DALITE Q2 - Boxplots, Standard Deviation and Normal Curves Solutions.

EPIB607 - Inferential Statistics[a]

[a]Fall 2018, McGill University

This version was compiled on September 18, 2019

This DALITE quiz will cover more descriptives such as boxplots, standard deviation, and introduce you to normal density curves.

Boxplots | Standard deviation | Normal curves

1. Boxplot properties Q1

A boxplot can show whether a data set is:

a) symmetric
b) skewed
c) **symmetric and skewed (Correct)**

1.1. Correct rationales.

- When the data is skewed, the box will be shifted towards one of the whiskers (maximum or minimum). Symmetric data will have a median that splits the box in half.
- If the data set is skewed, the median locates above or below the center of the box plot, and the box locates closer to the maximum or minimum values.

1.2. Incorrect rationales.

- A boxplot can show the mean, quartiles which can tell us about being symmetric
- Mean, quartiles, and the max/min values in the plot help to show the symmetry and skewness of the data

2. Boxplot properties Q2

If one side of the box is longer than the other, it means that side contains more data.

a) TRUE
b) **FALSE (Correct)**

2.1. Correct rationales.

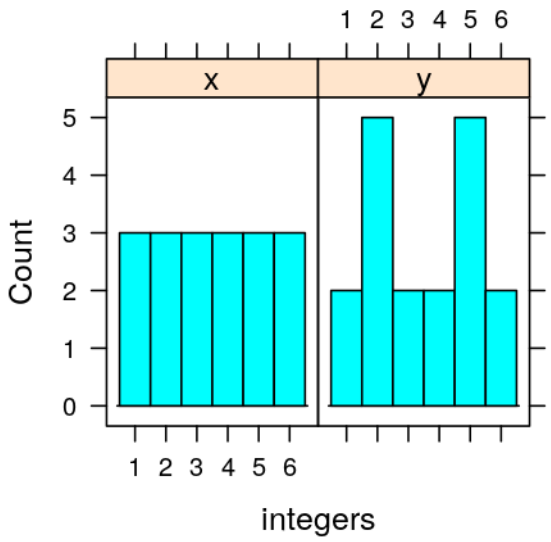
- The box is created by first finding the median which is the value half way between your ordered data. The quartiles are found by taking the medians of the upper and lower half of your data. Therefore, each quartile or side of the box contains the same amount of data it is just that if one side has larger values it will cause the boxplot to be skewed, making the box appear longer on one side.
- There is the same amount of data/observations in each quartile. The size of the quartiles is an indication of the spread of the observations within that quartile.

2.2. Incorrect rationales.

- The quartiles represent the number of data located within 25%, 50% 75% positions. Therefore the longer the box, the greater the number of data located within that percentile

3. Boxplot properties Q3

The figure below shows histograms from two different data sets, each one containing 18 values that vary from 1 to 6. The histogram on the left has an equal number of values in each group, and the one on the right has two peaks at 2 and 5. Which of the following statements is true?



a) The boxplots for each histogram will be different.
b) **The boxplots for each histogram will be the same (Correct)**
c) There is not enough information to tell us if the boxplots will be the same or if they will be different.

3.1. Correct rationales.

- They will be the same because the 5 summary statistics (min, Q1, median, Q3, and max) will be the same and thus, both boxplots will be visually identical.
- Both histograms are symmetric therefore the boxplots will be the same and won't difference in data distribution

3.2. Incorrect rationales.

- In the first histogram the values are equal which would create a very symmetric boxplot. Whereas in the second histogram, there are two counts which are higher than the rest and that would produce a more uneven boxplot.
- Although the median for both distributions is the same, the spread around the median differs, and this would account for differences in quartiles and therefore the shape of the boxplots

4. Question : Standard Deviations Q1

Researcher 1 takes a sample of 100 men age 18-24 in a certain town. In the same town, Researcher 2 takes a sample of 1000 men age 18-24. Which of the following statements is true?

a) The average height for the sample collected by Researcher 2 will be bigger than the average height for the sample collected by Researcher 1
b) The standard deviation of heights for the sample collected by Researcher 2 will be smaller than the standard deviation of heights for the sample collected by Researcher 1
c) The sample collected by Researcher 1 will likely contain the tallest of the 1,100 men.
d) **The sample collected by Researcher 2 will likely contain the shortest of the 1,100 men. (Correct)**

4.1. Correct rationales.

- Because Researcher 2 is sampling from more of the population he/she more likely to include the shortest and tallest of the population.
- this is true because having a bigger sample may allow greater inclusion of the outliers
- Over 90% of the population will be captured by Researcher 2, so the sample will likely contain the shortest of the men. Standard deviation will not necessarily be smaller because it is the SD of the sample, not of the sample means.

4.2. Incorrect rationales.

- Since sample size is the denominator to calculate standard deviation, a larger sample size will yield a smaller standard deviation.
- Because as you increase the sample size, you get the same or similar values more often.
- Assuming that heights are relatively normally distributed, collecting more samples would reduce the spread of their distribution (this can be visualized in the formula - when n increases, s decreases)

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

5. Standard Deviations Q2

If you add 7 to each entry on a list of numbers (which contains both positive and negative integers), that adds 7 to the standard deviation.

a) TRUE
b) **FALSE (Correct)**

5.1. Correct rationales.

- Adding a constant, 7, to all data values shifts the location of the data but does not affect its spread. -In the numerator of the formula for standard deviation, the added 7s would cancel out:

$$((x+7)-(\bar{x}+7))^2 = (x-\bar{x})^2$$

5.2. Incorrect rationales.

6. Normal Curves Q1

To completely specify the shape of a normal distribution, you must give:

a) the mean and standard deviation (Correct)
b) the five-number summary (min, Q1, median, Q3, max)
c) the mean and the median
d) the mean and the interquartile range

6.1. Correct rationales.

- This is a basic characteristic of normal distributions. If normal, they all have predictable properties and allow us to understand a great deal about them with these two values.

6.2. Incorrect rationales.

7. Normal Curves Q2

Which of the following statements is false regarding normal curves?

a) the mean of a normal density curve shifts the curve along the horizontal axis without changing its shape
b) increasing the standard deviation produces a flatter and wider bell-shaped curve and that decreasing the standard deviation produces a taller and narrower curve
c) area under a density curve over an interval represents the proportion of data that falls in that interval
d) **unlike the average, the standard deviation is not sensitive to outliers. (Correct)**

7.1. Correct rationales.

- The standard deviation is a measure of spread, if you have outliers, your data is more spread, thus increasing your standard deviation.

7.2. Incorrect rationales.