unbiased result
* If we just divide by n → biased information

What is discrete data?

dots, things that stop
People

Continuous data?

Connected
never stops
have decimals
Time, measurement