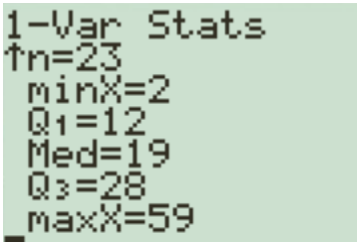
Name:_____ Date:_____

AP Statistics: Assignment #1.1 Boxplots, Exploring Data SOLUTIONS

1. A group of students from AP Statistics were timed when asked to complete the last question from the Gauss math contest. Use the stem leaf plot to answer the questions below:

0	9
1	1 1 2 2 3 4
1	5 5 7 9 9
2	1 1 3 3
2	6 8 9
3	1
3	7
4	1
5	9

a) Find the 5 number summary: Min, Q1, Median, Q3, and Max



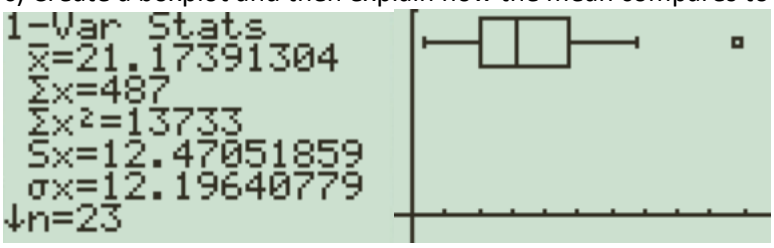
If we perform a 1 variable stats test, we see that the minimum is 2, Q1 is 12, Median is 19, Q3 is 28, and Max is 59.

b) Find the IQR and identify any outliers

TO find the IQR, subtract Q3 – Q1 = 28 – 12 = 16.
To find any outliers, multiply 1.5 x (Q3 – Q1) = 1.5 x 16 = 24.

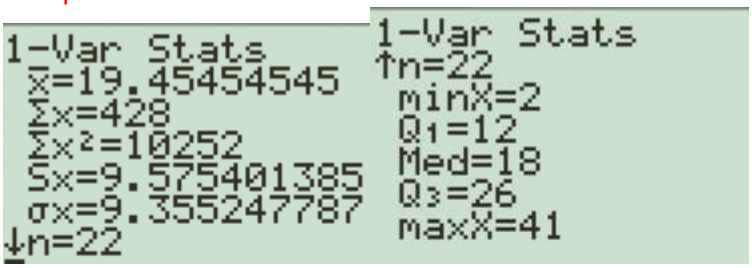
Any value that is greater than Q3 + 24 OR less than Q1 – 24 is an outlier.
Q3 + 24 = 28 + 24 = 52. And Q1 – 24 = 12 – 24 = -12 (neglect this)
So the value of 59 is an outlier.

c) Create a boxplot and then explain how the mean compares to the median



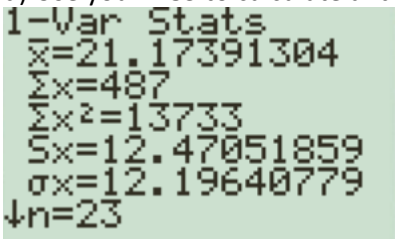
With an outlier on the right side, the distribution of this class of AP scores will be skewed right. The mean is greatly affected by outliers, so with an outlier on the right, the mean will be slightly greater than the median.
Mean is 21.1739 and Median is 19. The mean is greater than the median.

NOTE: One way to investigate the effects of the outliers is to remove it from the data set and see how it compares with the data with the outlier. You're not required to do this for this question.



From our output, without the outleir the mean would drop to 19.45 and the median is 18. The mean is more affected by the outlier than the median is.

d) Use your Ti83 to calculate and then interpret the mean.



The mean time for the students from an AP Statistics class to complete the last question from the Gauss contest is 21.17 minutes.

[Keep in mind that your responses need to be in the context of the question. Remember to identify who the subjects are, what is being tested, what is measured….]

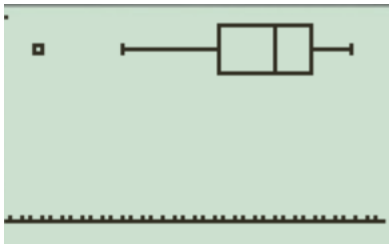
e) Use your Ti83 to calculate and then interpret the median

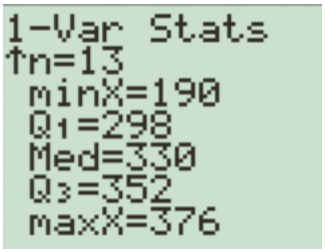
The median time for the students from an AP Statistics class to complete the last question from the Gauss contest is 19 minutes.

f) Which measure of center would be the most appropriate summary of the center of this distribution? Explain why:

Since there is an outlier, the median would be a more appropriate measure of central tendency. The median is not greatly affected by outliers. Whereas, the mean is greatly affected by outliers and would NOT be a good measure of central tendency in this case.

2. The number of pages of Mr. Cheong's favorite books are noted below:
   240 , 350, 310, 346, 320, 286, 336, 366, 190, 354, 318, 376, and 330.
   a) Use the data above to construct a boxplot. Describe the center and spread using the five number summary:





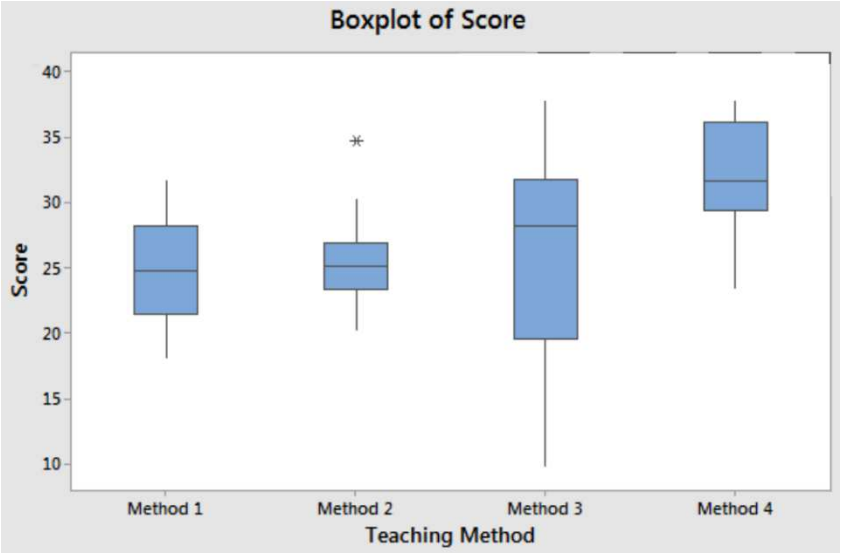
From out data output, the Minimum value is 190, Q1 is 298, Median is 330, Q3 is 352, Max is 376 and the range is 186. The middle 50% is ranged from 298 to 352, with an IQR is 54. There is an outlier on the lower end with a value of 190. The distribution is skewed to the left.

b) Calculate and interpret the mean and standard deviation for these data:

```
1-Var Stats
 x̄=317.0769231
 Σx=4122
 Σx²=1339780
 Sx=52.27246174
 σx=50.22175676
↓n=13
```

The mean number of pages in Mr. Cheong's favorite books have 317.077 pages with a standard deviation of 50.22 pages.

3. The following graph shows the distribution of scores from four classes that utilized different teaching methods. Each class had a class size of 50 students and were randomly assigned. Student scores ranged from 0 to 40. Use the graph to answer the following questions:



a) Describe the shape, center, and spread of the distribution of scores for each teaching method

Method 1: The distribution is symmetrical, centered at a median around 25, with Q1 at 22, Q3 at 27, Minimum at 17, and Maximum at 32. The middle 50% is between 22 to 27, with an IQR of 5. The range of scores using Method 1 is 15.

Method 2: The distribution is skewed to the right with an outlier at 35. The median of the distribution is at 25, Q1 24, Q3 at 26, Max at 35 and minimum at 20. The middle 50% is between 24 to 26, with an IQR of 2. The range of the data is also 15. If we removed the outlier, the distribution would be symmetrical, with a max of 30.

Method 3: The distribution is skewed left with a long tail on the left side. The median is around 27, Q1 at 19, Q3 at 31, Min is 10 and Max is 37. The middle 50% is between 19 to 31, with an IQR of 12. The range is 27.

Method 4: The median is around 32, Q1 at 30, Q3 at 35, Min at 25 , and Max at 37. The middle 50% is between 30 and 35, with an IQR of 5. 25% of the data point is between a small range of 35 and 37, and

another 25% is within 30 and 32. I would assume the distribution to be bimodal. The range of scores of this method is 12.

b) What do the data show, how would you rank the efficiency of teaching method? Justify your answer.
Teaching method 4 would have the highest efficiency because 75% of the class received a score of 30 or higher. In addition, the Median Q1 Q3 Max of this class are all higher than the rest of the other teaching methods.

Amongst the other three classes, it depends on how you define efficiency. For instance, method 3 has a higher median, Q3 and max than methods 1 and 2. AT the same time, it also has a higher failure rate than the other two methods. 25% of the students in Method 3 received a score of 20 or lower. I would rate this method as the least efficient.

Method 2 has a higher Q1 than Method 1, but a lower Q3. All students in Method 2 received a passing score of 20 or higher. In contrast, around 15 to 20% of students in method 1 received a score of 20 or lower in Method 2. Therefore, I would rank Method 2 to be more efficient than Method 1.

Keep in mind that these answers can vary. The quality of your answers would be based on your justification, rather than simply on the ranking.

4. For each of the following statements, indicate whether if it is true or false.
   a. If a distribution is skewed left, then the median is larger than the mean _____
   If the distribution is skewd left, the mean would be left of the median. Assuming that it is on a number line where the right side represents bigger values, then the statement would be true, Median would be greater than the mean.

   b. If a distribution is skewed right, then the median is larger than mean _____
   False, a right tail would pull the mean towards it and generate a larger mean than the median.

   c. If the mean is larger than the median, then the distribution is skewed right _____
   This is false. The mean can be larger than the median without being skewed right.

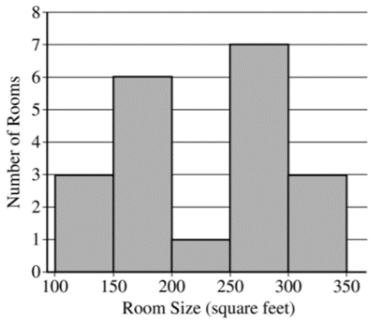
   d. If the mean is smaller than the median, the the distribution is skewed left _____
   False. Same reasoning as above.

   e. If a density curve is symmetric, then the mean is equal to the median _____
   False, you can have a symmetrical distribution that is bi modal.

5. If a set of data shows a distribution that is skewed right, would the "mean" or "median" be a better measure of central tendency? Explain:

If the distribution is skewed, then the mean would be NOT be a good measure of central tendency. Instead, we should use the median. This is because outliers can greatly affect the mean but not the median.

6. The sizes, in square feet, of the 20 rooms in a student residence hall at a certain university are summarized in the following histogram. The summary statistics for the sizes are also given in the table below:



Mean	Standard Deviation	Min	Q1	Median	Q3	Max
231.4	68.12	134	174	253.5	292	315

a) Using the histogram, write a few sentences describing the distribution of room size in the residence hall

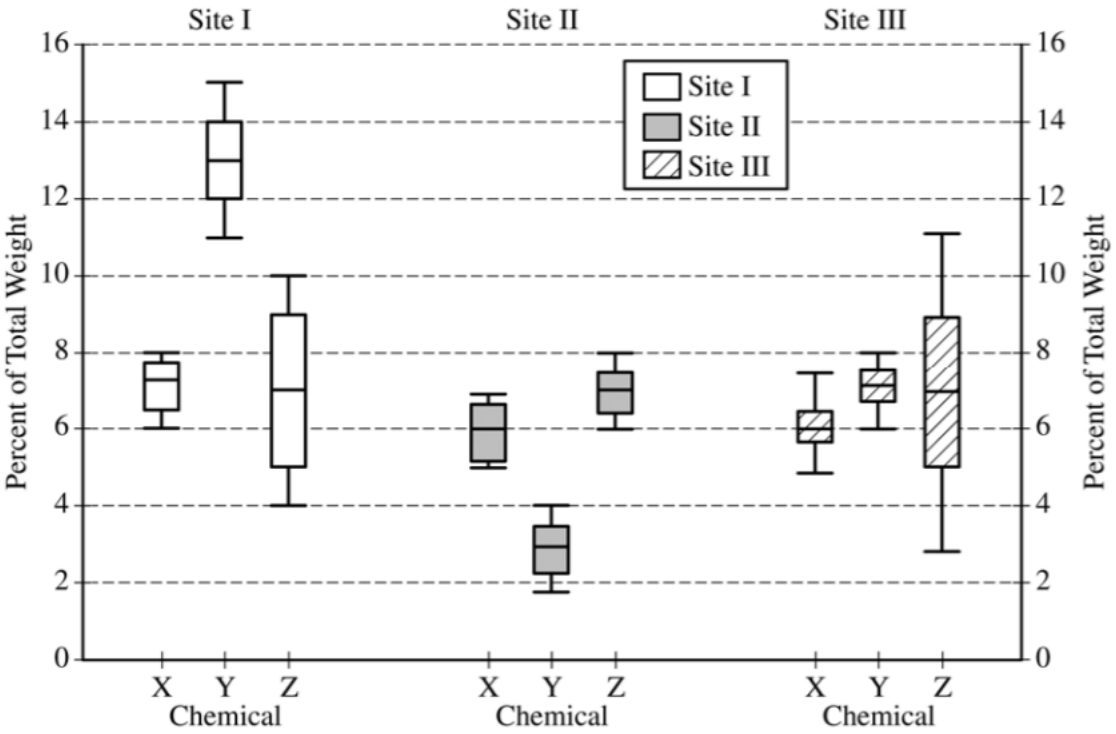
b) Determine whether there are potential outliers in the data. Use the grid below to draw a boxplot of the room size:

c) What characteristic of the shape of the distribution of room size is apparent from the histogram but not from the boxplot?

Question7

The chemicals in clay used to make pottery can differ depending on the geographical region where the clay originated. Sometimes, archaeologists use a chemical analysis of clay to help identify where a piece of pottery originated. Such an analysis measures the amount of a chemical in the clay as a percent of the total weight of the piece of pottery. The boxplots below summarize analyses done for three chemicals—X, Y, and Z—on pieces of pottery that originated at one of three sites: I, II, or III.

(cont. on next page)



(a) For chemical Z, describe how the percents found in the pieces of pottery are similar and how they differ among the three sites.

(b) Consider a piece of pottery known to have originated at one of the three sites, but the actual site is not known.

z

(i) Suppose an analysis of the clay reveals that the sum of the percents of the three chemicals X, Y, and Z is 20.5%. Based on the boxplots, which site—I, II, or III—is the most likely site where the piece of pottery originated? Justify your choice.

(ii) Suppose only one chemical could be analyzed in the piece of pottery. Which chemical—X, Y, or Z—would be the most useful in identifying the site where the piece of pottery originated? Justify your choice.