

STATISTICS WORKSHEET - 1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

Q.1 Bernoulli random variables take (only) the values 1 and 0.

- A) True B) False

Ans. A) True

Q.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 ?

- A) Central Limit Theorem
B) Central Mean Theorem
C) Central Limit Theorem
D) All of the mentioned

Ans. A) Central Limit Theorem

Q.3 Which of the following is incorrect with respect to use of Poisson distribution ?

- A) Modeling event/time data
B) Modeling bounded count data
C) Modelling contingency tables
D) All of the mentioned

Ans. B) Modeling bounded count data

Q.4 Point out the correct statement.

- A) The exponent of normally distributed random variables follows what is called the log-normal distribution
B) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
C) The square of a standard normal random variable follows what is called chi-squared distribution
D) All of the mentioned

Ans. D) All of the mentioned

Q.5 ____ random variables are used to model rates.

- A) Empirical B) Binomial C) Poisson D) All of the mentioned

Ans. C) Poisson

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

- A) True B) False

Ans. B) False

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

- A) Probability B) Hypothesis C) Causal D) None of the mentioned

Ans. B) Hypothesis

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

- A) 0 B) 5 C) 1 D) 10

Ans. A) 0

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

- A) Outliers can have varying degrees of influence
B) Outliers can be the result of spurious or real processes
C) Outliers cannot conform to the regression relationship
D) None of the mentioned

Ans. C) Outliers cannot conform to the regression relationship

Q10 and Q15 are subjective answer type questions. Answer them in your own words briefly.

Q.10 What do you understand by the term Normal Distribution ?

Ans. Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve. In a normal distribution the mean is 0 and the standard deviation is 1. It has zero skew and the normal distribution is the most important probability distribution in statistics because it fits many natural phenomena. For example - Heights, blood pressure,

measurement error, and IQ scores follow the normal distribution. The normal distribution model is motivated by the Central Limit Theorem.

Q.11 How do you handle missing data ? What imputation techniques do you recommend ?

Ans. Data can be missed due to the following ways that are shown below :

- Missing Completely At Random (MCAR) : When missing values are randomly distributed across all observations, then we consider the data to be missing completely at random.
- Missing At Random (MAR) : The key difference between MCAR and MAR is that under MAR the data is not missing randomly across all observations, but is missing randomly only within sub-samples of data.
- No Missing At Random (NMAR) : When the missing data has a structure to it, we cannot treat it as missing at random.

Missing data is a huge problem for data analysis because it distorts findings. It's difficult to be fully confident in the insights when you know that some entries are missing values. Hence, why they must be addressed. According to data scientists, there are three types of missing data that I have explained above.

Imputation Techniques :

1. Mean or Median Imputation
2. Multivariate Imputation by Chained Equations (MICE)
3. Random Forest

You could find missing/corrupted data in a dataset and either drop those rows or columns, or decide to replace them with another value. In Pandas, there are two very useful methods: `isnull ()` and `dropna()` that will help you find columns of data with missing or corrupted data and drop those values. If you want to fill the invalid values with a placeholder value (for example, 0), you could use the `fillna()` method.

Q.12 What is A/B testing ?

Ans. A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element,

etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drive business metrics. An AB test is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

Q.13 Is mean imputation of missing data acceptable practice ?

Ans. True, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased and that's a good thing. Further it is non-standard and it uses Random Forest. It is use to predict the missing data. It also can be used for both i.e. continuous as well as categorical data and so it makes advantageous over other imputations.

There are some limitations too :-

1. Mean imputation does not preserve the relationship among variables. It preserves the mean of observed data. If data is missing completely at random, the estimate of the mean remains unbiased.
2. Mean Imputation leads to an underestimate of standard errors.

Q.14 What is linear regression in statistics ?

Ans. Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model. A linear regression line has an equation of the form $Y = mx + c$, where X is the explanatory variable and Y is the dependent variable. The slope of the line is m, and c is the intercept (the value of y when x=0).

Types of linear regression :

1. Simple linear regression
2. Multiple linear regression
3. Logistic regression
4. Ordinal regression

5. Multinomial regression

Q.15 What are the various branches of statistics ?

Ans. Various branches of statistics are explained as follows :

1. Descriptive Methods :- This type of method consists of all the preliminary steps to final analysis and interpretation. As such this method includes the method of collection, methods of tabulation, measures of central tendency, measures of dispersion, measures of skewness, and analysis of time series. These methods bring out the various characteristics of data and help in summarizing and interpreting the salient features of the data. This method is also otherwise called Descriptive Statistics.
 2. Analytical Methods :- This type of method consists of all those methods which help in the matter of analysis and comparison between any two or more variables. This includes the methods of correlation, regression analysis, association of attributes and the like. This method is also otherwise called Analytical Statistics.
 3. Inductive Methods :- This type of method consists of all those procedures that help in the generalization or estimation over a phenomenon on the basis of random observation or partial data. This includes the procedure of interpolation, extrapolation, theory of probability and the like. This method is also otherwise called Inductive Statistics.
 4. Inferential Methods :- This type of method consists of those procedures which help in drawing inferences about the characteristics of the population on the basis of samples. As such, this method includes the theory of sampling, different tests of significance, statistical control etc. This method is also otherwise called Inferential Statistics.
 5. Applied Methods :- This type of method consists of those procedures which are applied to the problems of real life. This includes the method of statistical quality control, sample survey, linear programming and inventory control.
-