STATISTICS WORKSHEET-1

ANSWERS:

Q1→ a) True

Q2→ a) Central Limit Theorem

Q3→ b) Modelling bounded count data

Q4→ d) All of the mentioned above

Q5→ c) Poisson

Q6→ b) False

Q7→ b) Hypothesis

Q8→ a) 0

Q9→ c) Outliers cannot conform to the regression relationship

Q10→ explanation on normal distribution.

In Normal distribution , the distribution of a data set takes place in such a way that , the majority of the values lies near the centre or at the middle of the range and the intensity of data gradually decreases towards both of the ends to the extreme in a symmetrical manner. Normal distribution gives bell shaped curve when it is represented graphically.

- The total area under the normal curve is equal to one.
- In normal distribution values of mean, median and mode coincides i.e. mean=median=mode.
- Normal distribution has no skewness.

Q11→ Handling missing data

Handling the missing values in a data set is very important step while cleaning the data set. In some cases deletion of row which has the missing values can be done, but this cause the problem if the data set has only few rows.

The most common technique is used to deal with it is imputation in which the substitution of an estimate value takes place and considered as if imputed values are true observed values.

- Mean imputation :- mean of the observed values for that column is calculated and replaced in missing values. As it has certain drawbacks ,most often other techniques are used.

- Hot deck imputation

  A value picked at random from a sample member who has comparable values on other variables. To put it another way, select all the sample participants who are comparable on other factors, then choose one of their missing variable values at random.

- Regression imputation

The result of regressing the missing variable on other factors to get a predicted value. As a result, instead of utilising the mean, you're relying on the anticipated value, which is influenced by other factors. This keeps the associations between the variables in the imputation model, but not the variability around the anticipated values.

- Mode imputation

  This technique says to replace the missing value with the variable with the highest frequency or in simple words replacing the values with the Mode of that column.

Q12→ A/B testing

A/B testing is a process of statistical testing in which you compare two versions such as version A and version B of something to determine which performs more effective and also to understand if a difference between two versions is statistically significant. Its is also called as split testing or bucket testing.

Q13→ It depends on the data set we have or the problem we are dealing with . But in general it is considered as a bad idea because ,

- Mean imputation reduces the variance of the imputed variables.
- Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.
- Mean imputation does not preserve relationships between variables such as correlations.

Q14→ Linear Regression in statistics

In statistics linear regression shows the relationship between two variables by applying linear equation to observe the data and calculating a straight line that demonstrates relationship between two variables.

$y = a + bx$, where , x is input variable so called predictor variable , y is the output variable so called predicted variable and a is the intercept and b is the coefficient of x

Q15→ Branches of statistics.

Mainly there are two branches of statistics , both involves collection and analysing of data.

Descriptive statistics: if the data is small or can be easily described then it is called descriptive statistics. It involves summarizing the data.

inferential statistics : if the data is two big , here comes the concept of population and sample where small amount of data is randomly selection from the population and analysed and the conclusions of sample data is then applicable to population.

Apart from this there are various branches of statistics which is dependent on the field in which statistics is using , if the data is bio related then it is biostatistics , if the data is computer science related then it is data science ,machine learning etc. if it is economic data then it is economic statistics.

Like this statistics is a interdisciplinary subject which has a coordination with various subjects and their data related problems.