B – Graphs and Statistics, Lesson 2, Central Tendency and Dispersion (r. 2018)

GRAPHS AND STATISTICS

Central Tendency and Dispersion

Common Core Standards	Next Generation Standards
S-ID.A.2 Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (inter-quartile range, standard deviation) of two or more different data sets.	**AI-S.ID.2** Use statistics appropriate to the shape of the data distribution to compare center (median, mean) and spread (inter-quartile range, sample standard deviation) of two or more different data sets. Note: Values in the given data sets will represent samples of larger populations. The calculation of standard deviation will be based on the sample standard deviation formula $s=\sqrt{\frac{\left(x-\overline{x}\right)^2}{n-1}}$. The sample standard deviation calculation will be used to make a statement about the population standard deviation from which the sample was drawn.
S-ID.A.3 Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).	**AI-S.ID.3** Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).

LEARNING OBJECTIVES

Students will be able to:

1) Calculate measures of central tendency and dispersion for one variable data sets from a graphic representation of the data set, a table, or a context.
2) Compare measures of central tendency and dispersion for two or more one variable data sets.

Overview of Lesson

Teacher Centered Introduction	Student Centered Activities
Overview of Lesson **- activate students' prior knowledge** **- vocabulary** **- learning objective(s)** **- big ideas: direct instruction** **- modeling**	**guided practice ←Teacher: anticipates, monitors, selects, sequences, and connects student work** **- developing essential skills** **- Regents exam questions** **- formative assessment assignment (exit slip, explain the math, or journal entry)**

VOCABULARY

Center (measures of central tendency)
Mean
Median
Mode
Spread (measures of dispersion)
Interquartile Range

Standard Deviation
Normal Curve

Outliers (extreme data points)

BIG IDEAS

Measures of Central Tendency

A **measure of central tendency** is a *summary statistic* that indicates the typical value or center of an organized data set. The three most common measures of central tendency are the *mean*, *median*, and *mode*.

Mean A measure of central tendency denoted by $\bar{x}$, read "x bar", that is calculated by adding the data values and then dividing the sum by the number of values. Also known as the arithmetic mean or arithmetic average. The algebraic formula for the mean is:

$$\text{Mean} = \frac{Sum\ of\ \ items}{Count} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

Median A measure of central tendency that is, or indicates, the middle of a data set when the data values are arranged in ascending or descending order. ***If there is no middle number, the median is the average of the two middle numbers.***

Examples:
The **median** of the set of numbers: {2, 4, 5, 6, 7, 10, 13} is 6
The **median** of the set of numbers: {6, 7, 9, 10, 11, 17} is 9.5

Quartiles:
Q1, the **first quartile**, is the middle of the lower half of the data set.
Q2, the **second quartile**, is also known as the **median**.
Q3, the **third quartile**, is the middle of the upper half of the data set.

NOTE: To computer Q1 and Q2, find the middle numbers in the lower and upper halves of the data set. The median itself is not included in either the upper or the lower halves of the data set. When the data set contains an even number of elements, the median is the average of the two middle numbers and is excluded from the lower and upper halves of the data set.

Mode A measure of central tendency that is given by the data value(s) that occur(s) most frequently in the data set.

Examples:
The **mode** of the set of numbers {5, 6, 8, 6, 5, 3, 5, 4} is 5.
The **modes** of the set of numbers {4, 6, 7, 4, 3, 7, 9, 1,10} are 4 and 7.
The **mode** of the set of numbers {0, 5, 7, 12, 15, 3} is none or there is no mode.

Measures of Spread

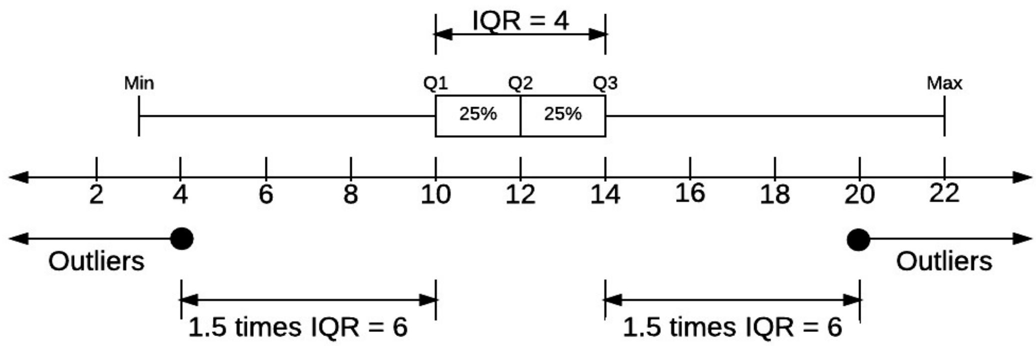
Measures of Spread indicate how the data is spread around the center of the data set. The two most common measures of spread are interquartile range and standard deviation.

<u>Interquartile Range</u>: The difference between the first and third quartiles; a measure of variability resistant to outliers.

$$IQR = Q3 - Q1$$

<u>Outlier</u> An observed value that is distant from other observations. Outliers in a distribution are 1.5 interquartile ranges (IQRs) or more below the first quartile or above the third quartile.

An **<u>outlier</u>** can significantly influence the measures of central tendency and/or spread in a data set.



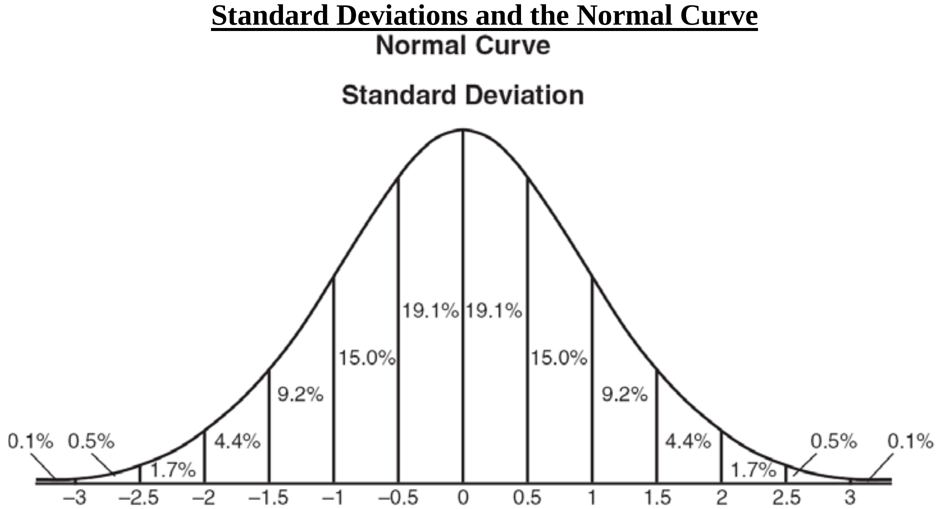
<u>Example</u>

In the above example, **<u>outliers</u>** would be any observed values less than or equal to 4 and/or any observed values greater than or equal to 20.

NOTE: Box plots, like the one above, are useful graphical representations of dispersion.

<u>Standard Deviation</u>: A measure of variability. **<u>Standard deviation</u>** measures the average distance of a data element from the mean. Typically, 98.8% of any set of univariate data can be divided into a total of six standard deviation units: three standard deviation units above the mean and three standard deviation units below the mean.

<u>Standard Deviations and the Normal Curve</u>



- **When a data set is normally distributed, there are more elements closer to the mean and fewer elements further away from the mean.**
- **The normal curve shows the distribution of elements based on their distance from the mean.**
- **Three standard deviation units above the mean and three standard deviation units below the mean will include approximately 98.8% of all elements in a normally distributed data set.**
- **Each standard deviation above or below the mean corresponds to a specific value in the data set.**
    - **In the above example, the distance associated with each standard deviation unit corresponds to a distance of approximately $2\frac{2}{3}$ units on the scale below the curve.**
- **Many things in nature, such as height, weight, and intelligence, are normally distributed.**

There are two types of **<u>standard deviations</u>**: population and sample.

<u>Population Standard Deviation</u>: If data is taken from the *entire population*, divide by n when averaging the squared deviations. The following is the formula for **<u>population standard deviation</u>**:

$$\sigma = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n}}$$

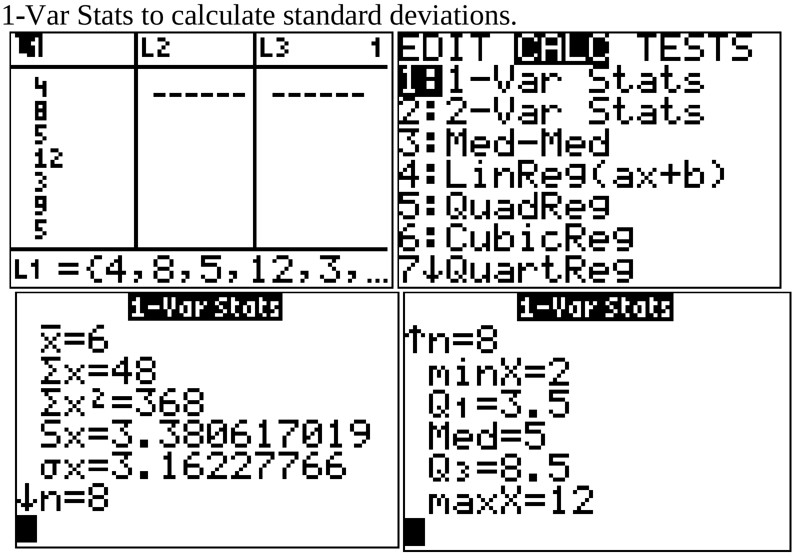
NOTE: Population standard deviation not included in Next Generation Standards.

<u>Sample Standard Deviation</u>: If data is taken from a *sample* instead of the *entire population*, divide by $n-1$ when averaging the squared deviations. This results in a *larger* standard deviation. The following is the formula for **<u>sample standard deviation</u>**:

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$$

<u>Tips for Computing Measures of Central Tendency and Dispersion</u>:
Use the STATS function of a graphing calculator to calculate measures of central tendency and dispersion.
INPUT VALUES: {4, 8, 5, 12, 3, 9, 5, 2}

1. Use STATS EDIT to input the data set.
2. Use STATS CALC 1-Var Stats to calculate standard deviations.



The outputs include:
$\overline{X}$, which is the mean (average), $\sum x$, which is the sum of the data set.

$\sum x^2$, which is the sum of the squares of the data set.
Sx, which is the **sample** standard deviation.
σx, which is the **population** standard deviation.
n, which is the number of elements in the data set

minX, which is the minimum value
Q1, which is the first quartile
Med, which is the median (second quartile)
Q3, which is the third quartile
maxX, which is the maximum value

DEVELOPING ESSENTIAL SKILLS

Use a graphing calculator to calculate one variable statistics for the following data sets:

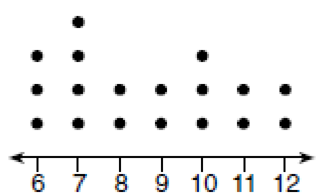
Set A

0.5	0.5	0.6	0.7	0.75	0.8
1.0	1.0	1.1	1.25	1.3	1.4
1.4	1.8	2.5	3.7	3.8	4
4.2	4.6	5.1	6	6.3	7.2

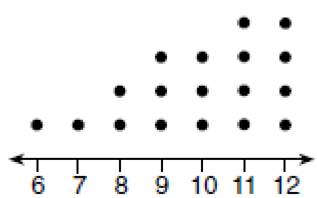
Set B

Number of Candy Bars Sold				
0	35	38	41	43
45	50	53	53	55
68	68	68	72	120

Sets C and D



Soccer Players' Ages



Basketball Players' Ages

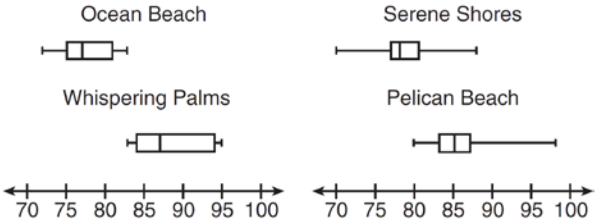
REGENTS EXAM QUESTIONS

S.ID.A.2-3: Central Tendency and Dispersion

4) Christopher looked at his quiz scores shown below for the first and second semester of his Algebra class.
Semester 1: 78, 91, 88, 83, 94
Semester 2: 91, 96, 80, 77, 88, 85, 92
Which statement about Christopher's performance is correct?

1) The interquartile range for semester 1 is greater than the interquartile range for semester 2.
2) The median score for semester 1 is greater than the median score for semester 2.
3) The mean score for semester 2 is greater than the mean score for semester 1.
4) The third quartile for semester 2 is greater than the third quartile for semester 1.

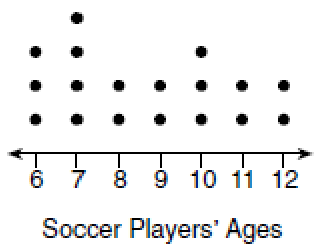
5) Corinne is planning a beach vacation in July and is analyzing the daily high temperatures for her potential destination. She would like to choose a destination with a high median temperature and a small interquartile range. She constructed box plots shown in the diagram below.

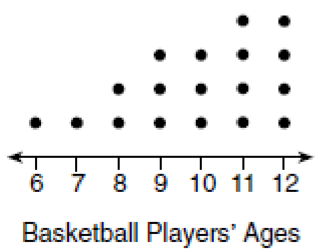


Which destination has a median temperature above 80 degrees and the smallest interquartile range?

1) Ocean Beach
2) Whispering Palms
3) Serene Shores
4) Pelican Beach

6) Noah conducted a survey on sports participation. He created the following two dot plots to represent the number of students participating, by age, in soccer and basketball.





Which statement about the given data sets is correct?

1) The data for soccer players are skewed right.
3) The data for basketball players have the same median as the data for soccer players.

2) The data for soccer players have less spread than the data for basketball players.

4) The data for basketball players have a greater mean than the data for soccer players.

7) The two sets of data below represent the number of runs scored by two different youth baseball teams over the course of a season.

Team *A*: 4, 8, 5, 12, 3, 9, 5, 2
Team *B*: 5, 9, 11, 4, 6, 11, 2, 7

Which set of statements about the mean and standard deviation is true?

1) mean A < mean B
standard deviation A > standard deviation B

3) mean A < mean B
standard deviation A < standard deviation B

2) mean A > mean B
standard deviation A < standard deviation B

4) mean A > mean B
standard deviation A > standard deviation B

8) Isaiah collects data from two different companies, each with four employees. The results of the study, based on each worker's age and salary, are listed in the tables below.

Company 1

Worker's Age in Years	Salary in Dollars
25	30,000
27	32,000
28	35,000
33	38,000

Company 2

Worker's Age in Years	Salary in Dollars
25	29,000
28	35,500
29	37,000
31	65,000

Which statement is true about these data?

1) The median salaries in both companies are greater than $37,000.

2) The mean salary in company 1 is greater than the mean salary in company 2.

3) The salary range in company 2 is greater than the salary range in company 1.

4) The mean age of workers at company 1 is greater than the mean age of workers at company 2.

9) The table below shows the annual salaries for the 24 members of a professional sports team in terms of millions of dollars.

0.5	0.5	0.6	0.7	0.75	0.8
1.0	1.0	1.1	1.25	1.3	1.4
1.4	1.8	2.5	3.7	3.8	4
4.2	4.6	5.1	6	6.3	7.2

The team signs an additional player to a contract worth 10 million dollars per year. Which statement about the median and mean is true?

1) Both will increase.
2) Only the median will increase.
3) Only the mean will increase.
4) Neither will change.

10) The heights, in inches, of 12 students are listed below.

61,67,72,62,65,59,60,79,60,61,64,63

Which statement best describes the spread of these data?

1) The set of data is evenly spread.
2) The median of the data is 59.5.
3) The set of data is skewed because 59 is the only value below 60.
4) 79 is an outlier, which would affect the standard deviation of these data.

11) The 15 members of the French Club sold candy bars to help fund their trip to Quebec. The table below shows the number of candy bars each member sold.

Number of Candy Bars Sold				
0	35	38	41	43
45	50	53	53	55
68	68	68	72	120

When referring to the data, which statement is *false*?

1) The mode is the best measure of central tendency for the data.
2) The data have two outliers.
3) The median is 53.
4) The range is 120.

SOLUTIONS

4)ANS: 3

Strategy: Compute the mean, Q1, Q2, Q3, and interquartile range for each semester, then choose the correct answer based on the data.

	Mean	Q1	Median (Q2)	Q3	IQR
Semester 1	86.8	80.5	88	92.5	12
Semester 2	87	80	88	92	12

PTS: 2 NAT: S.ID.A.2 TOP: Central Tendency and Dispersion

5) ANS: 4

Strategy: Eliminate wrong answers based on daily high temperatures, then eliminate wrong answers based on size of interquartile ranges.

Ocean Breeze and Serene Shores can be eliminated because they *do not* have median high temperatures above 80 degrees. Whispering Palms and Pelican Beach *do* have median high temperatures above 80 degrees, so the correct answer must be either Whispering Palms or Pelican Beach.

The interquartile range is defined as the difference between the first and third quartiles. Pelican Beach has a much smaller interquartile range than Whispering Palms, so Pelican Beach is the correct choice.

PTS: 2 NAT: S.ID.A.2 TOP: Central Tendency and Dispersion

6) ANS: 4

Strategy: Determine the skew, spread, median, and mean for both data sets, then eliminate wrong answers.

	Soccer Players	Basketball Players
Skew	???	Left Skewed
Spread	$12-6=6$	$12-6=6$
Median	8.5	10
Mean	$\frac{156}{18}$	$\frac{178}{18}$

~~a)~~ The data for soccer players are skewed right. Uncertain

~~b)~~ The data for soccer players have ~~less spread~~ than the data for basketball players. Not True. Both data sets have the same spread.

~~c)~~ The data for basketball players have the ~~same median~~ as the data for soccer players. Not True $8.5 \neq 10$

d) The data for basketball players have a greater mean than the data for soccer players. Definitely True

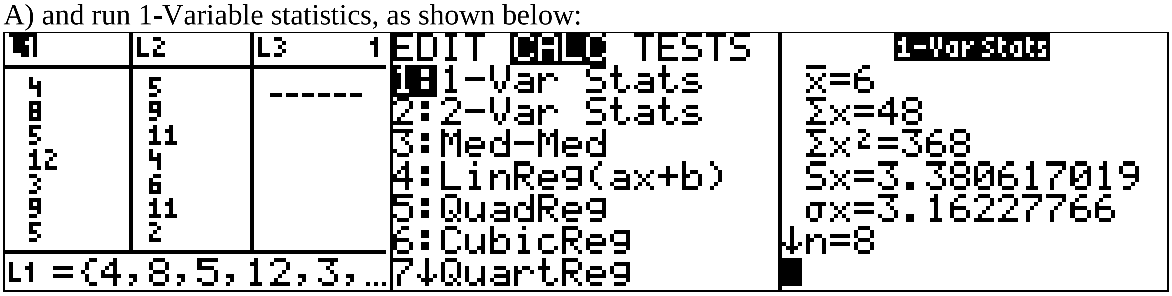
$\frac{178}{18} > \frac{156}{18}$

PTS: 2 NAT: S.ID.A.2 TOP: Central Tendency and Dispersion

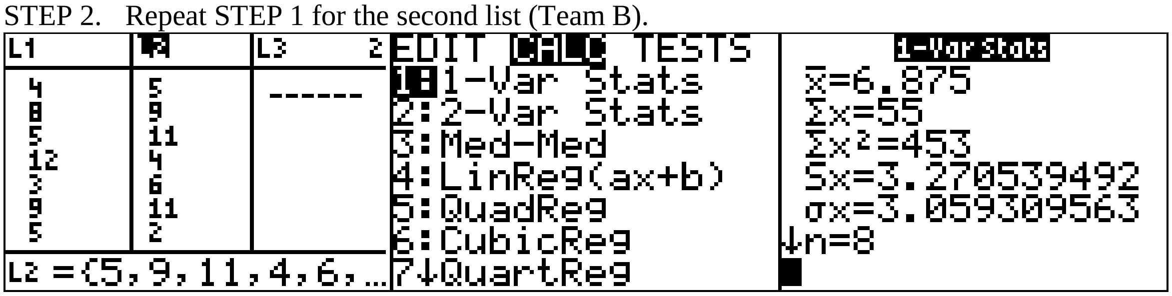
7) ANS: 1

Strategy: Compute the mean and standard deviations for both teams, then select the correct answer.

STEP 1. Enter the two sets of data into the STAT function of a graphing calculator, then select the first list (Team A) and run 1-Variable statistics, as shown below:



STEP 2. Repeat STEP 1 for the second list (Team B).



STEP 3. Use the data from the graphing calculator to choose the correct answer.

Choice a: mean A < mean B

$6 < 6.875$

standard deviation A > standard deviation B

$3.16227766 > 3.059309563$

Both statements in choice A are true.
$A: \bar{x} = 6;\ \sigma_x = 3.16 \quad B: \bar{x} = 6.875;\ \sigma_x = 3.06$

PTS: 2 NAT: S.ID.A.2 TOP: Central Tendency and Dispersion

8) ANS: 3
Strategy: Compute the median salary, mean salary, salary range, and mean age of employees for both companies, then select the correct answer.

		Company 1	Company 2
1	median salary	33,500	36,250
2	mean salary	33,750	44,125
3	salary range	8,000	36,000
4	mean age	28.25	28.25

PTS: 2 NAT: S.ID.A.2 TOP: Central Tendency and Dispersion

9) ANS: 3
Median remains at 1.4.

Strategy:
Compare the current median and mean to the new median and mean:

STEP 1. Compare the medians:
The data are already in ascending order, so the median is the middle number. In this case, the data set contains 24 elements - an even number of elements. This means there are two middle numbers, both of which are 1.4. When the data set contains an even number of elements, the median is the average of the two middle numbers, which in this case is $\frac{1.4 + 1.4}{2} = 1.4$

The new data set will contain 10 as an additional element, which brings the total number of elements to 25. The new median will be the 13th element, which is 1.4.

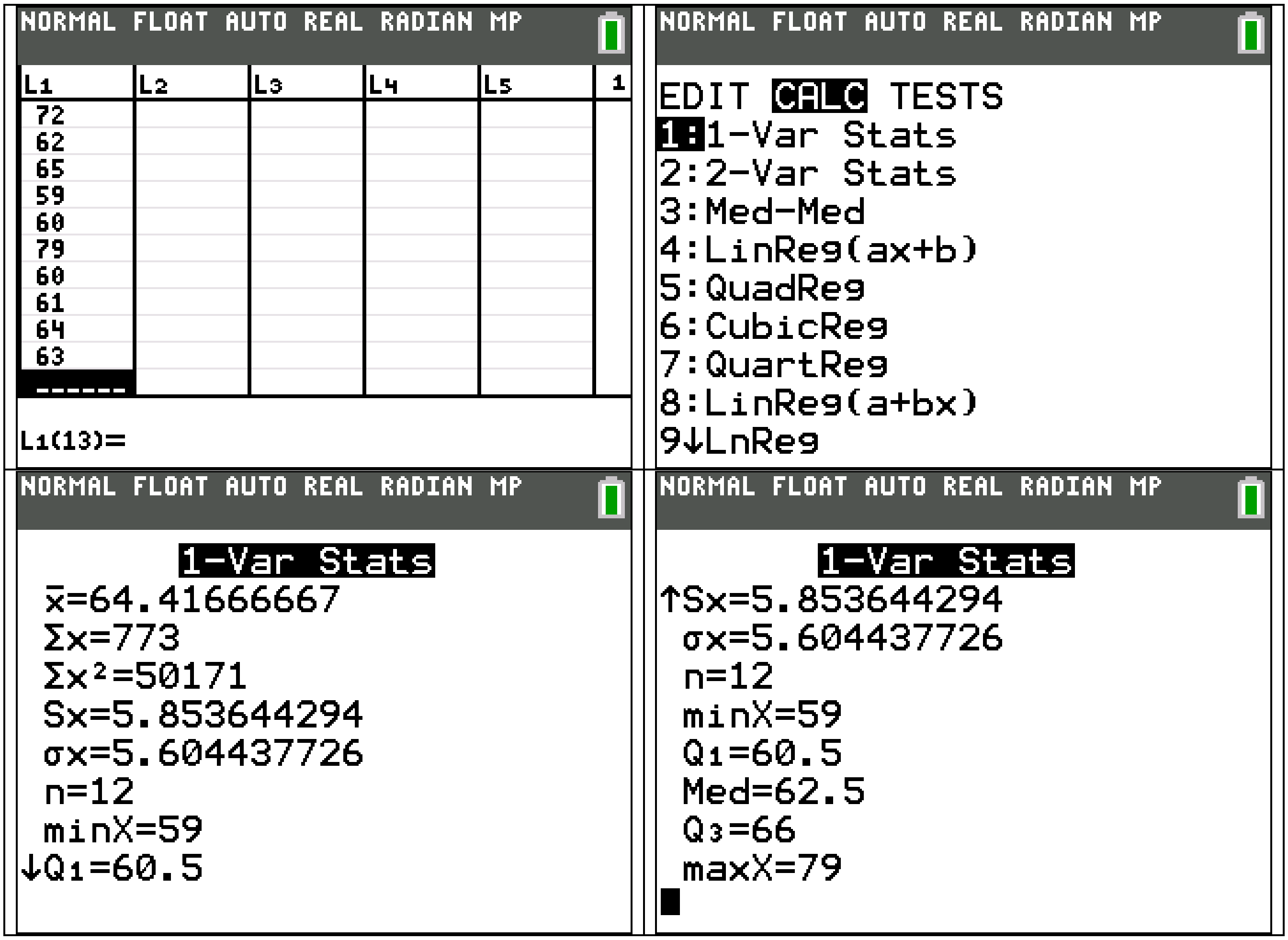
The current median and the new median are the same, so we can eliminate answer choices a and b.

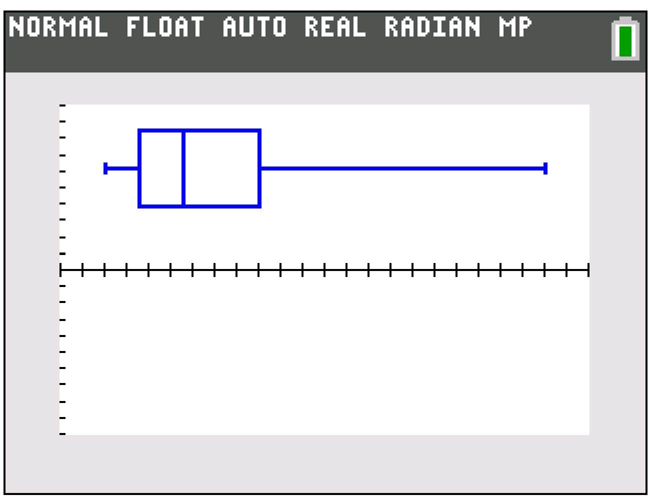
STEP 2. Compare the means:
The mean will increase because the additional element (10) is bigger than any current element. It is not necessary to do the calculations. We can eliminate answer choice d.

DIMS? Does it make sense that the answer is choice c?
Yes. The median will stay and 1.4 and only the mean will increase.

PTS: 2 NAT: S.ID.A.3 TOP: Central Tendency and Dispersion

10) ANS: 4
Input the data in a graphing calculator and obtain single variable statistics, then create a boxplot.





(1) The set of data is evenly spread. Wrong. The data is not evenly spread.
(2) The median of the data is 59.5. Wrong. The median of the data is 62.5.
(3) The set of data is skewed because 59 is the only value below 60. Wrong. The data is skewed, but the reason for skewdness is that the mean does not equal the median.
(4) 79 is an outlier, which would affect the standard deviation of these. True. Any value greater than Q3 plus 1.5 times the interquartile range is an outlier.

$$Q3 + 1.5(IQR) = \textit{Upper Outlier Fence}$$

$$66 + 1.5(66 - 60.5) = \textit{Upper Outlier Fence}$$

$$66 + 1.5(5.5) = \textit{Upper Outlier Fence}$$

$$66 + 8.25 = \textit{Upper Outlier Fence}$$

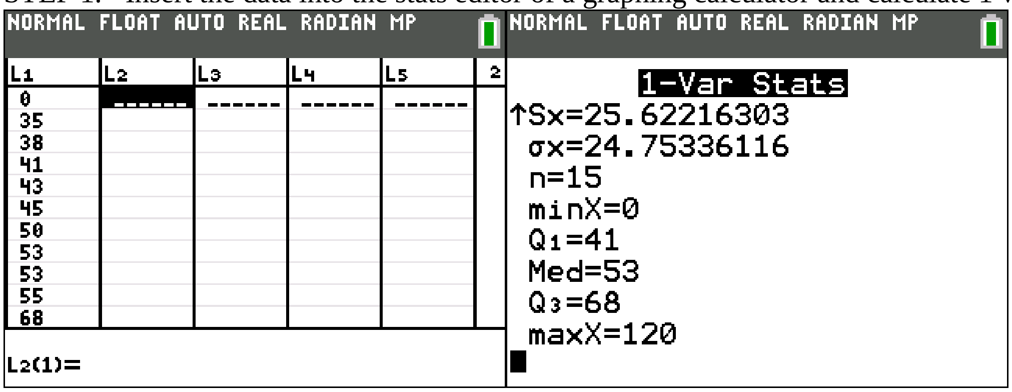
$$74.25 = \textit{Upper Outlier Fence}$$

79 is beyond the upper outlier fence.

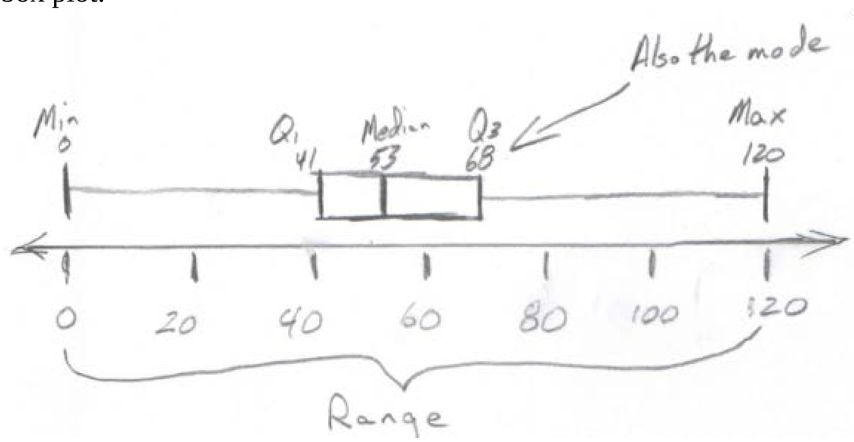
PTS: 2 NAT: S.ID.A.3 TOP: Central Tendency and Dispersion

11) ANS: 1

STEP 1. Insert the data into the stats editor of a graphing calculator and calculate 1 variable statistics.



Step 2. Construct a box plot.



STEP 3: Eliminate *true* answer choices.

It is trued that the data have two outliers. These are 0 and 120.

It is true that the median is 53.

It is true that the range is 120

PTS: 2 NAT: S.ID.A.3 TOP: Central Tendency and Dispersion