Data Analytics (CS40003)

Autumn Semester, 2016-2017 Session

Practice Set III

(Topic: Descriptive Statistics)

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?

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 is received.

3. Find the median and mode of the following data:

   (a) X = [5,2,3,2,5,5]

   (b) Y = [70,65,90,70,80,75,95]

4. Using the scaling transformation, find the GM of the following data.

   (a) X = [50,65, 72, 84, 90, 98]

   (b) Y = [22000, 25000, 34000, 40000, 63000]

5. The BMI and systolic BP of 6 patients are as follows:

   (BMI, BP) = [(24,114), (25,112), (22,110), (27,251), (26,132), (23,139)].

   Find the covariance.

6. Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows.

   Compute an *approximate median* value for the data.

Age	Frequency
1-5	200
5-15	450
15-20	300
20-50	1500
50-80	700
80-110	44

7. 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*?

(b) What is the *mode* of the data? Comment on the data's modality (i.e., bimodal, tri-model, etc.).

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

(d) Can you find (roughly) the first quartile (Q1) and he third quartile (Q3) of the data?

(e) Give the *five-number summary* of the data.

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

8. In many applications, new data sets are incrementally added to the existing large data sets. Thus an important consideration for computing descriptive data summary is whether a measure can be computed efficiently in incremental manner. Use *count*, *standard deviation*, and *median* as examples to show that a distributive or algebraic measure facilitates efficient incremental computation, whereas a holistic measure does not.

9. Which of the following measures of central tendency allow: a) distributive, b) algebraic and c) holistic measures:

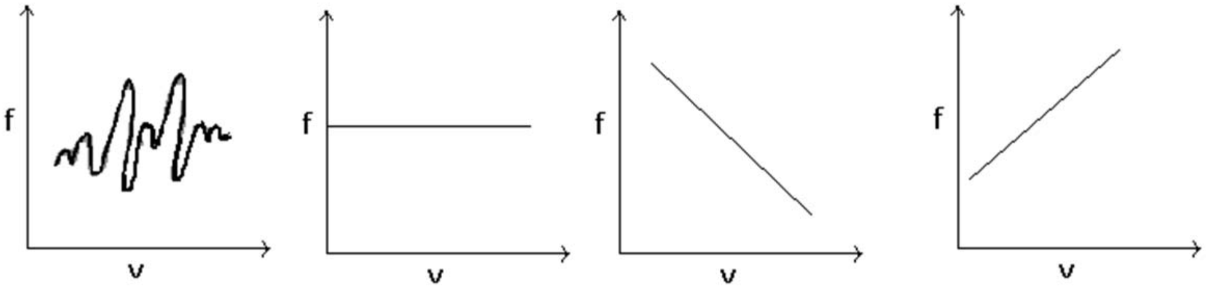
(a) Mean
(b) Median

(c) Mode

Which measure is faster as compared to the other?

10. Give three situations where AM, GM and HM are the right measures to provide a better central tendency.

11. 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.

12. Given a sample, how to decide whether it is :

(a) Symmetric
(b) Skew-symmetric(+ve or -ve)
(c) In-variate

13. How the box-plot will look like for the following type of samples:

(a) Symmetric
(b) Positive skew-symmetric
(c) Negative skew-symmetric
(d) In-variate data

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

$$Variance = \frac{1}{n-1}\sum_{i=1}^{n}(x'-x)^2$$

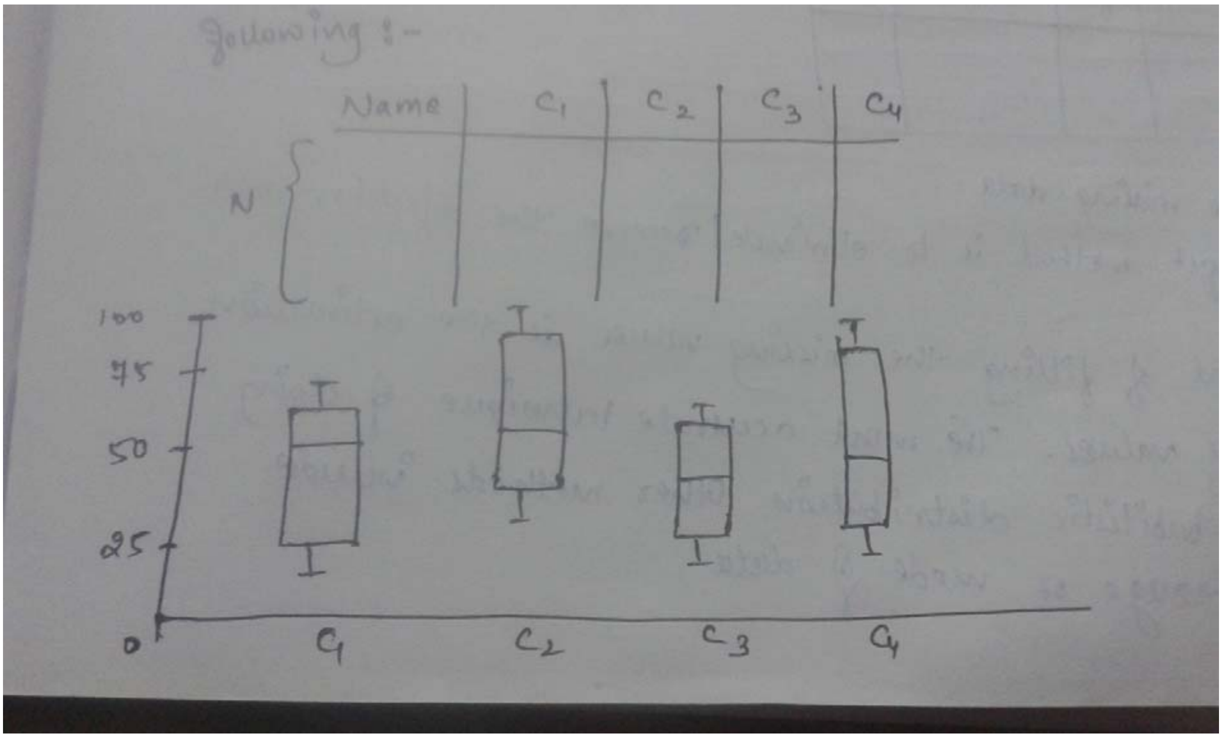
where x' is mean(x).

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

15. 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.

16. Give an example of X such that standard deviation is with a maximum value possible.

17. 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?

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

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

18. The following measurements were recorded for the drying time, in hours, of a certain brand of latex paint

3.4	2.5	4.8	2.9	3.6
2.8	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

Assume that the measurements are a simple random sample.

(a) What is the sample size for the above sample?
(b) Calculate the sample mean for this data.

(c) Calculate the sample median.
(d) Compute the 20% trimmed mean for the above data set.

19. A tire manufacturer wants to determine the inner diameter of a certain grade of tire. Ideally, the diameter would be 570 mm. The data are as follows:

572, 572, 573, 568, 569, 575, 565, 570.

(a) Find the sample mean and median.
(b) Find the sample variance, standard deviation, and range.
(c) Using the calculated statistics in parts (a) and (b), can you comment on the quality of the tires?

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. The concentration of DDT, in milligrams per liter, is:
(a) a nominal variable
(b) an ordinal variable
(c) an interval variable
(d) a ratio variable.

3. 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.

4. The species of each insect found in a plot of cropland is:

(e) a nominal variable
(f) an ordinal variable
(g) an interval variable
(h) a ratio variable.

5. 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

6. 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.

7. Identify which of the following is a measure of dispersion:
(a) median
(b) 90th percentile
(c) interquartile range
(d) mean

8. 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.
(e) None of the above is correct.

9. A measurable characteristic of a population is:

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

10. 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.

11. 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 small sample of automobile owners at IIT Kharagpur produced the following number of parking tickets during a particular year: 4, 0, 3, 2, 5, 1, 2, 1, 0. The mean number of tickets (rounded to the nearest tenth) is:

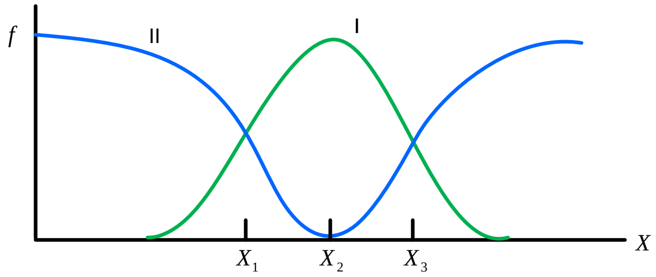
(a) 1.7
(b) 2.0
(c) 2.5

(d) 3.0

15. 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

16. 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.

17. 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