Situation 34: Mean Median

Prepared at Penn State

Mid-Atlantic Center for Mathematics Teaching and Learning

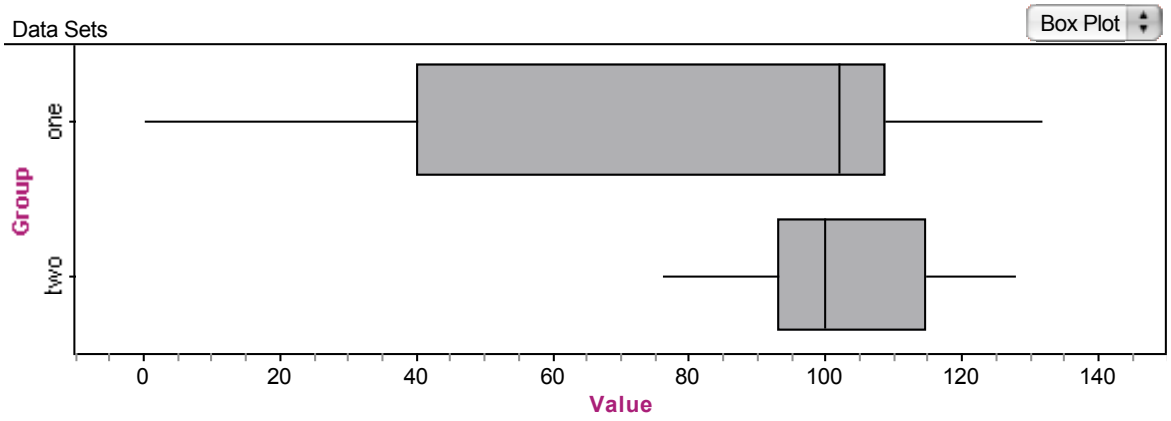
June 29, 2005 – Sue, Evan, Donna

Prompt

This prompt is based on a student's response to a task given at the end of the year in an AP Statistics class. The purpose of the task was to have students interpret box plots in relation to the mean.

The Task:

Consider the following boxplots 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 then to compare the corresponding expected values to determine which data set has the greater mean. The student responded that Data set two 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

Question: What information about mean, median, and the relationship between mean and median is revealed or obscured by box plots?

Commentary

Mathematical Foci

Mathematical Focus 1

Underlying mathematical understandings: The box plot display of a quartile gives no information about the "distribution" of data within that quartile. In particular, the mid-point of a quartile is no more or less informative than any other point in the quartile. Box plots are constructed by using the five summary statistics, the median, the first and third quartile, and the minimum and maximum values of the data set. Quartile endpoints may or may not be actual values in the data set. The mean can be calculated as an expected value using a frequency distribution approach, assuming the necessary information about the distribution (e.g. the mean of each quartile, the actual data values, etc.) is known.

For each data set, it seems that the student partitioned the data into four equal groups using the information about the quartiles from the box plots and then found the midpoint of each segment using the extreme values of each segment.

Finding the midpoint does not take into consideration how the data are distributed. Nor can we determine how the data are distributed in each segment given how the data is represented in a box plot. The midpoint will be representative of the data points in the segment in cases such as when the data are distributed normally, uniformly, or symmetrically. If each segment contains 25% of the data points and we know a representative value of the segment, then we can determine the expected value (mean) of the distribution by summing the products of the value and its probability. Each segment will contain 25% of the

data values when the size of the data set is a multiple of four. The median will be a member of the data set when the size of the data set is odd. The first and third quartiles will be members of the data set when the size of the data set is congruent to 2 mod 4 or congruent to 3 mod 4. The only values in these data sets that we know for certain are the lower and upper extremes. The other data points may be skewed in any segment, uniformly distributed across the segment, unimodal or multimodal within the segment, or distributed in any other way. Based on what is given in this problem, using the midpoint as representative of the values contained in the segment is a conservative approach (and the best we can do).

Mathematical Focus 2

Mean and median have particular relationships based on how the data are skewed.

These box plots provide information about the distributions of the two data sets. Data set two appears to be fairly symmetric. In that case, the mean and the median would be approximately equal. Data set one seemingly is skewed left, and if so, the mean will be less than the median. If the assumptions about skewness are met, then since the medians of the two data sets are approximately equal, data set two has the greater mean. Although reasoning via skewness is an approach that works for some distributions, it may be impossible to discern the relative locations of the means for some pairs of box plots.

Mathematical Focus 3

Stating definitive conclusions about a comparison of means is not possible for every pair of boxplots. In the case of the given pair of boxplots, however, definitive conclusions can be drawn. In order to compare boxplots to know when definitive conclusions can be drawn, one must know how to produce box plots, know how to compute means, know that the mean does not appear in a boxplot (else I could use the given information), and know that sample size does matter in drawing these conclusions – sample size affects mean and median.

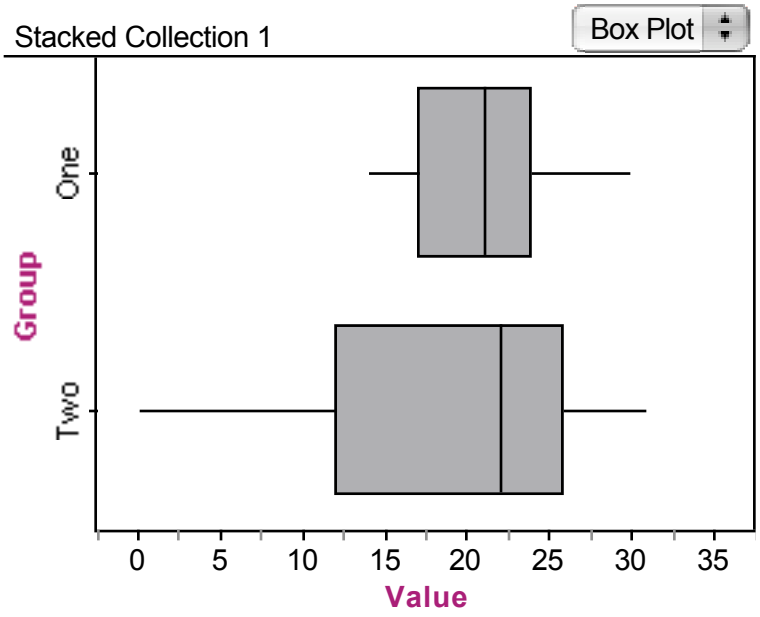
For the given boxplots, in the most extreme case scenario, to find the largest possible mean for data set 1, 25% of the values in data set 1 would be located, respectively, at Q1, at the median, at Q3, and at the upper extreme. Note: this is actually more extreme than the most extreme case, since we know that at least one data value is 0. In this extreme case, the mean of the data set would be: E(data set 1) = 0.25(40) + 0.25(102) + 0.25(109) + 0.25(132) = 95.75. Similarly, in order to find the smallest possible mean for data set 2, 25% of the values in data set 2 would be located, respectively, at the lower extreme, at Q1, at the median, and at Q3. Note: this is actually more extreme than the most extreme case, since we know that at least one data value is 128. In this extreme case,

the mean of the data set would be: E(data set 2) = 0.25(76) + 0.25(93) + 0.25(100) + 0.25(115) = 96. Even in this most extreme case, the mean of data set 2 is larger than the mean of data set 1.

Note: If data set 1 contains $4n$ values, then exactly 25% of the data values lie in each segment of the boxplot. The sample size in the above argument was assumed to be $4n$. If instead data set 1 contains $4n + 1$ data values, in order to maintain the same five-number summary, the extra data value would be located at the median of 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. 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. 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.

To produce a counterexample, consider the following boxplots.



	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 these data, if each set of data contained twelve values, and the values contained in data set 1 were 14, 14, 17, 17, 17, 21, 21, 21, 24, 24, 24, and 30, then the mean of data set 1 would be 20.333. If data set 2 contained the values of 0, 12, 12, 12, 22, 22, 22, 26, 26, 26, 31, and 31, then the mean of data set 2 would be 20.167. Thus, the mean of data set 1 would be larger than the mean of data set 2. However, if each set of data contained 100 values and five-number summaries were maintained, and data set 1 was distributed with 24 values at 14, 25 values at 17, 25 values at 21, 25 values at 24, and 1 value at 30, the mean of data set 1 would be 19.16. If data set 2 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 2 would be 22.44. Thus, the mean of data set 2 would be larger than the mean of data set 1. Stating definitive conclusions about a comparison of means is not possible for this pair of boxplots, and sample size impacted the relationship between the means for these distributions.

Mathematical Focus 4

Underlying the previous three mathematical foci seems to be an understanding of the mean as a balance point. The balance point is the point for which the sum of the distances between the point and each data value to the left of the point equals the sum of the distances between the point and each data value to the right of the point. This balance point is the mean.

One way to work this problem with mean as a balance point is,

Understanding the mean as balance point allows for the midpoint to be the mean given distributions that we've previously mentioned. But understanding mean as balance point can also allow us to think of the mean of a segment as falling anywhere in the segment. The minimum value for the mean of the distribution can be found by considering all of the weights being placed at the minimum value of each segment. The maximum value for the mean of the distribution can be found by considering all of the weight being placed at the maximum value of each

segment. In this manner a range of possible means for the two distributions can be found and comparisons can be made between them.

References

The inspiration for this setting came from a student response to question 1 of the 2004 AP Statistics examination. This task is considerably different from the actual problem on the examination.