1. Calculate the following using the data set below:

7,2,6,6,6,4,3,4

1. What is *n*?
2. What is the median?
3. What is the mean?
4. What is/are the mode(s)?
5. What is the range?
6. What is the sum of squares?
7. What is the variance?
8. What is the standard deviation?
9. Calculate the following using the data set below:

1, 16, 2, 2, 9

1. What is *n*?
2. What is the median?
3. What is the mean?
4. What is/are the mode(s)?
5. What is the range?
6. What is the sum of squares?
7. What is the variance?
8. What is the standard deviation?
9. Which is the only measure of central tendency (mean, median, or mode) that can be used with qualitative/categorical data? *Write your response.*
10. Which is (are) a measure(s) of spread?
    1. Standard deviation
    2. Range
    3. Variance
    4. Sum of squares
    5. All of these
11. Which measure(s) of central tendency are most likely to be values that are actually present in the data set? *Write your response.*
12. Why is the SS (sum of squared errors) not such a useful measure on its own? *Write your response.*
13. When the standard deviation is small, it implies that
    1. there are multiple scores in the data set that are identical.
    2. the scores in the data set are scattered pretty widely from the mean score.
    3. very few scores lie above the mean score.
    4. the mean estimates the typical data point quite well.
14. Your sample has 399 observations in it. What *position* will the median score be in? *Calculate.*
15. The variance is less useful than the standard deviation because
    1. in every sample, it will always equal zero.
    2. it tends to be very large when the sample size is large.
    3. the units for the variance are squared units (e.g., friends squared).
    4. it takes more steps to calculate than the standard deviation.
16. Which is the only measure of central tendency that incorporates the value of every single score in your data set? *Write your response.*
17. When the distribution is normal, 50% of data points will lie above and 50% of the data points will lie below the
    1. mode
    2. standard deviation
    3. mean
    4. range
18. If the scores in your data set are 1, 5, 7, 8, and 9, and the mean is 6, the standard deviation could be: (*Answer this question without actually calculating the standard deviation. Use your understanding of the meaning of “standard deviation.”)*
    1. 0.75
    2. 2.5
    3. 6.3
    4. 10.4
19. Assume that the sum of squares from Sample A and the sum of squares from Sample B are identical: 29,860. Sample A is larger than Sample B. Which will have the lower standard deviation?
    1. Sample A
    2. Sample B
    3. It’s impossible to tell
20. Which measure of central tendency is most influenced by outliers? *Write your response.*
21. A negatively-skewed variable will have a tail toward the low value end of the x axis (tail closer to the origin of the graph) – true or false?
22. A positively-skewed variable will have a median with a lower value than the mean – true or false?
23. Outliers and skew are very similar concepts; if there is an outlier, that means the distribution is skewed – true or false?

*Scroll to next page for key.*

KEY

1.

1. What is *n*? **8 (there are 8 scores)**
2. What is the median? **5 (the mean of the 4th and 5th scores is 5. (4+6)/2=5)**
3. What is the mean? **4.75 (sum of all scores = 38, divided by 8 scores = 4.75)**
4. What is/are the mode(s)? **6 (6 is the most frequent score)**
5. What is the range? **5 (7-2 = 5)**
6. What is the sum of squares (aka, sum of squared errors, aka SS)? **21.5 (subtract 4.75 from each of your 8 scores, square all of those differences, and sum all of those squared differences)**

|  |  |  |  |
| --- | --- | --- | --- |
| SCORE | MEAN | deviance (error) | Squared error |
| 2 | 4.75 | -2.75 | 7.5625 |
| 3 | 4.75 | -1.75 | 3.0625 |
| 4 | 4.75 | -0.75 | 0.5625 |
| 4 | 4.75 | -0.75 | 0.5625 |
| 6 | 4.75 | 1.25 | 1.5625 |
| 6 | 4.75 | 1.25 | 1.5625 |
| 6 | 4.75 | 1.25 | 1.5625 |
| 7 | 4.75 | 2.25 | 5.0625 |
|  |  |  |  |
|  | *SUM=* |  | 21.5 |

1. What is the variance? **3.07 (divide the SS (sum of squares) by n – 1, so 21.5 / (8-1))**
2. What is the standard deviation? **1.75 (take the square root of the variance)**

2.

1. What is *n*? **5**
2. What is the median? **2**
3. What is the mean? **6**
4. What is/are the mode(s)? **2**
5. What is the range? **15**
6. What is the sum of squares? **166 (see table below)**
7. What is the variance? **41.5 (166 / (5-1) = 166 / 4 = 41.5)**
8. What is the standard deviation? **6.44 (take square root of 41.5)**

|  |  |  |  |
| --- | --- | --- | --- |
| SCORE | MEAN | deviance (aka Error) | Squared error |
| 1 | 6 | -5 | 25 |
| 2 | 6 | -4 | 16 |
| 2 | 6 | -4 | 16 |
| 9 | 6 | 3 | 9 |
| 16 | 6 | 10 | 100 |
|  |  |  |  |
|  | *SUM=* |  | 166 |

1. The only measure of central tendency that can be used with categorical data is the **mode.** The mode is defined as the score (or response) that occurs most frequently in the data set.If you had trouble with this question, think of a categorical variable, like college major (English, Biology, Business, Theatre, etc.). Data for a categorical variable indicates how many students are majoring in each of these disciplines (e.g., 15 in English, 25 in Biology, 33 in Business, 9 in Theatre, etc.). The mode of this data set is *the discipline* that *most students* are majoring in – “Business” is the mode. If you created a histogram with these data, Business would have the tallest bar; it is the most frequent major. Median and mean would be incorrect for the following reasons. For the median, you’d need to first *order* the data, and categorical data are not able to be ordered. For mean, you need to be able to add up the different responses and divide by the total number of responses. Here, you cannot add up the responses, because responses are individual majors, which are not numeric. It does not make sense to calculate: “English” + “biology” + “Business” + etc.
2. E. All of these measure dispersion, or spread of the data; some of these are more frequently used by researchers than others.
3. The **mode(s)** is(are) scores that are actually in your data set (unless there is no mode, because all scores occur equally frequently). The median will be in your data set in some cases, for example, *if* the number of scores in your data set is an odd number. If your data set contains an even number of scores, the median may or may not be in your data set. The mean may happen to be in your data set (e.g., your mean is 15 and you happen to have a “15” in your data set), but it doesn’t have to be.
4. The sum of squares is strongly affected by the size of the sample you’re working with, with samples that contain larger numbers of scores producing a larger sum of squares. The purpose of the sum of squares is to estimate *dispersion*, not size of sample. If we wanted to compare dispersion across two samples of unequal size, we could not use the sum of squares because, independent of amount of dispersion, the larger sample would produce a larger sum of squares most likely. For these reasons, researchers usually convert the SS into the variance and, more often, the standard deviation. Another reason the sum of squares is not particularly useful (and neither is the variance for this same reason) is that the sum of squares yields a value of “units-squared”, and it is not intuitive to think about units in units-squared (e.g., 54.3 friends-squared, or 4.3 hours-squared).
5. D. Remember that the standard deviation for a data set is a measure of the typical, or average, distance of our scores from the mean. When the standard deviation is small, it implies a relatively small average distance between each score in our data set and that data set’s mean. Another way of thinking about this is that the mean is a good estimate of the typical score in our data set (i.e., the mean is close to the typical score in your data set).
6. The 200th score. Formula for position of the middle score is: (N+1)/2. (399+1) / 2 = 200.
7. C. Just like the sum of squares (SS), the variance involves using squared deviances, which makes the units for the variance squared units. If you answered B, you’re probably thinking about the sum of squares, which does happen to get larger as sample size gets larger.
8. The **mean**. The mode only incorporates the value of the most common (most frequent) score. Even a multi-modal data set will have modes that reflect only a few values in your data set, but not all scores. The median only reflects the middle score (or the average of the two middle scores), which means the median at most incorporates only 2 values from your data set. The mean, on the other hand, uses all values for all scores, because computing the mean involves summing every single score in your data set (and then dividing by the number of scores).
9. C
10. Remember that the standard deviation represents **how far the typical score in the sample is from the mean score**. We know the mean is 6, so we’re left to estimate how far away from 6 is the typical score in our sample. Let’s look at the scores in our sample. 1 is the score farthest from 6, and it’s **5 units away** from the mean, 7 and 5 are the scores closest to 6, and they are both **1 unit away** from the mean. Standard deviation, again, is defined as how far the typical score is from the mean. Therefore, we know the standard deviation must be between 1 and 5 because all scores in our sample are between 1 and 5 units away from the mean. The only answer choice between 1 and 5 is choice ***B***. 2.5.
11. A. Sample A will have the lower standard deviation. Standard deviation is calculated by, first, taking the SS and dividing by the total sample size minus 1, and then taking the square root. If you divide 29,860 by a large number (e.g., Sample A’s sample size minus 1), you will get a smaller result than if you divide 29,860 by a small number (e.g., Sample B’s size minus 1). Therefore, Sample A will have a smaller (lower) standard deviation; Sample A’s scores are clustered more closely around its sample mean, than Sample B’s scores are clustered around its mean.
12. The mean is the measure of central tendency most sensitive to the presence of outliers.
13. TRUE. A negatively-skewed variable will have a tail toward the low value end of the x axis (tail closer to where the x and y axes meet).
14. TRUE. (Statement: A positively-skewed variable will have a median with a lower value than the mean.) A positive skew means a tail toward the larger values, and the mean will always be pulled more strongly than the median in the direction of the tail (in this case, toward the larger values).
15. FALSE. (statement: Outliers and skew are very similar concepts; if there is an outlier, that means the distribution is skewed.) While both outliers and skew can influence the mean in a similar direction, these concepts are not related in a one to one fashion. Skewed distributions may not have any outliers (scores that are very different from the others), like the examples on the left, below. And if data for a variable have just one or a few people who have extreme values, that does not produce a tail (which is more gradual). The graph below on the right has one value of 20 (outlier), but no tail (so, no skew).

