S-ID.3 Describing Data Sets with Outliers

Task

For certain data sets, such as home prices and household or individual income, is often described using the median instead of the mean. The questions below explore the mean and median in some different situations to help you understand the information that they communicate.

a. Give an example of a set of five positive numbers whose median is 10 and whose mean is larger than 10.

b. Find the mean and median of the following set of numbers: $\{10, 15, 25, 30, 30, 50, 55, 55, 60, 80\}$. What happens to the mean and median of these numbers if 80 is replaced by 800?

c. The brightness of celestial bodies depends on many factors, two of the most important being the distance from Earth and size. The eight brightest objects in the night sky are listed below with their approximate distance from Earth (in light years).

Object	Distance in light years
Moon	0.000000038
Venus	0.0000048
Jupiter	0.000067
Mars	0.0000076
Mercury	0.0000095

Syrius	8.6
Canopus	310
Saturn	0.00014

Calculate the mean and median for these distances. Would the typical distance of these celestial bodies best be communicated using the mean or the median? Why?

d. What impact do the very large values in the data set have on the mean?

e. Suppose that a sample of 100 homes in the metropolitan Phoenix area had a median sales price of $300,000. The mean value of these homes was $1,000,000. Explain how this could happen. Why might the median price be more informative than the mean price in describing a typical house price?

f. Suppose the mean annual income for a sample of one hundred Minneapolis residents was $50,000. Do you think the median income for this sample would have been greater than, equal to, or less than $50,000? Explain.

IM Commentary

The goal of this task is to look at the impact of outliers on two important statistical measures of center: the mean and the median. In general, a small number of outliers have no impact on the median which is determined by the ''middle'' two data points when they are placed in increasing order. On the other hand, outliers can have a dramatic impact on the mean, especially if the data set is small or if these outliers are several orders of magnitude larger than most other data points. This is made particularly clear in the case of the distance of the stars. There are many interesting questions which the teacher can ask to stimulate conversation about the celestial bodies in part (c):

- Jupiter is about 9 times as far away from the earth as Mars but appears brighter in the sky. This means that Jupiter must be substantially larger than Mars and students may wish to investigate this.
- Given the shockingly large distance from the earth to the star Canopus, this must be an impressively large/bright star! What factors determine how ''bright'' a star appears in the sky? Size and distance must play a role but are there other factors?
- Are there other stars that are closer to the earth than Canopus but not as bright?

- The light we see today in the sky from Canopus was emitted in the very early 1700's!!! If some important changes took place to the star Canopus they would not be observed here until centuries later.

The information for the table on celestial bodies came from here for the stars http://en.wikipedia.org/wiki/List_of_brightest_stars and from here for the planets: http://www.pbs.org/deepspace/classroom/activity1.html

Solution

a. Answers will vary. One simple set of five numbers with a median of 10 would be $\{10, 10, 10, 10, 10\}$. Here, however, the mean is also 10. If we increase one of the 10's to 20, this will not impact the median while it will change the mean to 12. So $\{10, 10, 10, 10, 20\}$ is an example of a set of five numbers whose mean is larger than the median.

b. The sum of these ten numbers is 410 so the mean is 410 ÷ 10 = 41. The middle two numbers are 30 and 50 so the median is 40. If 80 is replaced by 800 then the sum of the ten numbers is now 1130 and the mean is 1130 ÷ 10 = 113. The median is still 40 as the middle two numbers have not changed, but the value of the mean increased.

c. Listed in increasing order of distance from the Earth the celestial bodies are: the moon, Venus, Mars, Mercury, Jupiter, Saturn, Syrius, Canopus. So for the median distance we take the average of Mercury and Jupiter, giving 0.0000382 light years. The mean distance is 39.825 light years.

A typical distance would be best communicated using the median. The mean distance is perplexing in this case as none of these celestial bodies is close to 40 light years away: most of them (the planets) are much closer, and one of them (Canopus) is much further away. The two stars, particularly Canopus, have greatly distorted the mean (relative to the median) because they are completely different orders of magnitude. The mean gives very little insightful information whereas the median tells us that the majority of the brightest objects in the sky are very close. This is clear in the case of the moon, less so in the case of the planets which do not appear, to the naked eye, much brighter (or ''closer'') than the brightest stars.

d. The two largest values in the data set (8.6 and 310) cause the mean to be much larger than most values in the data set. The mean is not a very good indication of a typical value.

e. To better evaluate the mean and median of the home prices we can begin by listing

them in non-decreasing size (with units of dollars): $h_1 \leq h_2 \leq \ldots \leq h_{100}$. We are given that the median home price is \$300,000. Since there are an even number of homes, the median is $\frac{h_{50}+h_{51}}{2}$. So we know that

$$h_{50} + h_{51} = 600,000.$$

The mean home price is the average of all of the home prices:

$$\text{Mean Home Price} = \frac{h_1 + h_2 + \ldots + h_{99} + h_{100}}{100}.$$

So if the mean home price value is \$1,000,000 this means that

$$h_1 + h_2 + \ldots + h_{99} + h_{100} = 100,000,000.$$

The only constraint the median puts on us is that $h_{50} + h_{51} = 600,000$. One way to satisfy these constraints would be to have $h_1, \ldots, h_{50}$ all be \$100,000, giving a total of \$5,000,000. If $h_{50} = \$100,000$ then we must have $h_{51} = \$500,000$ since these two homes together cost \$600,000. We could make $h_{51} = h_{52} = \ldots = h_{90} = \$500,000$. This adds \$20,000,000 in cost, leaving \$75,000,000 for the final ten homes. So we may take

$$h_{91} = \ldots = h_{100} = 7,500,000.$$

With this choice of home values, the median is \$300,000 but the mean is \$1,000,000. By adding in a small number of very expensive homes, the mean can be made as high as we wish with a given median value.

f. Answers will vary, but the explanation should depend on whether the student thinks that there will be unusual values in the data set. For example, if the student thinks that there may be a few people with very high incomes, they might argue that the median will be less than the mean.