

Suggested Answers for WarmUps for Lesson 06

You can find the article here: [www.indiana.edu/~econstat/pdf/The Median Isn't the Message.pdf](http://www.indiana.edu/~econstat/pdf/The%20Median%20Isn't%20the%20Message.pdf)

| | |
|--------|--|
| 1. | Gould's doctor anticipated that he would fall victim to what particular problem (often used to lie with statistics)? There is another example of this principle elsewhere in the article. Describe that situation. |
| Answer | The problem is incomplete or inadequate numerical representation. Citing a measure of center without a measure of spread or other context is a method often used to misrepresent the truth. The similar situation is that of the two politicians each quoting only one statistic and attempting to influence people with out complete information. |
| 2. | How did Gould avoid falling victim to the misperception you identified in the previous question? Don't quote him or repeat his actions, but briefly summarize his approach. |
| Answer | Gould analyzed the reported statistic based on his knowledge of the meaning of the median. He thought about its meaning as a measure of center, he thought about possible values and how they would be distributed that coincided with his knowledge of the meaning of the median. |
| 3. | What characteristic of distributions did Gould recognize that provided him additional solace? How did this characteristic change his understanding of his situation based on the known statistics? |
| Answer | Gould recognized that only a limited range of values below the median could be part of the distribution (truncated at 0) but that a virtually unlimited range of values could be possible above the median. He correctly concluded that the distribution must be right-skewed and that the upper limit had not been identified. He included information about the dispersion of the distribution in his analysis. He constructed and "read the graph correctly." |

Access this file.

| <i>Outfielder Salaries</i> | | <i>Pitcher Salaries</i> | |
|-----------------------------------|---|--------------------------------|--|
| Mean | 3.88 | Mean | =83.06/30 =2.769 |
| Median | 3.06 | Median | 2.85 |
| Standard Deviation | =SQRT(6.573)=2.564 | Standard Deviation | 1.04 Adjusted: =SQRT(1.045)=1.022 |
| Sample Variance | 6.8 adjust for population: =6.8*(29/30) =6.573 | Sample Variance | Calculate Variance and Adjust: =(1.04) ² *(29/30) =1.045 |
| Minimum | | Minimum | 0.76 |
| Maximum | | Maximum | 4.67 |
| Sum | =3.88*30=116.4 | Sum | 83.06 |
| Count | 30 | Count | 30 |
| Skewness | =3*(3.88-3.06)/2.564 = 0.959 | Skewness | =3*(2.769-2.85)/1.022 = - 0.238 |
| Coef. of Variation | =(2.564/3.88)*100 = 66.08% | Coef. of Variation | =(1.022/2.769)*100 =36.91% |

Relevant calculations are in red. Note that it is not always possible to fill in all the spaces in Descriptive Statistics tables, and some spaces you can fill in may be irrelevant.

| | |
|--------|--|
| 4. | Of Outfielder and Pitcher Salaries, which is more homogeneous? Explain your reasoning and any calculations you made to determine your answer. |
| Answer | <p>Homogeneous means "of the same kind" so the question is asking about dispersion, the type of summary statistic that addresses how close to one another observations in a data set are, how similar. While the range, variance and standard deviation all would measure this, the coefficient of variation is the most reliable measure of relative dispersion. CV is the ratio of standard deviation to the mean expressed as a percentage.</p> <p>In order to calculate this correctly both variances have to be adjusted to population variances ("all" teams), by multiplying by (29/30), then the square root must be taken of the population variance. The population standard deviation of Outfielder salaries is 2.56, and</p> |

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|--|---|----|-------------------------|------|-----------------|-------|--------------------|------|----|--------|----------|------|----|---|---|---|---|----|---|---|---|---|----|---|---|---|----|---|---|---|---|---|---|---|---|---|---|---|---|----|---|---|---|---|----|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|----|---|----|----|---|---|--|--|--|--|--|
| | <p>Pitcher salaries 1.02.</p> <p>The mean Pitcher salary must be calculated also. Since both variables represent observations from the same population (all Major League Baseball teams), the mean pitcher salary can be calculated by dividing the SUM of all pitcher salaries 83.06 by 30, giving a mean of 2.77 when rounded.</p> <p>Coefficients of variation can be calculated and are, respectively, 66.08% and 36.91%. Thus, pitcher salaries are more homogeneous because their relative dispersion is smallest. In other words, compared to outfielders, pitchers' salaries are more uniform, more like one another.</p> | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 5. | Referring to the data set you just looked at, describe as completely as possible the distribution of the two variables. Comment on the meaning of the data revealed in your descriptions | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Answer | <p>Outfielder salaries have a mean of 3.88 million dollars but a median of 3.06 million dollars, suggesting what is borne out by the skewness coefficient of 0.96 that at least one team's outfielders are being paid substantially more than the other teams. This likely contributes to the larger relative dispersion among outfielders when compared to pitchers. Thus outfielder salaries have a median of \$3.06M, a standard deviation of \$2.56M and are pretty right skewed.</p> <p>Pitcher salaries have a mean of \$2.77M and median of \$2.85M, indicating a slight negative skew, calculated to be -0.24. At least one clubs' pitchers are being paid at a slightly lower average rate than the other clubs. If a club is building its bull pen (hiring young, inexperienced pitchers) that might explain this slight skew. On the whole, however, pitchers' salaries are relatively similar compared to outfielders' salaries. Thus, pitcher salaries have a median of \$2.85M, a standard deviation of \$1.02M and a slight left skew.</p> | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 6. | Why does an individual statistic have little, if any, meaning? In what way can an individual "lie" by only citing individual statistics? | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Answer | A single value has no scale of comparison. | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 7. | <p>Access this file.</p> <table border="1"> <tr> <td colspan="10"><i>Lengths of 65 calls initiated during the last week of July</i></td> </tr> <tr> <td>1</td><td>2</td><td>10</td><td>5</td><td>3</td><td>3</td><td>2</td><td>20</td><td>1</td><td>1</td> </tr> <tr> <td>6</td><td>3</td><td>13</td><td>2</td><td>2</td><td>1</td><td>26</td><td>3</td><td>1</td><td>3</td> </tr> <tr> <td>1</td><td>2</td><td>1</td><td>7</td><td>1</td><td>2</td><td>3</td><td>1</td><td>2</td><td>12</td> </tr> <tr> <td>1</td><td>4</td><td>2</td><td>2</td><td>29</td><td>1</td><td>1</td><td>1</td><td>8</td><td>5</td> </tr> <tr> <td>1</td><td>4</td><td>2</td><td>1</td><td>1</td><td>1</td><td>1</td><td>6</td><td>1</td><td>2</td> </tr> <tr> <td>3</td><td>3</td><td>6</td><td>1</td><td>3</td><td>1</td><td>1</td><td>5</td><td>1</td><td>18</td> </tr> <tr> <td>2</td><td>13</td><td>13</td><td>1</td><td>6</td><td></td><td></td><td></td><td></td><td></td> </tr> </table> <p>Calculate the range, mean absolute deviation, variance, standard deviation, coefficient of variation and skewness for this data set. Compare and contrast the measures of spread, paying particular attention to the relationships between the various measures. Of these measures, which has the greatest influence on the skewness?</p> | <i>Lengths of 65 calls initiated during the last week of July</i> | | | | | | | | | | 1 | 2 | 10 | 5 | 3 | 3 | 2 | 20 | 1 | 1 | 6 | 3 | 13 | 2 | 2 | 1 | 26 | 3 | 1 | 3 | 1 | 2 | 1 | 7 | 1 | 2 | 3 | 1 | 2 | 12 | 1 | 4 | 2 | 2 | 29 | 1 | 1 | 1 | 8 | 5 | 1 | 4 | 2 | 1 | 1 | 1 | 1 | 6 | 1 | 2 | 3 | 3 | 6 | 1 | 3 | 1 | 1 | 5 | 1 | 18 | 2 | 13 | 13 | 1 | 6 | | | | | |
| <i>Lengths of 65 calls initiated during the last week of July</i> | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 2 | 10 | 5 | 3 | 3 | 2 | 20 | 1 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 6 | 3 | 13 | 2 | 2 | 1 | 26 | 3 | 1 | 3 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 2 | 1 | 7 | 1 | 2 | 3 | 1 | 2 | 12 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 4 | 2 | 2 | 29 | 1 | 1 | 1 | 8 | 5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | 4 | 2 | 1 | 1 | 1 | 1 | 6 | 1 | 2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3 | 3 | 6 | 1 | 3 | 1 | 1 | 5 | 1 | 18 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | 13 | 13 | 1 | 6 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Answer | <table border="1"> <tr> <td>Range</td><td>28</td> </tr> <tr> <td>mean absolute deviation</td><td>3.92</td> </tr> <tr> <td>Sample Variance</td><td>34.47</td> </tr> <tr> <td>Standard Deviation</td><td>5.87</td> </tr> <tr> <td>CV</td><td>131.15</td> </tr> <tr> <td>Skewness</td><td>1.27</td> </tr> </table> <p>The mean absolute deviation and the standard deviation are the same order of magnitude, which makes some sense considering that they are both looking specifically at average distance from mean to observation, however the standard deviation is 1.5 times the mean absolute deviation. The difference in the manner in which the two are calculated explains the difference. Remember that the standard deviation is the square root of the variance and the</p> | Range | 28 | mean absolute deviation | 3.92 | Sample Variance | 34.47 | Standard Deviation | 5.87 | CV | 131.15 | Skewness | 1.27 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Range | 28 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| mean absolute deviation | 3.92 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Sample Variance | 34.47 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Standard Deviation | 5.87 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| CV | 131.15 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Skewness | 1.27 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

variance sums squared deviations. The square root is calculated AFTER the squared deviations are summed, (a related problem is Example 3, page 4 of the lab manual.) The absolute deviations do not go through that process—we just use their absolute value—and the different process can make a big difference in the value calculated. The relationship between the variance and standard deviation is via the square root function, and comparing the two values to try to get more information is difficult. (This is why it is critical to compare standard deviation to standard deviation, variance to variance, and so on.) Curiously, the range and variance are the same order of magnitude, but then the coefficient of variation is 131%.

In summary, what have we learned? Comparing spread across measures is not a good idea. Each of these statistics was ostensibly measuring the same thing, but trying to make sense of all of them together gets us nowhere. The only way we can make sense of spread is by having either a given standard for comparison or another variable to compare this one to.

Which of the measures is most influential when it comes to skewness? Using Pearson's second skewness coefficient, it is clear that the ONLY spread measure used is the standard deviation.

8. Access this file (see above). Test the Empirical Rule using this data set. Do you think this data came from a bell-shaped, symmetric distribution? Cite your evidence.

Answer

| | |
|------------|----------|
| -3σ | -13.14 |
| -2σ | -7.27 |
| -1σ | -1.39 |
| mean | 4.476923 |
| $+1\sigma$ | 10.35 |
| $+2\sigma$ | 16.22 |
| $+3\sigma$ | 22.09 |

The table has the calculated values of plus and minus 1, 2 and 3 standard deviations from the mean. The first problem that leaps out is that all three of those that are below the mean are negative, and none of the data is negative. Furthermore, 72% (47 of 65) of the observations are below the mean, in fact they are between 1 and the mean. The Empirical Rule tells us we should have about 34% of the observations between the mean and -1σ , so the Empirical Rule is pretty much washed up. This data set is NOT bell-shaped nor symmetric. I created a graph which confirms this result.

