MAC-CPTM Situations Project

Situation 34: Mean Median

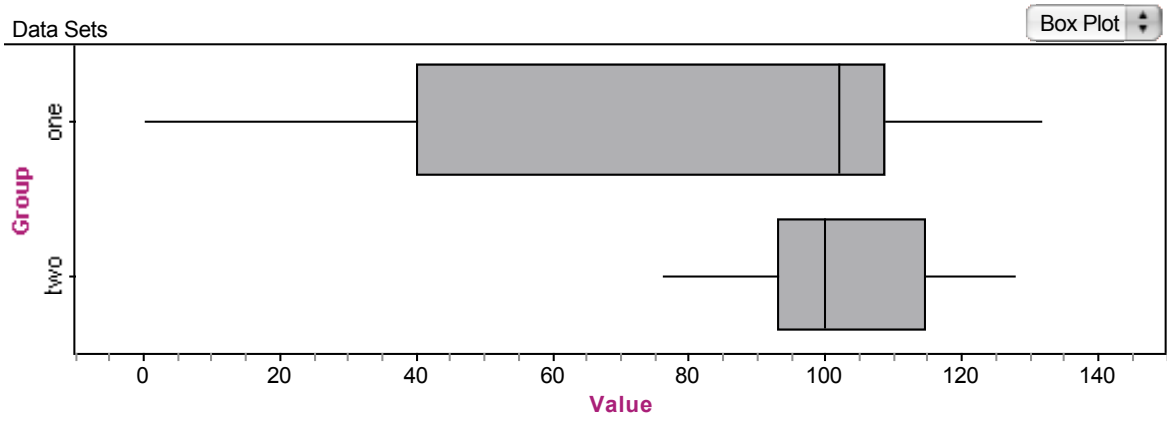
Prepared at Penn State
Mid-Atlantic Center for Mathematics Teaching and Learning
18 June 2005 – Sue, Evan McClintock, Donna Kinol

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

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 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

Commentary

This prompt seems to deal 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.

Mathematical Foci

Mathematical Focus 1

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

Using the mean of adjacent quartiles and extreme values to represent each of four quarters of data does not take into consideration how the data are distributed. How the data are distributed in each segment is not represented in a box plot. The midpoint will be representative of the data points in the segment in cases such as when the data is 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. There is also a problem in that the quartiles may not be data values. The only values in these data sets that we know for certain are the lower and upper extremes. The median will be a member of the data set when the number of data points 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 to 3 mod 4. Each segment will contain 25% of the data values when the size of the data set is a multiple of four.

Mathematical Focus 2

The skewness of data in a set 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 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. This is because the lesser values that result in the distribution being skewed left will have more effect on the mean than the greater values. On the other hand, the median is the "middle" value of the data set after the data set is ordered in either increasing or decreasing order, and thus is not affected in the same way by the spread of the lesser values. If data set 2 is not skewed like data set 1, then because 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
When exact values of two quantities are not known, comparisons between the two quantities can sometimes be may by comparing the ranges of their possible values.

For the given box plots, 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. In this extreme case, the mean of the data set would be 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 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. In this extreme case, the mean of the data set would be given by

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

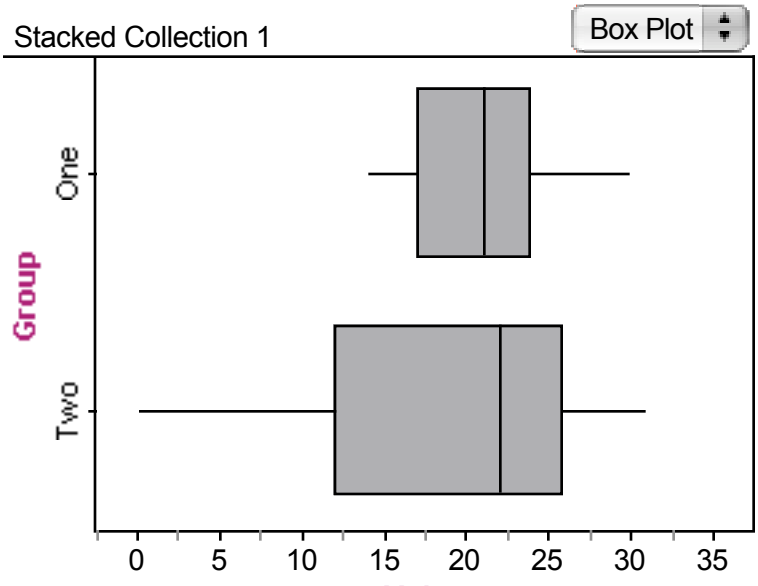
So, any mean of a data set that has the box plot shown for data set 2 is greater than the mean of any possible data set 1.

It is important to note that if data set 1 contains 4n values, then exactly 25% of the data values lie in each segment of the box plot. 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 box plots.



	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 sets, if each 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 box plots, and sample size impacted the relationship between the means for these distributions.

Mathematical Focus 4
Using the idea that other statistics (e.g. range) depend on only a few data points can be used to provide gross estimations of some statistics (e.g., mean) that depend on all values in a data set.

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

Given only a box plot to represent a data set, we do not know any data values other than the minimum and maximum values. Without knowing where the points lie along between the extreme values, we do not have enough information to locate the balance point.

Think about the data as divided into four segments, each of which begins and ends with adjacent values among the quartiles and extreme values. The minimum value for the mean of the distribution can be found by considering all of the data values within a segment as the minimum value of each segment, without changing the maximum value. The maximum value for the mean of the distribution is similarly found by using 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 these ranges.

Post-Commentary

Each of the foci underscores a need to attend to separating information that allows us to make claims with certainty from information that allows us only to make general claims about possibilities. Although exact calculations are possible only in the former case, estimations are possible in the latter case.