Page No

Date

Statistics - 1 worksheet

By:
- Yatika Taneja

① Bernoulli random ~~values~~ variables take only the values 1 or 0

Sol) True (a)

② Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases

Sol) (a) Central limit theorem

③ Which of the following is incorrect with respect to use of poisson distribution

Sol) (b) Modeling bounded count data

④ Point out the correct statement

Sol) (d) All of the above mentioned

⑤ _______ Random variables are used to model rates

Sol) (c) Poisson distribution

6. Usually replacing the standard error by its estimated value does change the CLT

Sol> (b) False

7. Which of the following testing is concerned with making decisions using data.

Sol> (b) HyPothesis

8. Normalized data are centered at ___ and have units equal to standard deviations of the original data.

Sol> (a) 0

9. Which of the following statement is incorrect with respect to outliers

Sol> (c) Outliers cannot conform to the regression relationship

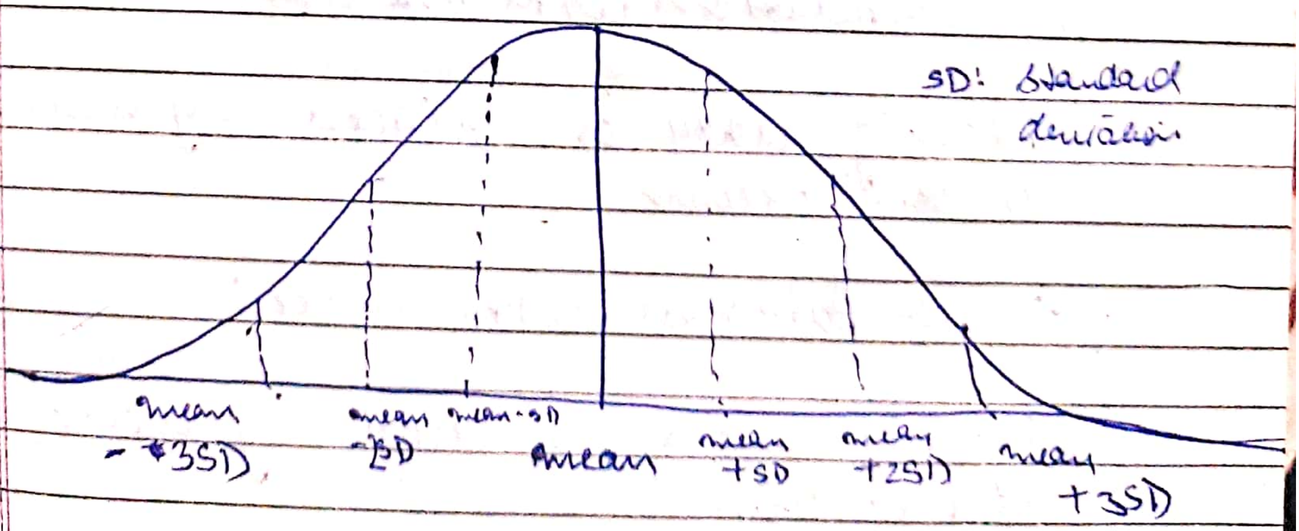
10. What do you understand by the term Normal distribution ?

- Normal distribution is a probability distribution that is symmetric about the mean

- The standard normal distribution has 2 Parameters : mean & standard deviation

- The empirical rule states that :
  (i) 68% of values fall within ~~first~~ 1 standard deviation of the mean
  (ii) 95% of values fall 2 standard deviation of the mean
  (iii) 99.7% of values are within 3 standard deviation of the mean

- In graph form, it appears as bell curve



(11) How do you handle missing data? What imputation ~~data~~ techniques do you recommend

(1) Remove all rows with missing data if there are not too many rows with missing data

(2) If more than 50% of rows of a specific column have missing data it is common to remove the particular column

(3) Imputation Techniques:-

(a) Imputation with mean:-

Missing data is replaced by mean of the column

(b) Imputation with median:

Missing data is replaced by median of the column

(c) Imputation with mode!

Missing data is replaced by mode of the column

(11) Imputation with Linear Regression:

The missing value is replaced by applying linear regression based on other feature values

(12) What is A/B testing?

- It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics

- A/B testing is a way to compare two versions of a variable to find out which perform better in a controlled environment.

(13) Is mean imputation of missing data acceptable practice?

- Mean imputation should only be used for small numerical datasets

- For larger datasets, it is not suitable because:

(i) It does'nt factor the correlation between features. It only works on the column level

(ii) ~~daas~~ Gives poor result on encoded categorical features

(iii) Not accurate and doesnot account for uncertaincity in the imputations

(14) What is linear regression in Statistics

- Linear regression is a linear model i.e a model that assumes linear relationship between input variables (x) and single output variable (y).

- Linear regression can be expressed as:

$$y = B_0 + B_1 * x$$

where B_0 is the intercept & B_1 is the coefficient

- In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane

(15) What are various branches of statistics?

Sol) The 2 branches of statistics are :-

I Descriptive statistics

- Analysis of data that helps describe, show or summarize data in a meaningful way.
- It does not allow us to make conclusions beyond the data we have analyzed
- 2 types of descriptive statistics are :-

→ Measure of central tendency
- mean
- median
- mode

→ Measure of spread.
- range
- quartiles
- variance
- standard deviation

II Inferential statistics

- Inferential statistics are often used to compare the differences between treatment groups.

- Inferential statistics use measurement from sample of subjects in the experiment to compare the treatment groups and make generalizations about the larger population of subjects

Statistics - 2 Worksheet

By:
~Yatika Taneja

(1) What represent a Population Parameter?

Sol) (B) mean

(2) What will be the median of following set of scores (18, 6, 12, 10, 15) ?

Sol) (C) 12

(3) What is standard deviation?

Sol) (D) All of the above

Scanned with CamScanner

(4) The intervals should be — in a grouped frequency distribution

Sol) c) Both of these

(5) what is the goal of descriptive ~~of~~ statistics?

Sol) (B) Summarizing & explaining specific set of data

(6) A set of data organized in a participant by variables format is ~~o~~ called

Sol) (B) Data Set

(7) In multiple regression, dependent variables are used

Sol) (c) 1

(8) which of the following ~~is~~ is used when you want to visually examine the relationship between 2 quantitative variables?

Sol) B) Scatter Plot

9 2 or more group means are compared by using

Sol> D) Analysis of variance

10 __ is a raw score which has been transformed into standard deviation units.

Sol) (a) Z-score

11 __ is the value calculated when you want arithmetic average?

Sol> c) mean

12 find the mean of these set of numbers (4,6,7,9,2000000).

Sol) D) 400005.2

13 _ is a measure of central tendency that takes into account magnitude of scores

Sol> D) Mean

Page No.

Date

(14) __ Is focus on describing or explaining data whereas __ involves going beyond immediate & making inferences

Sol) (A) Descriptive & inferences

(15) What is the formula of range

Sol) (D) H-L