Data Analytics (CS3203N)

Problem Set III (Topic:Descriptive Statistics)

Leaño, Jomar
Rosales, Jade
Samson, Jose Glen
Stewart, Christian

I. Concept Questions

1. If all values in a sample are the same constant (say c) what is the standard deviation? What is the mean? Does the mode exist?

   If all values in a sample are the same constant (c), then the standard deviation would be 0. This is because there is no variation in the sample. The formula for standard deviation involves computing the differences between each value and the mean. This means that in this case, every difference would be 0 which would result in a standard deviation of 0.

   The mean of the sample would be equal to the constant value (c), since the mean is calculated by summing up all the values in the sample and dividing by the total number of values. In this case, the sum of all values in the sample is c times the number of values in the sample, so the mean is also equal to c.

   The same applies to the mode, since the mode is the value that appears most frequently in the sample. All the values are the same constant (c).

2. The arithmetic mean of the 15 customer orders is 54. Find the new (combined) arithmetic mean in each of the following situations:

   (a) A new order for amount 70 is received.

   (b) An order for amount 38 is cancelled.

   (c) 3 new orders with mean = 56 are received.

   Arithmetic mean = total customer orders / number of customers
   a= (15 * 54 + 70) / 16
   = 880 / 16
   = 55.06

   b= (15 * 54 - 38) / 14
   = 772 / 14
   = 55.14

   c= ((15 * 54) + (3 * 56)) / 18
   = 978 / 18
   = 54.33

3. Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

   (a) What is the *mean* of the data? What is the *median*? **Mean: 29.96 ; Median: 25**

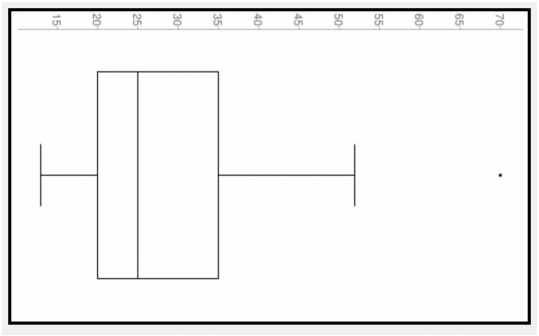
   (b) What is the *mode* of the data? Comment on the data's modality (i.e., bimodal, tri-model, etc.).
   Mode 25,35 both appeared 4 times which makes the data's modality bimodal

   (c) What is *midrange* of the data? **41.5**

   (d) Can you find (roughly) the first quartile (Q1) and he third quartile (Q3) of the data?
   First Quartile: 20
   Third Quartile: 35

   (e) Give the *five-number summary* of the data.
   Minimum: 13
   Quartile Q1: 20
   Median: 25
   Quartile Q3: 35
   Maximum: 70

   (f) Show a *boxplot* of the data.

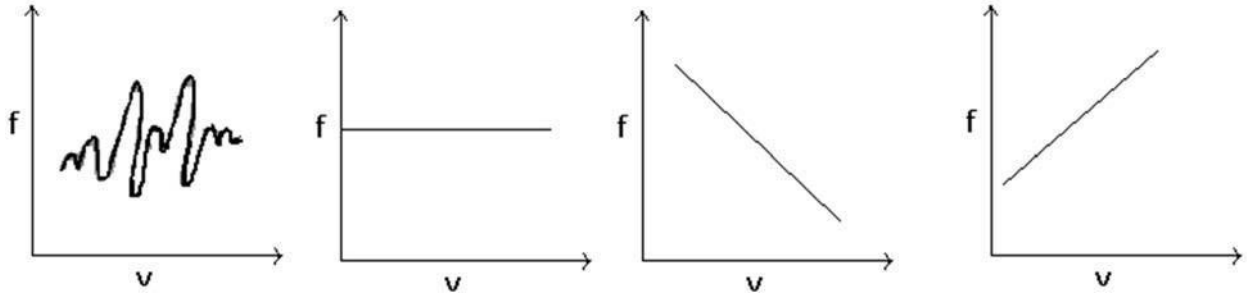


4. **Which** of the following measures of central tendency allow: a) distributive, b) algebraic and c) holistic measures:
   (a) Mean => **Algebraic**
   (b) Median => **Holistic**

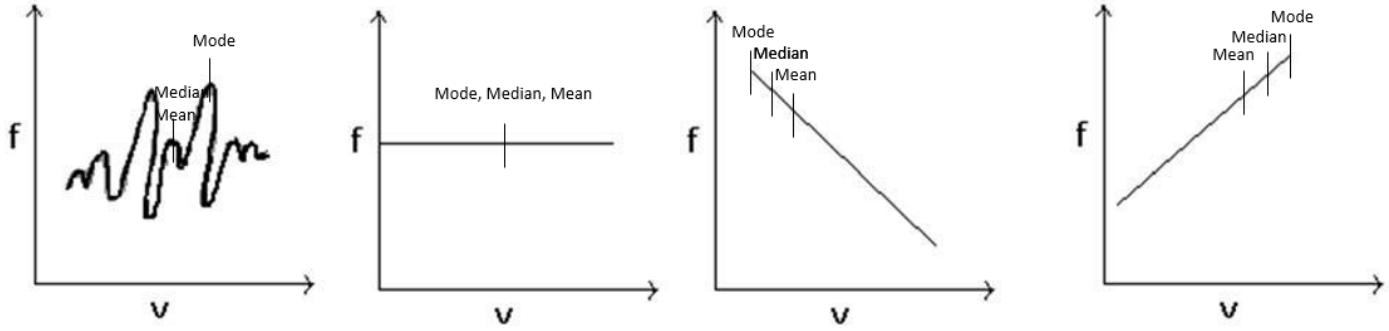
(c) Mode => **Holistic**

Which measure is faster as compared to the other? **Median is faster than others**

5. Suppose frequency distribution of two samples are shown in the following graphs:

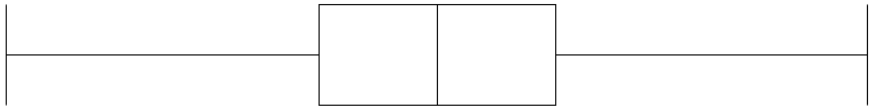


Locate the position of 1) Mean 2) Median 3) Mode in each of the above mentioned graphs.

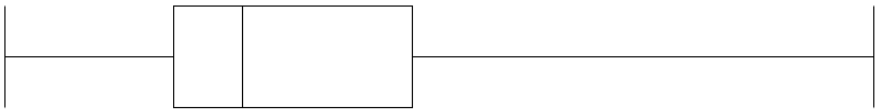


6. Given a sample, how to decide whether it is :
   (a) Symmetric: **It is symmetric if mean, median and mode is in the same point**
   (b) Skew-symmetric(+ve or -ve): **For positive skew-symmetric: mode < median < mean**
   For negativity skewed-symmetric: mean<median< mode

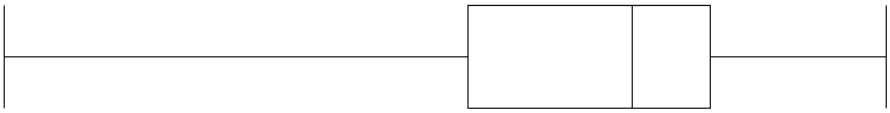
7. How the box-plot will look like for the following type of samples:
   (a) Symmetric



   (b) Positive skew-symmetric



(c) Negative skew-symmetric



8. Variance of a sample X={x1, x2, x3, …., xn} is calculated using the following formula:

$$\text{Variance} = \frac{1}{n-1}\sum_{i=1}^{n}(x^{i} - $$

where x' is mean(x).

In the above formula, why is (*n*-1) in the denominator instead of *n*?

The example above uses (*n*-1) in the denominator because it is only looking for an estimate of the variance of the sample as opposed to only using *n* which is used to compute the exact variance.

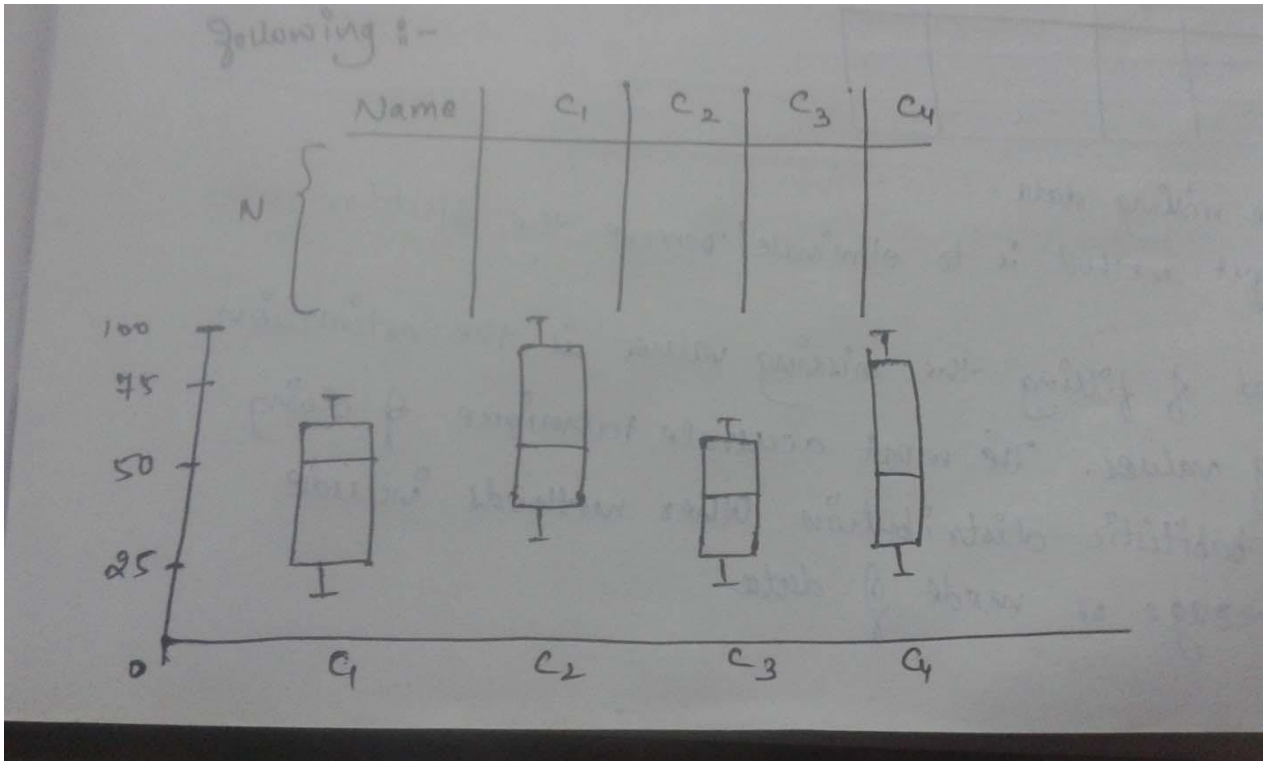
9. The standard deviation of the sample X is zero, is it possible? If possible, then what it does mean? Under what type of distribution of data in X it is possible? Give an example.

If the standard deviation of a sample X is zero, this means that all the values in the sample are the same hence there is no variation in the data.

It is possible for the standard deviation of a sample to be zero, but this only occurs when all the values in the sample are identical. This can happen with any distribution of data, as long as all the values in the sample are the same.

Consider a sample of 8 test scores: 80, 80, 80, 80, 80, 80, 80, 80. The mean of this sample is also 80, and the standard deviation is zero because all the values are the same.

10. From the tabulation of marks of students participated in four courses c1,c2,c3 & c4, box- plots are shown in the following figure:



From the plots give answer to the following questions:

a. In which course the performance of the student is very good? **C2**

b. In which course the performance of the student is very bad? **C3**

c. Which course(s) give(s) better average performance? **C2**

II Objective Questions

1. The scores of eight persons in an IQ test were:

   95 87 96 110 150 104 112 110

   The median is:

   (a) 107
   (b) 110
   (c) 112
   (d) 104
   (e) None of the above.

2. If the interquartile range is zero, you can conclude that:

   (a) the range must also be zero
   (b) the mean is also zero
   (c) at least 50% of the observations have the same value
   (d) all of the observations have the same value
   (e) none of the above is correct.

3. The "average" type of grass used in Texas lawns is best described by

   (a) the mean
   (b) the median
   (c) the mode.
   (d) the standard deviation

4. A sample of 100 IQ scores produced the following statistics: mean = 95 lower quartile = 70 median = 100 upper quartile = 120 mode = 75 standard deviation = 30

   Which statement(s) is (are) correct?

   (a) Half of the scores are less than 95.
   (b) The middle 50% of scores are between 100 and 120.
   (c) One-quarter of the scores are greater than 120.
   (d) The most common score is 95.

5. Identify which of the following is a measure of dispersion:

   (a) median
   (b) 90th percentile
   (c) interquartile range
   (d) mean

6. A sample of pounds lost in a given week by individual members of a weight reducing clinic produced the following statistics:

mean = 5 pounds,	first quartile = 2 pounds
median = 7 pounds,	third quartile = 8.5 pounds
mode = 4 pounds,	standard deviation = 2 pounds

Identify the correct statement:

(a) One-fourth of the members lost less than 2 pounds.
(b) The middle 50% of the members lost between 2 and 8.5 pounds.
(c) The most common weight loss was 4 pounds.
(d) All of the above are correct.
None of the above is correct.

7. A measurable characteristic of a population is:

(a) a parameter
(b) a statistic
(c) a sample
(d) an experiment.

8. What is the primary characteristic of a set of data for which the standard deviation is zero?
(a) All values of the variable appear with equal frequency.
(b) All values of the variable have the same value.
(c) The mean of the values is also zero.
(d) All of the above are correct.
(e) None of the above is correct.

9. Let X be the distance in miles from their present homes to residences when in high school of individuals at a class reunion. Then X is:

(a) a categorical (nominal) variable
(b) a continuous variable
(c) a discrete variable
(d) a parameter
(e) a statistic.

12 A subset of a population is:

(a) a population
(b) a statistic
(c) a sample
(d) none of the above.

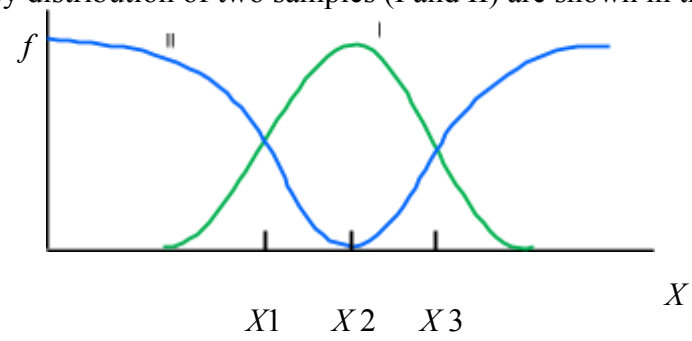
13. The median is a better measure of central tendency than the mean if:

(a) the variable is discrete
(b) the distribution is skewed
(c) the variable is continuous
(d) the distribution is symmetric
(e) none of the above is correct.

14. A set of data points follow a simple linear relation y= 3x + 2, where x is any integer number. The mean of the values of y for all values of x in the range [1 ... 100] is

(a) 50
(b) 50.5
(c) 152
(d) 152.5

15. Suppose frequency distribution of two samples (I and II) are shown in the following figure:



(a) The means, medians and modes for both I and II will be located at X2.
(b) The means of both I and II are at X1 and median and mode of II are at X1 and X3, respectively.
(c) The means of both I and II are at X1 and mode and median of II are at X1 and X3, respectively.
(d) Data II does not have neither median nor mean.

16. Number of wickets obtained by a bowler in 10 Test matches are shown in the following table.

Number of wickets	0	1	2	3	4
Number of Test matches	1	3	4	1	1

The mode of the above observation is

(a) 1
(b) 2
(c) 3
(d) 4