
Problem 1

1. **The head circumference (in centimeters) of 15 college-age males was obtained, resulting in the following measurements: 55, 56, 56, 56.5, 57, 57, 57, 57.5, 58, 58, 58, 58.5, 59, 59, 63. If the last measurement (63 cms) were incorrectly recorded as 73, which one of the following statistics would change?**

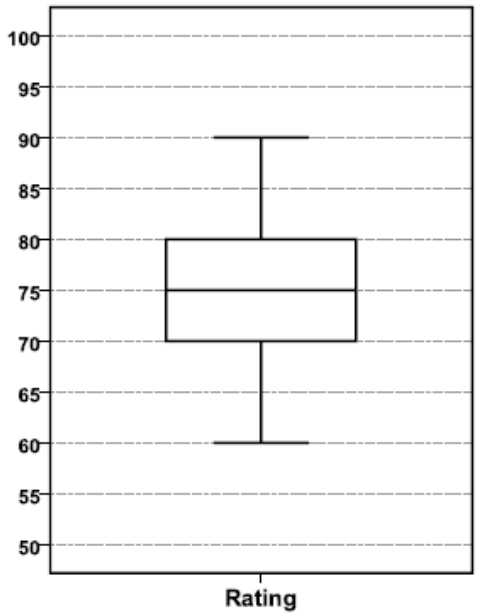
a) Q1 (1st quartile)
b) Standard deviation
c) Median
d) Q3 (3rd quartile)

The standard deviation is a measure of the amount of variation or dispersion in a set of values. A low standard deviation means that the values tend to be close to the mean (average) of the set, while a high standard deviation means that the values are spread out over a wider range. Changing a single value from 63 to 73 would change the mean of the data points, as well as the overall spread of the data points.

Problem 2

2. **The following boxplot gives the distribution of the ratings of a new brand of peanut butter for 50 randomly selected consumers (100 points possible with higher points corresponding to a more favorable rating).**

Identify the five-number summary and the description of each of them using the example.



Min - 60 points
Q1 - 70 points
Median - 75 points
Q3 - 80 points
Max - 90 points

Problem 3

This is a standard deviation contest. You must choose four numbers from the whole numbers 0 to 10, with repeats allowed.

3. **Choose four numbers that have the smallest possible standard deviation.**

a) 7, 7, 7, 8
b) 3, 5, 7, 9
c) 1, 1, 1, 1
d) 1, 2, 3, 4

Recall standard deviation is a measure of spread. This being said, the data with the smallest spread will also have the smallest standard deviation. As all four observations are the same in answer choice C, there is no spread and the standard deviation, σ, equals 0.

4. **Choose four numbers that have the largest possible standard deviation.**

a) 0, 3, 6, 10
b) 9, 9, 10, 10
c) 1, 4, 7, 10
d) 0, 0, 10, 10

Recall the formula for standard deviation. If we square the distance between each observation and the mean value for each answer choice, add the squared distances, divide by 3 (the number of observations - 1 in each answer choice), and take the square root, then we can see answer choice D will have the largest value. Numerically,

SD FOR A) $\sqrt{1/3\,[\,(0-4.75)^2+(3-4.75)^2+(6-4.75)^2+(10-4.75)^2\,]} = 4.272$

SD FOR B) $\sqrt{1/3[(9-9.5)^2+(9-9.5)^2+(10-9.5)^2+(10-9.5)^2]}=.577$

SD FOR C) $\sqrt{1/3[(1-5.5)^2+(4-5.5)^2+(7-5.5)^2+(10-5.5)^2]} = 3.87$

SD FOR D) $\sqrt{1/3[(0-5)^2+(0-5)^2+(10-5)^2+(10-5)^2]} = 5.773$

Problem 4

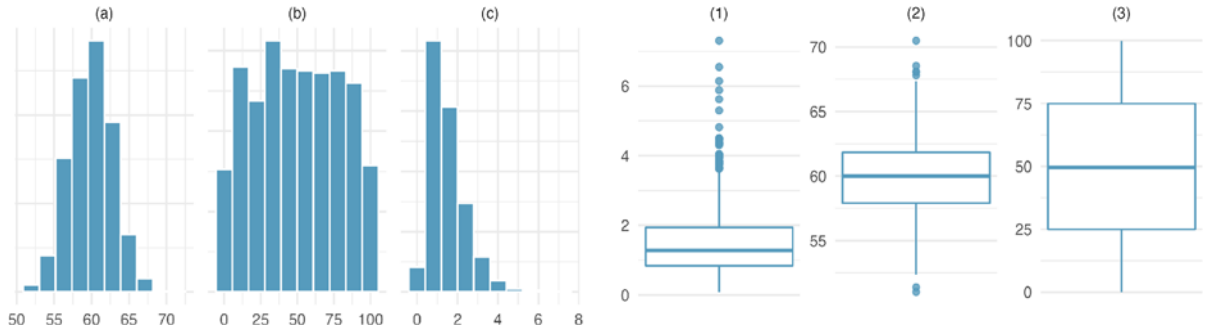
Researchers are interested in how crime rates are different for southern states.

5. **The standard deviation of unemployment for 35 to 39-year-old for southern states is 8.01 and the standard deviation for non-southern states is 8.76. What do these standard deviations tell you about the southern and non-southern states?**

a) The average unemployment rate for 35 to 39-year old in southern states is about the same as in non-southern states.
b) The average unemployment rate for 35 to 39-year old in southern states is slightly lower than in non-southern states.
c) The average spread from the mean rate for 35 to 39-year-old in southern states is lower than the non-southern states.

Problem 5

6. **Describe (in words) the distribution in the histograms below and match them to the box plots.**



The histogram **(a)** can be matched to boxplot **(2)**
The histogram **(b)** can be matched to boxplot **(3)**
The histogram **(c)** can be matched to boxplot **(1)**

By looking at the location of the distributions, we can immediately assign histogram C to boxplot 1 (note the median is much less than the other two distributions). To distinguish between the remaining two distributions we should look at the tails of the distributions. It does not appear that histogram B has any outliers, thus we should pair it with boxplot 3. Therefore, histogram A should be paired with boxplot 2. Note, when we look at the tails of the distributions, it confirms that we paired histogram C correctly, as the paired boxplot displays notable skewness.

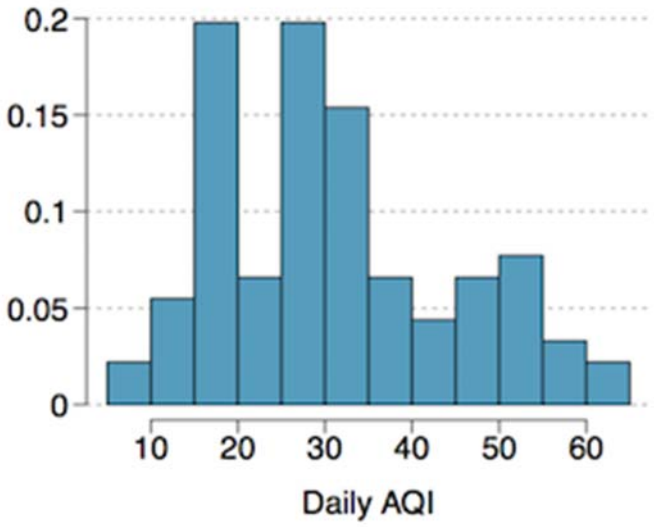
Problem 6

7. **College students in a statistics class were asked how many hours of television they watch per week, including online streaming services. This sample yielded an average of 8.28 hours, with a standard deviation of 7.18 hours. Is the distribution of the number of hours students watch television weekly symmetric? If not, what shape would you expect this distribution to have? Explain your reasoning.**

No, we would expect this distribution to be right skewed. There are two reasons for this: there is a natural boundary at 0 (it is not possible to watch less than 0 hours of TV) and the standard deviation of the distribution is very large compared to the mean.

Problem 7

Daily air quality is measured by the air quality index (AQI) reported by the Environmental Protection Agency. This index reports the pollution level and what associated health effects might be a concern. The index is calculated for five major air pollutants regulated by the Clean Air Act and takes values from 0 to 300, where a higher value indicates lower air quality. AQI was reported for a sample of 91 days in 2011 in Durham, NC. The histogram below shows the distribution of the AQI values on these days.



8. **Estimate the median AQI value of this sample.**

Between 25 and 30

9. **Would you expect the mean AQI value of this sample to be higher or lower than the median? Explain your reasoning.**

Recall, the direction of the skewness pulls the mean with it. Thus, since the distribution is right skewed the mean is pulled to the right and is higher than the median.

10. **Estimate Q1, Q3, and IQR for the distribution.**

Q1: between 15 and 20,
Q3: between 35 and 40,
IQR: between 20 and 25

Each bar's height represents the proportion of days that fall within a certain AQI range. To estimate Q1, you need to figure out where the lower 25% of the data lies in terms of AQI. Start from the lowest AQI range and add up the proportions represented by each bar. The first quartile (Q1) is the value below which 25% of the data fall. So, you're looking for the point where the cumulative proportion first reaches or exceeds 0.25 (25%).The first bin has about 0.04, the second bin has about 0.08, and the third bin has 0.20. These add up to 0.32. So the first three bins, or between 0 and 20, contain a proportion .32 of the data. Thus we know that Q1, which is a proportion of 0.25, is between the second and third bin, thus 15 and 20.

We would do the same thing for Q3, except we are looking for where the higher 25% of the data lies. IQR is the difference between Q3 and Q1.

11. **Would any of the days in this sample be considered to have an unusually low or high AQI? Explain your reasoning.**

Values that are considered to be unusually low or high lie more than 1.5×IQR away from the quartiles.

Upper fence: Q3 + 1.5 × IQR = 37.5 + 1.5 × 20 = 67.5;

Lower fence: Q1 - 1.5 × IQR = 17.5 + 1.5 × 20 = −12.5;

The lowest AQI recorded is not lower than 5 and the highest AQI recorded is not higher than 65, which are both within the fences. Therefore none of the days in this sample would be considered to have an unusually low or high AQI.

Problem 8

In a class of 25 students, 24 of them took an exam in class and 1 student took a make-up exam the following day. The professor graded the first batch of 24 exams and found an average score of 74 points with a standard deviation of 8.9 points. The student who took the make-up the following day scored 64 points on the exam.

12. **Does the new student's score increase or decrease the average score?**
    Decrease, the new value is lower than the previous mean
13. **What is the new average?**
    73.6; we can find this by doing the following: (24/25)(74) + (1/25)(64)
14. **Does the new student's score increase or decrease the standard deviation of the scores?**
    Increase

Problem 9

In each of the following situations, which is the explanatory variable and which is the response variable? Are they categorical or quantitative (quantitative means "numerical")?

15. **The typical number of calories a person consumes per day and that person's percent of body fat.**
    - **a)** Number of calories consumed per day: response, quantitative. Percent of body fat: explanatory, quantitative.
    - **b)** Number of calories consumed per day: explanatory, quantitative. Percent of body fat: response, quantitative.
    - **c)** Number of calories consumed per day: response, quantitative. Percent of body fat: explanatory, categorical.
    - **d)** Number of calories consumed per day: explanatory, categorical. Percent of body fat: response, categorical.

16. **Water temperature is controlled at different levels and growth (measured by weight) of corals in aquariums.**
    a. Water temperature: response, quantitative. Growth: explanatory, categorical.
    b. Water temperature: explanatory, categorical. Growth: response, categorical.
    c. Water temperature: response, categorical. Growth: explanatory, quantitative.
    d. Water temperature: explanatory, quantitative. Growth: response, quantitative

Problem 10

Coffee is a leading export from several developing countries. When coffee prices are high, farmers often clear forests to plant more coffee trees. Here are data on

prices paid to coffee growers in Indonesia and the rate of deforestation in a national park that lies in a coffee-producing region, for five years

Price (cents per pound)	Deforestation (percent)
29	0.49
40	1.59
54	1.69
55	1.82
72	3.10

17. **Coffee is currently priced in dollars. If it were priced in euros, and the dollar prices in the above table were translated into the equivalent prices in euros, would the correlation between coffee price and percent deforestation change?**
    - **a)** The correlation would remain zero, because the two variables are independent
    - **b)** Yes, units affect correlation
    - **c)** No, units do not affect correlation
    - **d)** It is impossible to calculate the correlation, because coffee price is categorical.

Problem 11

A study shows that there is a positive correlation between the size of a hospital (measured by its number of beds (x)) and the median number of days (y) that patients remain in the hospital.

18. **What lurking variable could be present in this study?**
    - **a)** cost: it's more expensive to run larger hospitals.
    - **b)** severity of disease: since large hospitals have better facilities and more doctors to cope with severe illness.
    - **c)** number of visitors: since larger hospitals receive more visitors.
    - **d)** facilities: since larger hospitals have better facilities, patients choose to stay longer

Problem 12

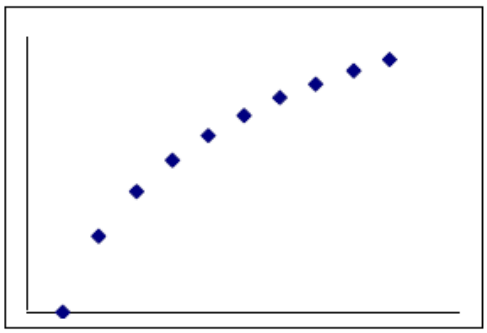
Milk use is positively correlated to cancer rates. While this is not a popular finding within the milk industry, there is a moderately positive correlation with drinking milk and getting cancer (Paulos, 1990). Milk consumption is greater in wealthier countries. In wealthier countries people live longer. Greater longevity means people live long enough to eventually get some type of cancer.

19. **Which is the response and explanatory variable?**
    Response – cancer. Explanatory- milk consumption
20. **Which is a lurking variable?**
    Longevity
21. **Will you conclude that drinking more milk increases the chance of getting cancer? Explain your reasoning.**
    No, because it is an observational study and it doesn't considered the lurking variables.

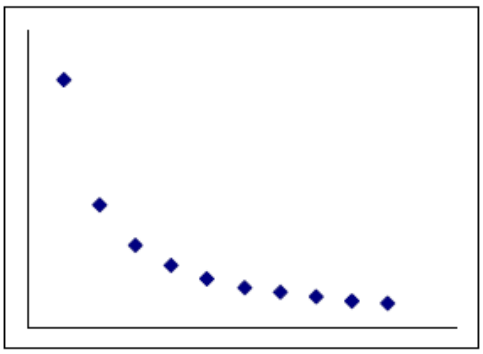
Problem 13

22. **Which of the following plots will have a correlation coefficient of .85? A**

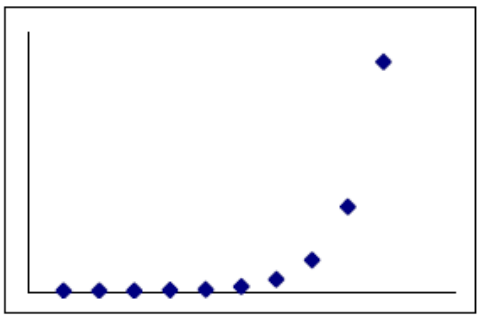
A.



B.



C.



D.