Classroom Voting Questions: Elementary Statistics

Describing Distributions with Numbers

1. In a certain university there are three types of professors. Their salaries are approximately normally distributed within each of the following types:

   - Assistant Professors make a median salary of \$50K, with a minimum of \$40K and a maximum of \$60K.
   - Associate Professors make a median salary of \$65K per year, with minimum of \$57K and a maximum of \$80K.
   - Full Professors make a median salary of \$90K per year, with a minimum of \$70K and a maximum of \$110K.

   There are 1600 total Professors at this University, with the following distribution: 50% of all Professors are Assistants, 30% are Associates, and 20% are Fulls.

   What can we say about the average salary at this university?

   (a) mean $<$ median

   (b) mean $=$ median

   (c) mean $>$ median

   (d) insufficient information

   Answer: (c). Note: To answer this question correctly, the student needs to recognize the shape of the distribution. Since the salary distribution is more heavily populated on the lower end, the student should recognize the we have a right-skewed distribution, which leads to the median being less than the mean.

   (A) Students may not understand the relationship between the shape of the distribution and measures of central tendency.

   (B) Students may assume mean and median are always equivalent.

   (C)* correct Since half of the population consists of assistant professors, so they are going to define the median salary. They have the lowest salary, and since their max salary is \$60K so the median cant be higher than \$60K. If the 50% were lower, like 20%, the reasoning could possibly change. So, if the assistant professors were 20% and the full professors were 50%, then the skew would be reversed.

   (D) Students do not understand minimum amount of information needed to determine shape.

   by Murphy, McKnight, Richman, and Terry

STT.01.02.010

CC HZ MA336 S10: 32/8/**51**/5
AS DH MA3321 Su12: 40/0/**60**/0 time 2:20

2. Many individuals, after the loss of a job, receive temporary pay unemployment compensation until they are re-employed. Consider the distribution of time to re-employment as obtained in an employment survey. One broadcast reporting on the survey said that the average time until re-employment was 4.5 weeks. A second broadcast reported that the average was 9.9 weeks. One of your colleagues wanted a better understanding of the situation and learned (through a Google search) that one report was referring to the mean and the other to the median and also that the standard deviation was about 14 weeks. Knowing that you are a statistically-savvy person, your colleague asked you which is most likely the mean and which is the median?

   (a) 4.5 is the mean and 9.9 is the median.

   (b) 4.5 is the median and 9.9 is the mean.

   (c) Neither (A) nor (B) is possible given the SD of the data.

   (d) I am not a statistically-savvy person, so how should I know?

   Answer: (b). (A) This answer would imply that the distribution is left-skewed, which it is not (cf. (B)).

   (B)* correct The data must be right-skewed since the distribution is truncated at 0 weeks on the left-side of the distribution. Data that are truncated at one-end tend to have a skew in the direction away from the truncated end.

   (C) Students are thinking that the distribution must be normal. This thinking is very low-level and requires remediation.

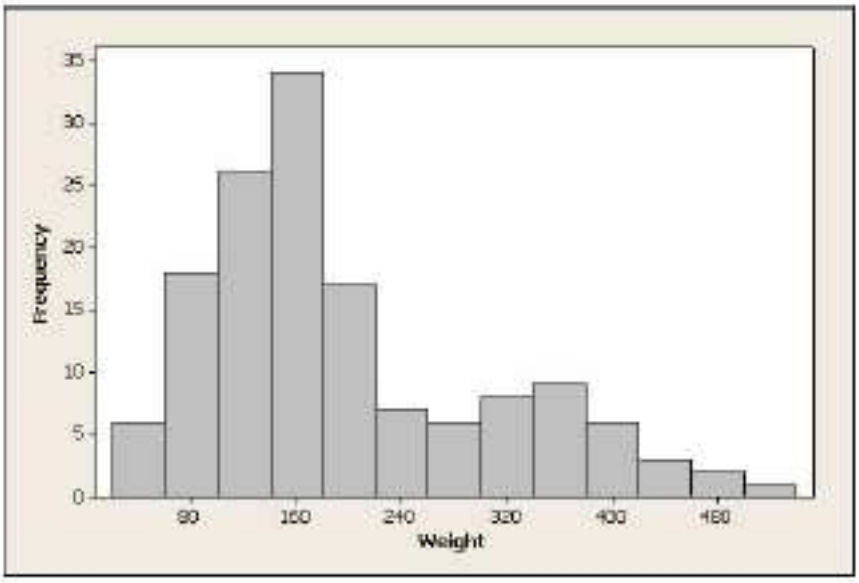
   (D) Students who give this answer need additional help to become statistically savvy.

   by Murphy, McKnight, Richman, and Terry

   STT.01.02.020

   CC HZ MA336 S10: 11/**35**/27/27
   AS DH MA3321 Su12: 7/**27**/53/13 time 3:30
   AS DH 3321 010 F14: 65/**12**/8 time 15 ,

3. For the data set displayed in the following histogram, which would be larger?



(a) mean

(b) median

(c) Can't tell from the given histogram.

Answer: (a).

by Roxy Peck for the textbooks: Roxy Peck and Jay Devore, Statistics: The Exploration and Analysis of Data, 6th Edition, Brooks/Cole Cengage Learning 2008 and Roxy Peck, Chris Olsen and Jay Devore, Introduction to Statistics and Data Analysis, 3rd Edition, Brooks/Cole Cengage Learning 2008.

STT.01.02.030

CC HZ MA207 F09: **45**/45/10 time 1:50
AS DH MA3321 Su12: **73**/13/13 time 2:30
AS DH MA1333 010 F12: **47**/53/0 time 2:30
AS DH MA1333 020 F12: **78**/11/11 time 2:00
AS DH 1333 010 S13: **90**/10/0 time 3:00
AS DH 1333 020 S14: **82**/18/0 time 2:20 ,
AS DH 3321 010 S14: **83**/17/0 time 1:30 ,
AS DH 1333 010 F14: **95**/5/0 time 2:30 ,
AS DH 3321 010 F14: **79**/21/0 time 2:20 ,
AS DH 1333 020 S15: **88**/12/0 time 2:20 ,
AS DH 3321 010 S15: **89**/11/0 time 2:20 ,

4. Why is the term $(n-1)$ used in the denominator of the formula for sample variance?

(a) There are $(n-1)$ observations.

(b) There are $(n-1)$ uncorrelated pieces of information.

(c) The $(n-1)$ term gives the correct answer.

(d) There are $(n-1)$ samples from the population.

(e) There are $(n-1)$ degrees of freedom.

Answer: (e). (A) There are n observations, not $n-1$.

(B) The formula does not require uncorrelated pieces of information.

(C) This statement is true but does not answer "Why?".

(D) The formula for sample variance is not related to the number of samples from the population.

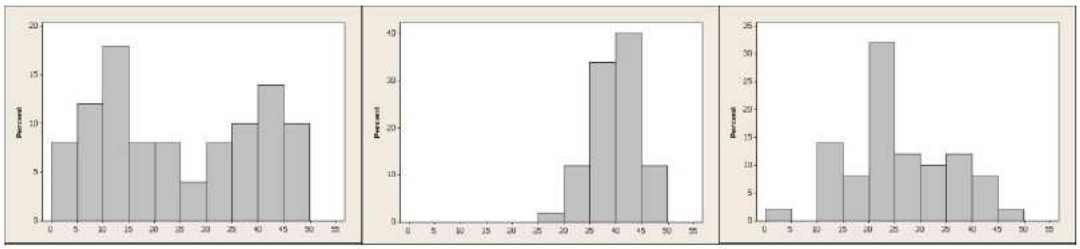
(E)* correct Use of $n-1$ makes sample variance an unbiased estimator for population variance. If we have any statistic that uses the mean in its calculation, we only need $n-1$ of the data pieces to determine the other information since we can use algebra to rearrange the formula for sample mean: $\bar{x} = \frac{x_1+\cdots+x_n}{n}$.

by Murphy, McKnight, Richman, and Terry

STT.01.02.040

AS DH MA3321 Su12: 7/7/67/7/**13** time 2:50

5. Which of the three histograms shown summarizes the data set with the smallest standard deviation?



Answer: (b). Graph 2 shows the smallest standard deviation.

by Roxy Peck for the textbooks: Roxy Peck and Jay Devore, Statistics: The Exploration and Analysis of Data, 6th Edition, Brooks/Cole Cengage Learning 2008 and Roxy Peck, Chris Olsen and Jay Devore, Introduction to Statistics and Data Analysis, 3rd Edition, Brooks/Cole Cengage Learning 2008.

STT.01.02.050

CC HZ MA207 F09: 33/**53**/14
AS DH MA3321 Su12: 33/**60**/7 time 1:40

AS DH MA1333 010 F12: 35/**60**/5 time 1:40
AS DH MA1333 020 F12: 6/**82**/12 time 2:00
AS DH 1333 020 S14: 0/**100**/0 time 2:00 ,
AS DH 3321 010 S14: 4/**96**/0 time 2:00 ,
AS DH 1333 010 F14: 19/**81**/0 time 2:30 ,
AS DH 3321 010 F14: 17/**83**/0 time 1:30 ,
AS DH 1333 020 S15: 17/**78**/4 time 2:00 ,
AS DH 3321 010 S15: 0/**100**/0 time 2:10 ,

6. Suppose your statistics instructor tells you that you scored 70 on an exam and that the class mean was 74. You should hope that the standard deviation of exam scores was _____.

   (a) Small

   (b) Large

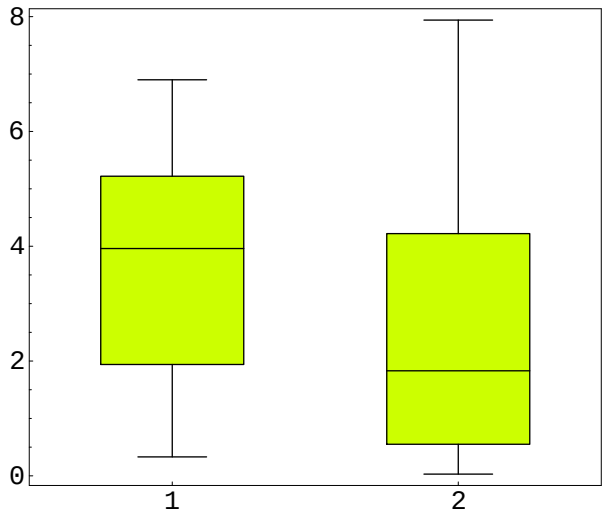
   Answer: (b).

   by Roxy Peck for the textbooks: Roxy Peck and Jay Devore, Statistics: The Exploration and Analysis of Data, 6th Edition, Brooks/Cole Cengage Learning 2008 and Roxy Peck, Chris Olsen and Jay Devore, Introduction to Statistics and Data Analysis, 3rd Edition, Brooks/Cole Cengage Learning 2008.

   STT.01.02.060

   CC HZ MA207 F09: 57/**43** time 1:10
   AS DH MA3321 Su12: 87/**13** time 2:20
   AS DH MA1333 010 F12: 35/**65** time 1:40
   AS DH MA1333 020 F12: 35/**65** time 2:10
   AS DH 1333 020 S14: 65/**35** time 2:30 ,
   AS DH 1333 010 F14: 68/**32** time 2:50 ,
   AS DH 3321 010 F14: 46/**54** time 2:10 ,
   AS DH 1333 020 S15: 83/**17** time 2:10 ,
   AS DH 3321 010 S15: 25/**75** time 2:00 ,

7. Below are boxplots for two data sets.



TRUE or FALSE: There is a greater proportion of values outside the box for the set on the right than for the set on the left.

(a) True, and I am very confident.

(b) True, and I am not very confident.

(c) False, and I am not very confident.

(d) False, and I am very confident.

Answer: (False). These are boxplots, so the box represents the middle 50% of data in both cases, meaning that what's outside of the box is also 50% in both cases. (The only exception is if the data set has a lot of repeated values right at the first or third quartile. These values would be "in" the box and could increase the proportion of data in the box beyond the standard 50%).

by Derek Bruff

STT.01.02.070

CC HZ MA207 F09: 70/**30** time 0:40
AS DH MA3321 Su12: 7/**93** time 0:50
AS DH MA1333 010 F12: 0/56/39/**6** time 2:20
AS DH MA1333 020 F12: 10/48/29/**14** time 3:00
AS DH 1333 010 S13: 32/47/5/**16** time 3:00
AS DH 1333 020 S14: 9/41/32/**18** time 2:00 ,
AS DH 3321 010 S14: 5/14/32/**50** time 1:30 ,
AS DH 1333 010 F14: 6/24/18/**53** time 2:50 ,
AS DH 3321 010 F14: 0/25/32/**43** time 2:20 ,
AS DH 1333 020 S15: 25/13/31/**31** time 2:30 ,
AS DH 3321 010 S15: 6/60/14/**20** time 2:30 ,

8. The five-number summary for all student scores on an exam is 29, 42, 70, 75, 79. Suppose 200 students took the test. How many students had scores between 42 and 70?

   (a) 25

   (b) 28

   (c) 50

   (d) 100

   Answer: (c). Note: The five-number summary represents the min, 25th percentile, median, 75th percentile, and max, respectively and in order. Since 42 is the 25th percentile score and 70 is the median score, then 25% must have had scores between 42 and 70.

   (A) 25 is the percentage of the sample of scores that is between 42 and 70, but the question asks for the number (not percentage).

   (B) 28 is the difference between 42 and 70, which does not give the number of students.

   (C)* correct 25% of n = 200 students is 50.

   (D) As 70 is the median, there are 100 students whose scores are below 70.

   by Murphy, McKnight, Richman, and Terry

   STT.01.02.080

   CC HZ MA207 F09: 68/0/**26**/5 time 1:30
   AS DH MA3321 Su12: 7/13/**60**/20 time 2:00
   AS DH MA1333 010 F12: 11/44/**39**/6 time 3:10
   AS DH MA1333 020 F12: 38/0/**57**/5 time 2:50
   AS DH 1333 010 S13: 13/13/**65**/9 time 2:50
   AS DH 1333 020 S14: 0/19/**81**/0 time 2:50 ,
   AS DH 3321 010 S14: 0/9/**82**/9 time 3:00 ,
   AS DH 1333 010 F14: 0/24/**59**/18 time 3:00 ,
   AS DH 1333 020 S15: 0/35/**65**/0 time 3:40 ,

9. The five-number summary for all student scores on an exam is 40, 60, 70, 75, 79. Suppose 500 students took the test. What should you conclude about the distribution of scores?

   (a) Skewed to the left.

   (b) Skewed to the right.

   (c) Not skewed.

   (d) Not enough information given to determine skew.

Answer: (a). Note: The five-number summary represents the minimum, the first quartile, the median, the third quartile, and the maximum, respectively and in order.

(A)* correct By examining the scores, the student should be able to recognize that the scores are more squished together at the top end of the scale, and more spread out at the bottom end of the scale. For example, there is only a 9-point difference between the median and the maximum score, but a 30-point difference between the median and the minimum score. Thus, the distribution must be left-skewed.

(B) If the distribution were right-skewed, then the first quartile would be closer than the third quartile to the median or the minimum would be closer than the maximum to the median.

(C) If the distribution were not skewed, then the first and third quartiles would be approximately equi-distant from the median and the minimum and maximum would be approximately equi-distant from the median.

(D) Students may default to this kind of answer when they dont understand, but there is, in fact, sufficient information given to determine skew.

by Murphy, McKnight, Richman, and Terry

STT.01.02.090

CC KC MA207 F09: **52**/45/3/0 time 2:00
AS DH MA3321 Su12: **85**/15/0/0 time 2:20
AS DH 1333 010 S13: **87**/9/0/4 time 3:00
AS DH 1333 020 S14: **74**/23/3/0 time 3:00 ,
AS DH 1333 010 F14: **71**/26/3/0 time 2:40 ,
AS DH 3321 010 F14: **93**/4/0/4 time 2:30 ,
AS DH 1333 020 S15: **96**/4/0/0 time 2:30 ,
AS DH 3321 010 S15: **97**/3/0/0 time 2:20 ,

10. Jack uses a calculator to find the sample standard deviation of a data set and ends up getting a negative number as the result. This implies that

    (a) on average, the deviations from the mean are negative.

    (b) the mean of the deviations is negative.

    (c) the mean is negative.

    (d) Jack made a mistake.

    Answer: (d).

    (a) (Students might find it easier to discuss answer choice (b) first.) The average deviation from the mean is always 0.

    (b) Students can recall that the deviations always sum to 0. This is, in fact, the essence of the mean.

(c) Students can consider whether shifting a data set on the real number line changes 1) the standard deviation and 2) the mean.

(d) Jack has made a mistake somewhere in his calculation. Formally: Students can recall that "square roots are always positive," and note that there is no plus-or-minus sign in front of the square root. More conceptually: Standard deviation is a (non-negative) distance, the measure of "on average, how far the data points are from the mean." [Follow-up question: What if you get a negative number inside the square root?]

by David A. Huckaby

STT.01.02.100

11. Which set of two observations would you expect to have the smaller standard deviation, the weights of two randomly-selcted professional ballerinas, or the weights of two randomly-selected sumo wrestlers?

(a) The weights of the ballerinas, because ballerinas are lighter.

(b) The weights of the sumo wrestlers, because sumo wrestlers are heavier.

(c) The weights of the ballerinas; their weights are more likely to be closer together.

(d) The weights of the ballerinas; their weights are more likely to be farther apart.

(e) The weights of the sumo wrestlers; their weights are more likely to be closer together.

(f) The weights of the sumo wrestlers; their weights are more likely to be farther apart.

Answer: (c). The population distribution of ballerina weights has a smaller standard deviation than the population distribution of sumo wrestler weights.* Sample distributions look like the population distribution from which they are drawn. We therefore expect the two ballerina weights to be closer together than the two sumo wrestler weights.

*Establishing this point will perhaps constitute the bulk of the discussion. Students might sketch the distributions and then imagine drawing two weights from each, or students might use or view a computer simulation. Answer choices (a) and (b) are included primarily to catch any misconceptions that the pertinent concept is central tendency rather than spread. Of course the relative sizes of the means play a role in establishing the answer to this voting question. Most students who had the correct intuition but answered either (a) or (b) will probably admit—especially in light of the fact that there is no answer choice of "more than one of the above"—that (c) is the best answer. [Follow-up question: We understand that the data set $\{4, 5, 6\}$ has a smaller standard deviation than the data set $\{40, 50, 60\}$. But shouldn't we be able to say something like, "Yes, but they have the same spread relative to the size of the data?" Can you come up with a way of quantifying this? Answer: How about dividing the standard deviation by the mean? This ratio is called the coefficient of variation.]

by David A. Huckaby

STT.01.02.110

12. When a professional sumo wrestler joined a group of people, the standard deviation of the weights of the new group members was substantially less than the standard deviation of the weights of the original group members. Which of the following is most likely?

    (a) The original group consisted of 3 professional sumo wrestlers.

    (b) The original group consisted of 100 professional sumo wrestlers.

    (c) The original group consisted of 3 professional ballerinas.

    (d) The original group consisted of 100 professional ballerinas.

    Answer: (a).

    (a) Imagine three sumo wrestlers the range of whose weights is fairly large. Adding a fourth sumo wrestler whose weight is about the mean of the three weights would substantially decrease the standard deviation.

    (b) The larger the group, the less influence an additional member will have on the standard deviation.

    (c) The weight of the added sumo wrestler will be much greater than the mean of the 3 ballerinas' weights. An outlier is being added to a small data set. The standard deviation would increase substantially.

    (d) The weight of the added sumo wrestler will be much greater than the mean of the 100 ballerinas' weights. An outlier is being added to a large data set. The standard deviation would increase, but not as much as in (c).

    by David A. Huckaby

    STT.01.02.120

13. A multi-billionaire decides to retire back in the small town in which she grew up. All of the houses in this town are modest and inexpensive. On the outskirts of town, she builds a huge, luxurious mansion. Consider house prices in the town before and after she builds her mansion. Which of the following measures of central tendency changes the most?

    (a) mean

    (b) median

    (c) mode

Answer: (a). The mean is not resistant to the influence of outliers, while the median is. The mode (if there was a mode to begin with) is not changed with the addition of one observation whose value is not in the orginal data set. [Follow-up question: What would you estimate the probability to be that there was no mode to begin with? Answer: Recall that if all of the numbers in a data set are unique, then the set has no mode. So our question is: What is the probability that at least two houses share a common price? Can you think of factors that would affect the probability? How about the size of the town? The standard deviation of the prices? The tendency to round house prices to the nearest thousand dollars?]

by David A. Huckaby

STT.01.02.130

14. Which of the following measures is resistant to the influence of outliers?

    (a) mean

    (b) median

    (c) standard deviation

    (d) Q_3

    (e) interquartile range

    (f) two out of (a) through (e)

    (g) three out of (a) through (e)

    (h) four out of (a) through (e)

    Answer: (g).

    (a) The mean is not resistant to outliers.

    (b) The median is resistant to outliers.

    (c) The standard deviation is not resistant to outliers. (This can be illustrated with a small, compact data set to which an extreme outlier is added. A more general, though loose, plausibility argument: If we view standard deviation as a sort of mean of distances ("On average, how far are the observations from the mean?"), then the fact that the mean is not resistant suggests that the standard deviation will not, in general, be resistant, either.)

    (d) The third quartile is a resistant measure. (Q_3 is a median. Since a median is resistant to outliers, so is Q_3. The example in (c) can also be used to illustrate this, as long as the data set used is not *too* small.)

    (e) The interquartile range is resistant to outliers. (Q_1 and Q_3 are both medians, hence resistant to outliers. Thus so is their difference. The example in (c) can again be used to illustrate this, again as long as the data set used is not *too* small.)

    (f) Incorrect. Choices (b), (d), and (e) are all valid.

(g) Correct. Choices (b), (d), and (e) are valid.

(h) Incorrect. Only choices (b), (d), and (e) are valid.

by David A. Huckaby

STT.01.02.140

15. In a history class with over 500 students, a professor gave a very easy test, so that the distribution of scores was highly left-skewed. Which measure of central tendency and measure of variation should be used to summarize the scores?

(a) median and standard deviation

(b) median and interquartile range

(c) mean and standard deviation

(d) mean and interquartile range

Answer: (b). The mean and standard deviation form a natural pairing, as do the median and interquartile range. In the absence of a compelling reason to do otherwise, these are the only pairings that should be considered. That leaves choices (b) and (c). Since the mean and standard deviation are not resistant to outliers, they should not be used due to the fat left tail of the distribution. (In general, the mean and standard deviation should not be used for highly-skewed distributions.) This leaves (b), the median and interquartile range, which are resistant measures.

by David A. Huckaby

STT.01.02.150

16. Consider a data set that consists of the following four numbers: 2, 5, 6, and a certain number that is less than negative one million. For this data set, rank the following from least to greatest: mean, median, standard deviation.

(a) mean, median, standard deviation

(b) mean, standard deviation, median

(c) median, mean, standard deviation

(d) median, standard deviation, mean

(e) standard deviation, mean, median

(f) standard deviation, median, mean

Answer: (a). The median is clearly 3.5. The mean and standard deviation cannot be found precisely, but it can be seen that the mean is a negative number, while the standard deviation is a positive number much larger than 3.5. [Follow-up questions: Can you see why the mean is a negative number? Maybe consider the data set $\{0, 0, 0, -1000000\}$. Can you tell what the mean is? (Ideally students can do more

than just perform the simple calculation here but also have a feel for what the mean truly is and can see what the mean is by, for example, looking at a number line and using a "weighting" argument.) Using the interpretation of the standard deviation as "on average, how far the data points are from the mean," can you see roughly what size of number the standard deviation is (and, of course, what sign it has)? Is the distribution in the question left-skewed or right-skewed? Answer: Left-skewed. So would you expect the mean to be larger or smaller than the median? Answer: Smaller.]

by David A. Huckaby

STT.01.02.160