MAC-CPTM Situations Project

Situation 34: Mean Median

Prepared at Penn State
Mid-Atlantic Center for Mathematics Teaching and Learning
18 June 2005 – Sue, Evan McClintock, Donna Kinol

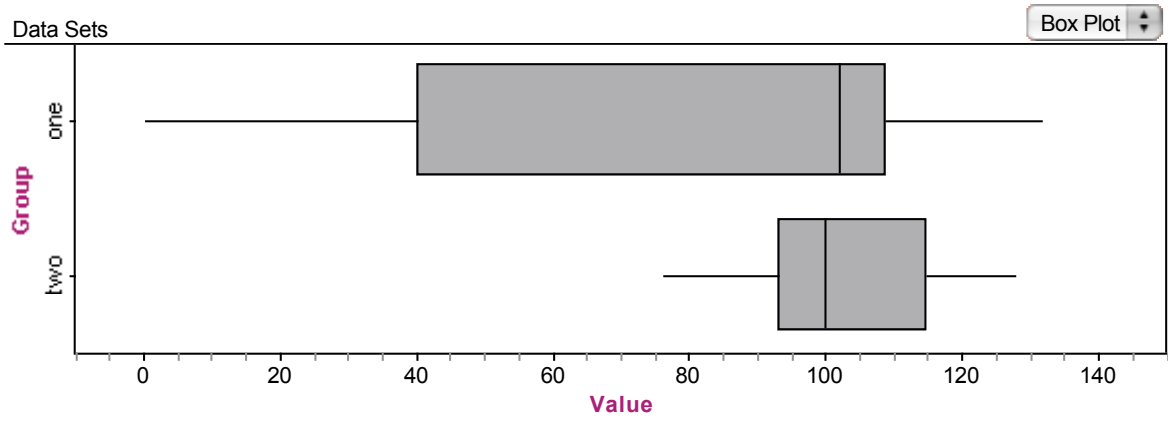
Edited at Penn State
Mid-Atlantic Center for Mathematics Teaching and Learning
14 February 2007 – Shiv Karunakaran
23 February 2007 – Rose Mary Zbiek, M. Kathleen Heid

Edited at University of Georgia
16 March 2009 – Laura Singletary and Sarah Donaldson

Prompt

The following task was given to students at the end of the year in an AP Statistics class.

Consider the following box plots and five-number summaries for two distributions. Which of the distributions has the greater mean?



Data Sets

	Group	
	one	two
Value	0	76
	40	93
	102	100
	109	115
	132	128

S1 = min ( )
S2 = Q1 ( )
S3 = median ( )
S4 = Q3 ( )
S5 = max ( )

One student's approach to this problem was to construct the following probability distributions for each data set and compare the corresponding expected values to determine which data set has the greater mean. The student calculated the mean for each interval, where the 4 intervals were formed using the five-number summaries given above (i.e. he defined his intervals as the four quarters of the distributions each quarter containing 25% of the values for the distribution). Using the midpoint of each interval as the X value for that interval, he then calculated the weighted mean for each probability distribution. After conducting his calculations, the student responded that the second data set had the larger mean.

Data set one:

E(X) = 79.25

	0 – 40	40 - 102	102 - 109	109 - 132
X	20	71	105.5	120.5
P(X)	0.25	0.25	0.25	0.25

Data set two:

E(X) = 102.75

	76-93	93-100	100-115	115-128
X	84.5	96.5	107.5	121.5
P(X)	0.25	0.25	0.25	0.25

Commentary

This Prompt deals with the differences and similarities between the mean and median of a particular data set when the data set is displayed as a box plot using its five-number summary. It is likely that the intent of the question was not to encourage mathematical calculations, but rather to ask students to predict which distribution would have a greater mean based on what is expected from the visual display of the box plots.

Another important aspect to consider is that the problem given in the Prompt is given without a context. Without knowing the context of the data given, we do

not know if the data is continuous or discrete. Also, we do not know the statistical question being considered; i.e. why was the data collected?

Mathematical Foci

Mathematical Focus 1
The skewness of a data distribution affects the relationship between the mean and median of that set of data.

These box plots provide information about the distributions of the two data sets. Data set 2 appears to be roughly symmetric with possible right skewness (note that the median pulled to the left of central box). For this case, we would expect the mean and the median to be approximately equal, or the mean slightly greater than the median. Data set 1 appears skewed left. Therefore, we expect the mean will be less than the median if right skewed. If the distribution for the first data set is skewed to the left, then smaller values have a stronger impact on the mean than the larger values. On the other hand, the median is the "middle" value of the data set after the data set is arranged in increasing order, and is resistant to the larger spread in the smaller values. Since the medians are similar in each distribution, we expect data set 2 has the larger mean. Although reasoning via the shape of the distribution is an approach that typically works when making comparisons about distributions, it is not always possible to make conclusive statements about the relative locations of the means for some pairs of box plots.

Mathematical Focus 2
A box plot display of data does not necessarily give the data values or information about the "distribution" of the data within each quarter.

How the data are distributed within each interval determined by the five-number summary is not represented in a box plot. The mean of a particular interval represented by the midpoint would be representative of the data points in that interval only when the data is distributed normally, uniformly, or symmetrically within that interval. The information given in the Prompt does not allow us to make such an assumption.

In the Prompt the student assumed that each interval contains exactly 25% of the data points. However this is only true when the number of data points is divisible by 4. Also, it is important to note that some of the numbers in the five-number summary may not be members of the data set. The only values in the data set that we know for certain from the box plot are the minimum and maximum values. The median will only be a member of the data set when the number of data points is odd. Q_1 and Q_3 are members of the data set only when the size of the data set has a remainder of 2 or 3 when divided by 4.

In the Prompt, the student seems to have made contradictory assumptions about the data set. In his calculations, he assumed that each interval contained exactly

25% of the data set, indicated in the assignment of probability P(X)= 0.25 for each interval. Another assumption he seems to have made is that Q_1 , the median, and Q_3 are values in the data set. As discussed previously, these assumptions cannot both be valid simultaneously, because the number of values in the data set is either even or odd, but not both.

Mathematical Focus 3
When exact values of two quantities are not known, comparisons between the two quantities can sometimes be made by comparing the ranges of their possible values.

For the given box plots, we can calculate an upper bound and a lower bound for the means of the data sets. Because of the apparent skewness discussed in Focus 1, we assume that data set 2 has a greater mean than data set 1. In order to investigate this, let us consider the upper bound for the mean of data set 1 and compare it to the lower bound for the mean of data set 2. We will see that the highest possible value of the mean of data set 1 is strictly less than the lowest possible value of the mean of data set 2. Therefore, we can conclude that the mean of data set 2 is greater than the mean of data set 1.

To find an upper bound for the mean of data set 1, assume that in each interval the data points are located at the greatest possible value within the interval. (Of course, strictly speaking, since 0 is the minimum value of the data set, 0 is also a data point in the first interval; however, since we are investigating an upper bound and *n* is not known, we will not include 0 in the calculation.) Let 25% of the values in data set 1 be located at the maximum value of each interval, that is, at Q1, at the median, at Q3, and at the maximum. In this extreme case, the mean of the data set is given by:

$$E(\text{data set 1}) = 0.25(40) + 0.25(102) + 0.25(109) + 0.25(132) = 95.75$$

Similarly, in order to find a lower bound of the mean for data set 2, assume that in each interval the data points are located at the least possible value within the interval. (By using this method we will not include the maximum value, 128, in our calculation.) Let 25% of the values in data set 2 be located at the minimum value of each interval, that is, at the minimum, at Q1, at the median, and at Q3. In this extreme case, the mean of the data set is given by:

$$E(\text{data set 2}) = 0.25(76) + 0.25(93) + 0.25(100) + 0.25(115) = 96.00$$

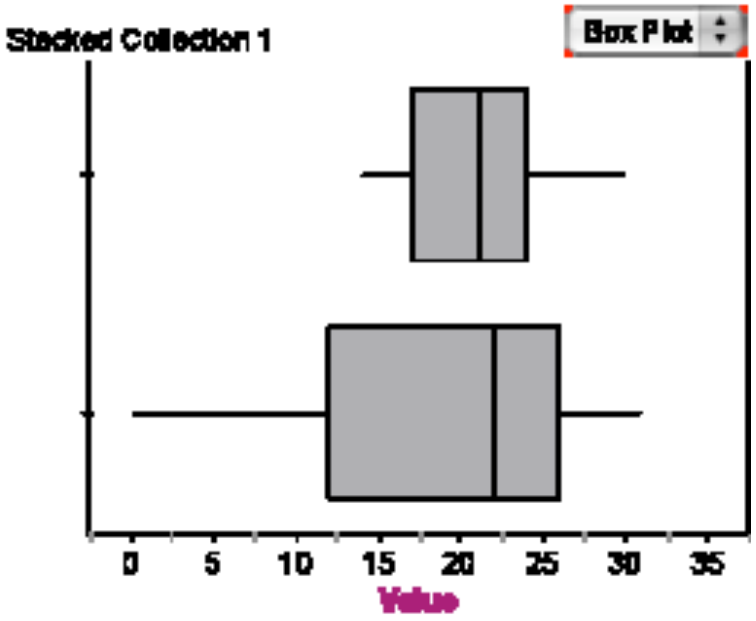
Since the lower bound of the mean of data set 2 is greater than the upper bound of the mean of data set 1, we can conclude that the mean of data set 2 is greater than the mean of data set 1.

It is important to note that the previous example assumed that each interval contained exactly 25% of the data set. This assumption may easily be wrong, because we know that this situation only occurs when the size of the data set is

equivalent to 4k (i.e. is divisible by 4). The Post-Commentary contains further investigation of data set 1 taking into consideration the possibility that the data set is of size 4k+1, 4k+2, or 4k+3. Even in these cases, however, the mean of data set 2 is shown to be greater than the mean of data set 1.

Mathematical Focus 4
Stating a definitive conclusion about a comparison of the means using the five-number summary and box plots is not always possible because the size of the data set may influence the relationship between the means for these distributions.

The following example will portray the importance of sample size and its effect on the relationships between the means and the five-number summaries for the given distributions.



	Group	
	One	Two
Value	14	0
	17	12
	21	22
	24	26
	30	31

S1 = min( )
S2 = Q1( )
S3 = median( )
S4 = Q3( )
S5 = max( )

For this example, consider the possibility that each data set contained twelve values. If the values in data set A were 14, 14, 17, 17, 17, 21, 21, 21, 24, 24, 24, and 30, then the mean of data set A would be 20.333. If data set B contained the

values of 0, 12, 12, 12, 22, 22, 22, 26, 26, 26, 31, and 31, then the mean of data set B would be 20.167. Thus, the mean of data set A would be larger than the mean of data set B.

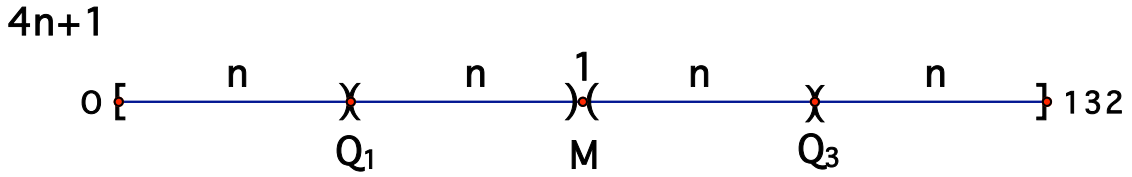
However, what if each set of data contained 100 values and the five-number summaries were maintained? If data set A was distributed with 24 values at 14, 25 values at 17, 25 values at 21, 25 values at 24, and 1 value at 30, then the mean of data set A would be 19.16. If data set B was distributed with 24 values at 31, 25 values at 26, 25 values at 22, 25 values at 12, and 1 value at 0, the mean of data set B would be 22.44. Thus, the mean of data set B would be larger than the mean of data set A.

Since the sizes of the data sets may influence the relationship between their means, stating definitive conclusions about a comparison of means is not always possible for a pair of box plots.

Post-Commentary

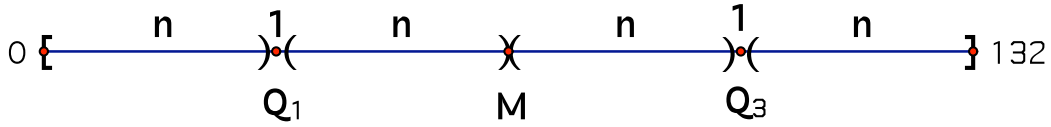
In Focus 3, we calculated the upper bound of the mean of data set 1 and the lower bound of the mean of data set 2 assuming the size of each data set was divisible by 4. The following pictures illustrate the possibilities not considered in Focus 3 (i.e. that the size of a data set might be 4k+1, 4k+2, or 4k+3).

Consider the possibility that data set 1 contains 4k + 1 data values. In order to maintain the same five-number summary, the extra data value would be located at the median, 102. The mean value of 95.75 calculated above assumed 25% of the data values would lie in each segment; however, the lower extreme value of 0 was not accounted for in the calculation. Note that accounting for the lower extreme of 0 (by effectively removing a value of 40) will lower the mean more than the addition of a value at 102 will increase the mean. Thus, the mean of data set 2 is still larger than the mean of data set 1.



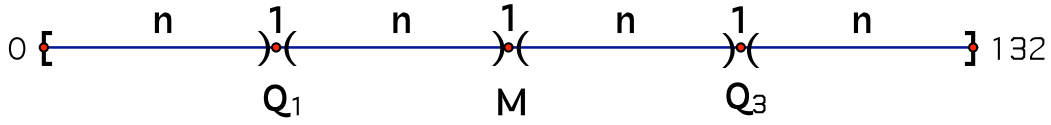
If data set 1 contains 4n + 2 values, then one additional value will be located at Q1 and one additional value will be located at Q3. Accounting for the minimum of 0, this situation nets the addition of a value at 0 and a value at 109. The value added at 0 lowers the mean more than the value of 109 increases the mean, and the mean of data set 2 is still larger than the mean of data set 1.

4n+2



Lastly, if data set 1 contains 4n + 3 values, then a value is added at Q1, the median, and Q3. Accounting for the minimum of 0, this situation nets the addition of a value at 0, a value at 102, and a value at 109. The value added at 0 lowers the mean more than the addition of values at 102 and 109. And therefore, the mean of data set 2 is still larger than the mean of data set 1.

4n+3



In all cases, the mean of data set 2 is larger than the mean of data set 1. Similar arguments can be made for different sample sizes and their effects on the mean of data set 2.

Each of the Foci highlight a difference between information that allows conclusions to be made with mathematical precision, and information that only allows for general claims to be made. For example, Focus 3 contains a conclusive argument for the relative sizes of the two means, whereas Focus 1 describes how claims can be made based on what is expected using reasoning about the shape of the distribution.