AP Stats
Unit 01 – Univariate Data
Day 3 Notes

Name Key

Measures of Center, Measures of Spread, Outliers, Linear Transformations, and Comparing Distributions

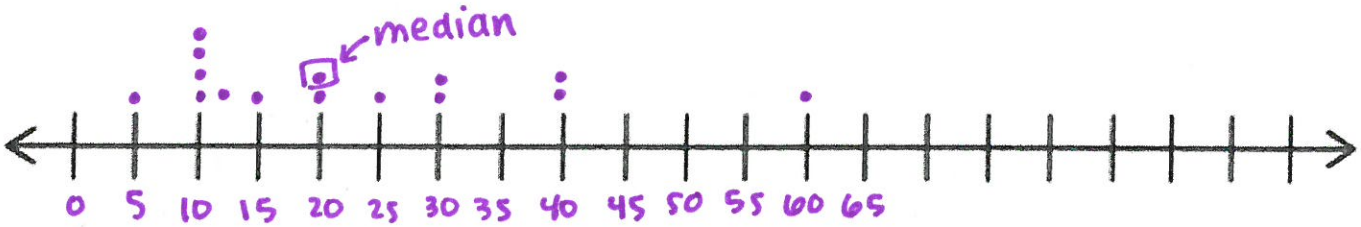
The following data are travel times for fifteen people to get to work:

20 30 10 40 25 20 10 60 15 40 5 30 12 10 10

Rewrite the numbers in order from least to greatest:

5 10 10 10 10 12 15 20 20 25 30 30 40 40 60
min, Q1, median, Q3, max

Make a dotplot of the data for a visual representation:



Measures of Center:

Median: the middle # when written in order
→ if odd data values, middle # is median
→ if even data values, average 2 middle #s for median

ex: 20 is the middle #

Mean: the average of the data

$\bar{x}$

$$\bar{x} = \frac{\sum x_i}{n}$$

ex: $\frac{337}{15} \approx 22.46$ minutes

Measures of Spread:

5 # summary with IQR:
min: smallest # (5)
Q_1: between min and median (10)
median: middle # when written in order
Q_3: between median and max (30) (20)
max: largest #
IQR: range of middle 50% of data
$IQR = Q_3 - Q_1$ $30 - 10 = 20$
range: max − min

Standard Deviation: looks at how far the observations are from the mean.
↳ 2 types:
σ_x = population is known
S_x = sample of a population
↳ paired w/ mean when describing data
→ $S_x \geq 0$
→ same units as mean
→ $S_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$

Comparing the Mean and Median:

The mean and median of a roughly symmetric distribution are close together. If the distribution is exactly symmetric, the mean and median are exactly the same. In a skewed distribution the mean is usually farther out in the long tail than its median. If the outliers were to increase, it would increase the mean but the median would stay the same.

ex: In the data above, $\bar{x} = 22.46$ and the median is 20 which means that the data is skewed right because the mean is closer to the "tail" end.

Standard Deviation by hand:

Given the data set of: 1, 3, 4, 4, 8 find the standard deviation by hand.

① Find $\bar{x} = \frac{1+3+4+4+8}{5} = 4$

② Data #	③ Deviation	④ Deviation^2
1	$1-4=-3$	$(-3)^2=9$
3	$3-4=-1$	$(-1)^2=1$
4	$4-4=0$	$0^2=0$
4	$4-4=0$	$0^2=0$
8	$8-4=4$	$4^2=16$
	+ / 0	+ / ⑤ Sum = 26

⑥ $\frac{\text{sum}}{n-1} = \frac{26}{5-1} = \frac{26}{4} = 6.5$

⑦ $\sqrt{6.5} \approx 2.55$

Roughly the average distance of the values from the mean.

Standard deviation/1 variable statistics in calculator:

1.2 Describing Distributions with Numbers

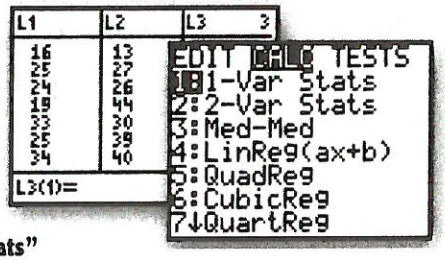
When you first encounter a dataset, it is a good habit to study a graphical display and estimate the SOCS. However, for a more detailed understanding of data, we must calculate numeric summaries of the center and spread. *Note: Be sure you understand how the following measures are calculated before relying on the TI to do the mechanics for you.*

The most common measures of center for a dataset are **mean** ($\bar{x}$) and **median** (Q2). The most common measures of spread/variability for a dataset are **range** (max-min), **interquartile range "IQR"** (Q3-Q1), and **standard deviation** (s_x).

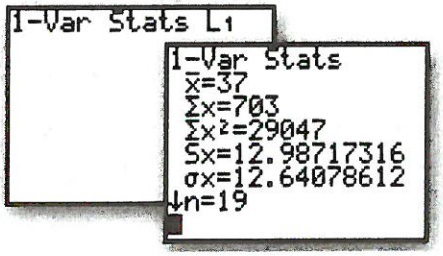
Calculating Numeric Summaries

The calculation of each of these measures, especially the standard deviation, can be quite tedious. Thankfully, the TI can automate those calculations for us. Like plotting data, the calculator requires that you enter the dataset before it can report a numeric summary. If you haven't done so already, enter the Bonds and Aaron data into [STAT] **Edit…** **L1** and **L2**, respectively.

1. Enter data in to into [STAT] **1:Edit…**
2. Press [STAT] **CALC** **1:1-Var Stats** [ENTER]

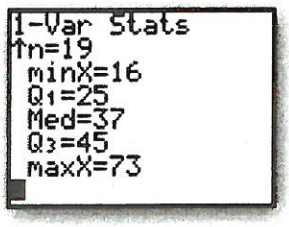


3. Your homescreen should read **"1-Var Stats"**
4. Press **2ND** [1] **(L1)** [ENTER]
5. A numeric summary of the Bonds data should appear.
6. Repeat Steps 2 through 4 for **L2** to get a numeric summary of the Aaron data.



7. Scroll down on each numeric summary to see the 5-number summary.

Remember to interpret the numeric summary in the context of the problem!



SD:

the closer to zero your SD is, the less variation it has.

ex: 5, 5, 5, 5, 5

$\bar{x}=5$

SD = 0 no variation

* always include range

WHICH MEASURE OF CENTER/SPREAD DO YOU USE?

	Reasonably Symmetric/No Outliers	Strongly Skewed or has Outliers
Center	mean ($\bar{x}$)	median
Spread	standard deviation	IQR
Outliers	2 x SD	1.5 x IQR

Determine if there are outliers from the problem on the previous page.

Outliers:

① $IQR = Q_3 - Q_1 =$
$30 - 10 = 20$

② $1.5 \times 20 = 30$

③ $Q_1 - 1.5IQR$
$10 - 30 = -20$

$Q_3 + 1.5IQR$
$30 + 30 = 60$

So if there were any values below -20 or above 60, they would be outliers.

Ex. The follow data compares the amount of text messages that males and females send in two days:

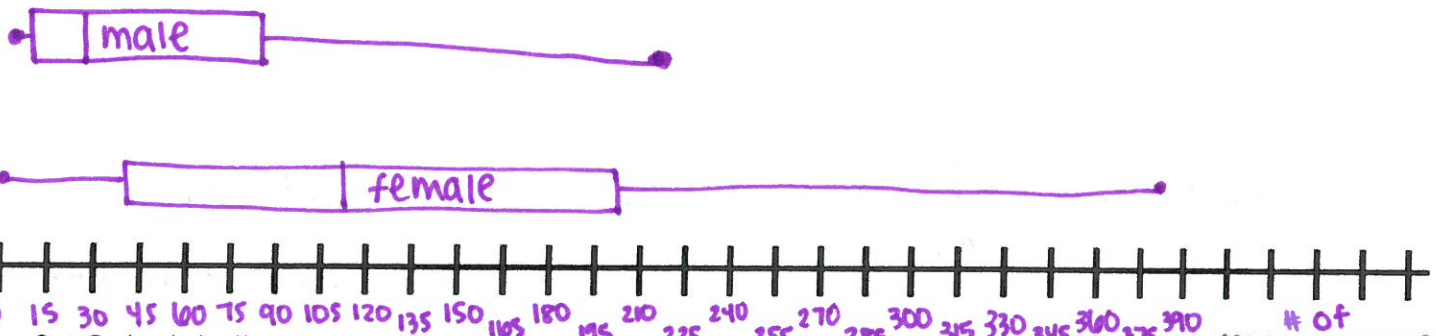
Males	127	44	28	83	0	6	78	6	5	213	73	20	214	28	11
Females	112	203	102	54	379	305	179	24	127	65	41	27	298	0	130

Question:

Does the data give convincing evidence that the females text more than males?

1. Create a side by side box plot comparing the two data sets

texting



2. Calculate the one-variable stats

	Mean	SD	Min	Q1	Median	Q3	Max	IQR
Male	62.4	71.37	0	6	28	83	214	77
Female	136.4	115.18	0	41	112	203	379	162

3. Which "center of spread" test should best be used to describe this data and why?

due to the fact that the data is skewed and has outliers, we are going to use the median, IQR, and range.

4. Finally compare SOCS to determine of the data is proof that females text more.

S: both skewed right

O: male = 213, 214 female = no outliers

C: females text more than males. female median = 112 is 4x larger than males(28). female median is higher than Q_3 males, meaning over 75% of males text less than median females.

S: Sx is higher for females. Female IQR = 162 is more than double males(77). m range = 214, f range = 379

Changing the Unit of Measurement:

A linear transformation changes the original variable x into the new variable x_{new} by $x_{new} = a + bx$ (changes size of unit; shifts ↑↓)

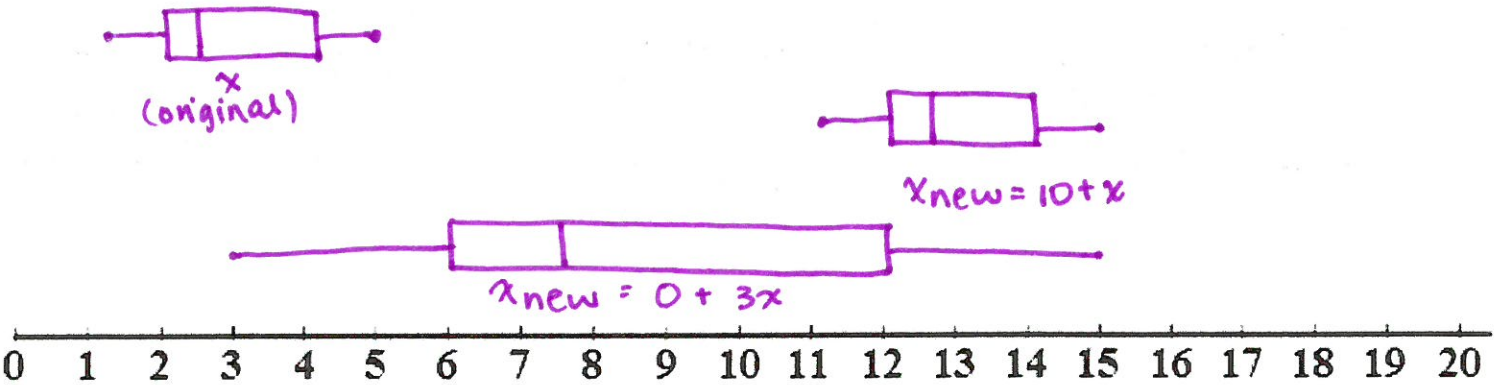
The following data represents the number of people who ran for senior class president in the past 10 years:

1, 5, 3, 2, 4, 4, 2, 2, 1, 3

1 1 2 2 2 : 3 3 4 4 5 (2.5)

$S_x = 1.337$

1. Create a box plot and calculate mean and standard deviation for this data.



2. Now let's say that we add 10 to each number. What would this do to the box plot? Compare this box plot, mean, and standard deviation with the one above (graph on same number line above).

11 11 12 12 12 : 13 13 14 14 15 (12.5)

$S_x = 1.337$

3. Go back to the original data and multiply each number by 3. What would this do to the box plot? Compare this box plot, mean, and standard deviation with the ones above (graph on the same number line above).

3 3 6 6 6 : 9 9 12 12 15 (7.5)

$S_x = 4.0124$

($S_x \times 3 = 1.337 \times 3 = 4.0124$)

WHAT DO YOU NOTICE?

→ adding the same # (+/−) to each data value adds "a" to the measure of center (mean/med) and to quartiles, but not to SD.

→ multiplying (× or ÷) to each data value multiplies "b" by measure of center (mean/med) and measures of spread (SD & IQR).

Resistance:

ability to resist the influence of extreme observations

Resistant measures: median, IQR

Non-resistant measures: mean, SD, range