Content Specific Tips & Traps

1. **DESCRIBING DATA:** When you analyze one-variable data, always discuss **S**hape, **U**nusualfeatures, **C**enter**,** and **S**pread **(SUCS)**.

When commenting on shape:

* Symmetric is not the same as “equally” or “uniformly” distributed
* Do not say that a distribution is normal just because it looks symmetric and unimodal. **TREAT “NORMAL” AS A FOUR LETTER WORD**. You should use it only if you are really ABSOLUTELY sure that it is appropriate in the given situation.
* When referring to “normal”, **ALWAYS USE THE WORD “APPROXIMATELY”**
1. **PATTERNS:** Look for patterns in the data, and then for deviations from those patterns.
2. **CENTERS:** Don’t be confused by median and mean. They are both measures of center, but for a given data set, they may differ by a considerable amount.
* data is skewed **RIGHT**……..**mean greater than the median**
* data is skewed **LEFT**…..**mean less than median**.

BUT mean > median is NOT sufficient to show that the distribution is skewed right

AND mean < median is NOT sufficient to show that the distribution is skewed left.

1. **SPREAD: Don’t confuse standard deviation and variance**. Remember that the standard deviation units are the same as the data units, while variance is measured in square units.
2. **TRANSFORMATION:** Know how transformations of a data set affect summary statistics.

* **Adding** (or **subtracting**) the same positive number, k, to (from) each element in a data set increases (decreases) the mean and median by k. The standard deviation and IQR do not change.
* **Multiplying** all numbers in a data set by a constant k multiplies the mean, median, IQR, and standard deviation by k. For instance, if you multiply all members of a data set by 4, then the new set has a standard deviation that is 4 times larger than that of the original data set, but the variance is 16 times the original variance.

**Simple examples**:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Original data | Mean | St.Dev. | Variance | Median | IQR | Range |
| 1,2,3,4,5 | 3 | 1.414 | 2 | 3 | 2 | 4 |

**Add 7 to each element for the original set:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Original data | Mean | St.Dev. | Variance | Median | IQR | Range |
| 8,9,10,11,12 | 10 | 1.414 | 2 | 10 | 2 | 4 |

**Multiply each element of the original data set by 4**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Original data | Mean | St.Dev. | Variance | Median | IQR | Range |
| 4,8,12,16,20 | 12 | 5.6569 | 32 = 2\*42 | 12 | 8 | 16 |

**Multiply each element of the original data set by 4 then add 7.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Original data | Mean | St.Dev. | Variance | Median | IQR | Range |
| 11,15,19,23,27 | 19 | 5.6569 | 32 | 19 | 8 | 16 |

1. **CALCULATOR SKILLS:**
* Be able to use 1-var stats with and without summarized data

SYNTAX: 1-var stats list or 1-var-stats list, freqlist

* Be able to use the STATPLOT menu to create any of the plots available….be able manipulate the window if needed ..or Zoom 9
* Get the dot plot, stem plot programs.
* Be able to manipulate lists

Multiple Choice Questions

1. Which one of the following is true about a histogram?
2. It is exactly the same as a bar graph.
3. the number of intervals depends on the range of the data
4. accurate histograms can display either cutpoints or midpoints of the interval.
5. the area under a histogram is 1.
6. all of the above are false.
7. Dot plots
8. can be used for either discrete or continuous data
9. do not indicate spread of data as well as histograms
10. are the same as stem and leaf plots but do not give as much detail.
11. are excellent plots for discrete data sets.
12. are not as immediately informative as cumulative relative frequency plots.
13. Which is true of a boxplot?
14. one must have a box but may be missing one or both whiskers.
15. it is possible to have a box plot with no box or whiskers
16. it is possible to determine the data values if you only have a box plot.
17. if there is one outlier, then there are at least two whiskers
18. all of these are true for all boxplots.
19. A distribution that has the boxplot shown could be described as



1. skewed left
2. skewed right
3. roughly symmetric but with outliers
4. not possible
5. inconclusive
6. If the largest value of a data set is doubled, which of the following is not true?
7. the mean increases
8. the standard deviation increases
9. the interquartile range increases
10. the range increases
11. the median remains unchanged
12. Which of the following is reasonable about a distribution of the distances that employees drive to work when the mean is 13.1 miles and the standard deviation is 8.2 miles? Assume the distribution is symmetric and mound-shaped.
13. approximately 95% of the employees drives are within 16.4 miles of the mean value
14. the minimum and first quartile are the same value.
15. the distribution is skewed right.
16. the standard deviation is too large.
17. the distribution possesses outliers.
18. If the test scores of a class of 30 students have a mean of 75.6 and the test scores of another class of 24 students have a mean of 68.4, then the combined group mean is

a. 72 b. 72.4 c. 72.8 d. 74.2 e. none of these

1. Which observation has a higher z-score?
2. $x=25.4, \overbar{x}=12.9, s=3.7$
3. $x=137.5, \overbar{x}=73.7, s=17.1$
4. I
5. II
6. z-scores are equal
7. cannot be determined since we do not know the standard deviation of the population
8. cannot be determined since we don’t know if the populations are normal.
9. If and a score of 75 is in the 30th percentile, what is the percentile position of a score of 100?

 a. 45th b. 50th c. 63rd d. 75th e. cannot be determined

1. The standard deviation of the set {5,7,7,8,10,11} is 2. Which of the following sets also has a standard deviation equal to 2? This may be done without a calculator.
2. {4,5,5,8,12,14}
3. {2,4,6,8,10,12}
4. {3,5,5,6,8,9}
5. {10,14,14,16,20,22}
6. none of these.

**Free Response**

* 1. **The graph below displays the scores of 32 students on a recent exam. Scores on this exam ranged from 64 to 95 points.**

 6 \* \*

 6 \* \*

 7 \* \* \*

 7 \* \* \* \*

 8 \* \* \* \*

 8 \* \* \* \* \* \*

 9 \* \* \* \* \* \* \*

 9 \* \* \* \*

* + - * 1. Describe the shape of the distribution.
				2. In order to motivate her students, the instructor of the class wants to report that, overall, the class’s performance on the exam was high. What summary statistic, the mean or the median, should the instructor use to report that overall exam performance was high? Explain.
1. The midrange is defined as . Compute this value using the data above. Is the midrange considered a measure of center or a measure of spread? EXPLAIN.
	1. **The goal of a nutritional study was to compare the caloric intake of adolescents living in rural areas with the caloric intake of those living in urban areas. A simple random sample of ninth-grade students from one high school in a rural area was selected. Another random sample of ninth-graders from one high school in an urban area was also selected. Each student in each sample kept records of all the food he or she consumed in one day. The back-to-back stemplot displays the number of calories consumed per kilogram of body weight for each student in the study on a given day.**

 URBAN RURAL

 9 9 9 9 8 8 7 6 2

 4 4 3 1 0 3 2 3 3 4

 9 7 6 6 5 3 5 6 6 6 7

 2 0 4 0 2 2 2 4

 4 5 6 8 8 9

 5 1

1. Write a few sentences comparing the daily caloric intake of students in urban versus rural areas.
2. Is it reasonable to generalize the findings of this study to all rural and urban ninth-grade students in the United States? Explain.
3. Researchers who want to conduct a similar study are debating which of the following two plans to use.

Plan I: Have each student in the study record all the food he or she consumed in one day. Then researchers would compute the number of calories of food consumed per kilogram of body weight for each student for that day.

Plan II: Have each student record all the food he or she consumed over the same 7-day period. Then researchers would compute the average daily number of calories of food consumed per kilogram of body weight for each student during that 7-day period.

Assuming that the students keep accurate records, which plan, I or II, would better meet the goal of the study? Justify your answer.

* 1. **The amount of time a NHS AP Statistics student spends on homework each evening is approximately normally distributed. If 22% of the 75 students spend at least 47 minutes doing homework, and Erin, who spends 65 minutes has a standardized score of 1.932, determine:**
1. the mean and standard deviation of the distribution
2. the approximate number of students who spend less than 15 minutes on homework.

Multiple Choice KEY

|  |  |  |
| --- | --- | --- |
| 1 | **C** | NOT A GREAT QUESTION. |
| 2 | **D** | Self-explanatory |
| 3 | **B** | If all the values of the data is the same value, the IQR = 0, so there is no box. It is possible to have no whiskers. |
| 4 | **B** |  |
| 5 | **C** | The IQR is resistant |
| 6 | **A** | The Empirical Rule says, for a normal distribution, approximately 95% of the dta is within ±2σ. 2σ = 2(8.2) = 16.4. Aside: this is the only solution that makes sense. |
| 7 | **B** | Use LIST 1 for score and LIST 2 for Frequency.  |
| 8 | **B** | Score 1 = $\frac{25.4-12.9}{3.7}=3.378378$. Score 2 = $\frac{137.5-73.7}{17.1}=3.701754$ |
| 9 | **E** | Don’t know the distribution. You can calculate a z-score even if NOT normal. It’s just that we don’t know how many std.dev  |
| 10 | **C** | This data set is equal to the values in original data set less 2. This transformation does not change the variance or standard deviation. |

Free Response KEY

|  |  |
| --- | --- |
| **1.** | * 1. The shape of the distribution is skewed left. NOTE the question did not ask to describe the data set or the distribution, so using SUCS model would be an overkill.
	2. The instructor should use the MEDIAN as the measure of central tendency. The MEDIAN is an appropriate measure of center for non-symmetric distributions like this one. Also, the MEDIAN would be greater than the MEAN value since the data is skewed left.
	3. The midrange = ½ (95 + 64) = 79.5. The midrange would be a measure of center because it’s an estimate of the midpoint of the distribution. It would not be considered a measure of spread since it is not providing and “range” of values. The midrange is a point of reference, not a measure of dispersion.

  |
| **2.** | 1. The distribution of caloric intake for the sample from the urban high school is skewed right, compared to the distribution from the rural high school which is symmetric and approximately uniform. The measure of center for the urban sample is the median of 32, compared to the mean of the rural distribution being 40.45. The spread of the data is broader for the rural, which has a range of 19 vs. a range for the urban of 16. There are no outliers in either data sets.
2. Using the results to generalize the findings of this study across the US would be unwise. First, there is no indication that the schools were picked at random. While the students were picked at random within the school, they were all grade nine student, which would not necessarily represent the high school as a whole, which in turn would make it difficult to infer anything beyond the grade 9 age group. Finally, there may be confounding data that could affect the data: family income (the higher the income, the greater the ability to buy food); gender (there may be differences between male and female consumption); school location (different foods consumed in different parts of the country).
3. Plan II would be a better plan of data collection. There can be considerably more variability in measuring daily intakes (Plan 1) compared to the use of averages and lower variability over a 7-day period (Plan 2).
 |

|  |  |
| --- | --- |
| **3.** | 1. From Erin, we know that the std. z-score is 1.932 for x = 65 minutes of study. So, $\frac{x- μ}{σ}=Z$ and $\frac{65- μ}{σ}=1.932$. OR **65 - μ = 1.932σ.**

From the stmt of 22% spend at least 47 minutes, we have $\frac{47- μ}{σ}=0.77219$ where 0.77219 is the INVNORM (0.78, 0, 1). Thus, **47 - μ = 0.77219σ.** Using two linear equations and two unknowns, we have **μ = 35.02 minutes** and σ = **15.52 minutes**. Don’t forget the units of measure. Checking our math, P(X>47) for N(35.02, 15.52) is 0.22, and the Z-score for Erin would be $\frac{65- 35.02}{15.52}=1.932$1. P(X < 15) = P(Z < $\frac{15- 35.02}{15.52}$) = P (Z < -1.30678) = 0.0985343

0.0985343 x 75 students = 7.39 ≈ 7 studentsThus, the estimated that 7 student studied than 15 minutes. Much too high! |