STATISTICS WORKSHEET-1

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

1. Bernoulli random variables take (only) the values 1 and 0.
   Answer: a) True

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?
   Answer: a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?
   Answer: b) Modeling bounded count data

4. Point out the correct statement.
   a) The exponent of a 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
   Answer: d) All of the mentioned

5. ______ random variables are used to model rates
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned
   Answer: c) Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False
   Answer: b) False

7. Which of the following testing is concerned with making decisions using data?
   a) Probability
   b) Hypothesis
   c) Causal
   d) None of the mentioned
   Answer: b) Hypothesis

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
Answer: a) 0

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
Answer: c) Outliers cannot conform to the regression relationship

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

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

Answer:

The normal distribution is also referred to as Gaussian or Gauss distribution. The distribution is widely used in natural and social sciences. It is made relevant by the Central Limit Theorem, which states that the averages obtained from independent, identically distributed random variables tend to form normal distributions.

Shape of Normal Distribution

A normal distribution is symmetric from the peak of the curve, where the mean is. This means that most of the observed data is clustered near the mean, while the data become less frequent when farther away from the mean. The resultant graph appears as bell-shaped where the mean, median, and mode are of the same values and appear at the peak of the curve.

Parameters of Normal Distribution

The two main parameters of a (normal) distribution are the mean and standard deviation. The parameters determine the shape and probabilities of the distribution. The shape of the distribution changes as the parameter values change.

All forms of (normal) distribution share the following characteristics:

1. It is symmetric

A normal distribution comes with a perfectly symmetrical shape. This means that the distribution curve can be divided in the middle to produce two equal halves. The symmetric shape occurs when one-half of the observations fall on each side of the curve.

2. The mean, median, and mode are equal

The middle point of a normal distribution is the point with the maximum frequency, which means that it possesses the most observations of the variable. The midpoint is also the point where these three measures fall. The measures are usually equal in a perfectly (normal) distribution.

3. Empirical rule

In normally distributed data, there is a constant proportion of distance lying under the curve between the mean and specific number of standard deviations from the mean. For example, 68.25% of all cases fall within +/- one standard deviation from the mean.
95% of all cases fall within +/- two standard deviations from the mean, while 99% of all cases fall within +/- three standard deviations from the mean.

4. Skewness and kurtosis

Skewness and kurtosis are coefficients that measure how different a distribution is from a normal distribution. Skewness measures the symmetry of a normal distribution while kurtosis measures the thickness of the tail ends relative to the tails of a normal distribution.

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

Answer:

Missing data appear when no value is available in one or more variables of an individual. Due to Missing data, the statistical power of the analysis can reduce, which can impact the validity of the results.

<u>Reason behind missing data:</u>

Missing data can occur due to many reasons. The data is collected from various sources and, while mining the data, there is a chance to lose the data.

<u>Types of missing data:</u>

Missing Completely at Random (MCAR)
Missing at Random (MAR)
Missing Not at Random (MNAR)

Detecting Missing value

Detecting missing values numerically

Detecting missing data visually using Missingno library

Finding relationship among missing data

Using matrix plot

Using a Heatmap

Treating Missing values

1. **Deletions**

   Pairwise Deletion

   List wise Deletion/ Dropping rows

   Dropping complete columns

2. **Basic Imputation Techniques**

   Imputation with a constant value

   Imputation using the statistics (mean, median, mode)

   K-Nearest Neighbour Imputation

12. What is A/B testing?

Answer:
A/B testing (also known as bucket testing or split-run testing) is a user experience research methodology. A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

As highlighted in the AB testing definition, it helps increase profits by improving conversions and allowing the business to reach more people. About 60 percent of businesses believe it helps improve conversion. In addition to this, A/B test results can improve bounce rates and increase engagement.

A/B testing, also known as split testing, is a marketing experiment wherein you split your audience to test a number of variations of a campaign and determine which performs better. In other words, you can show version A of a piece of marketing content to one half of your audience, and version B to another.

Reasons not to run a test

You don't yet have meaningful traffic

You can't safely spend the time.

You don't yet have an informed hypothesis.

There's low risk to taking action right away.

The typical approach for testing a null hypothesis is to select a statistic based on a sample of Fixed size, calculate the value of the statistic for the sample and then reject the null Hypothesis if and only if the statistic falls in the critical region.

A/B testing can take a lot longer to set up than other forms of testing. Setting up the A/B system can be a resource and time hog, although third-party services can help. Depending on the company size, there may be endless meetings about which variables to include in the tests.

A/B testing allows individuals, teams and companies to make careful changes to their user Experiences while collecting data on the results. This allows them to construct hypotheses And to learn why certain elements of their experiences impact user behaviour.

A/B testing is an optimisation technique often used to understand how an altered variable Affects audience or user engagement. It's a common method used in marketing, web design, Product development, and user experience design to improve campaigns and goal Conversion rates.

13. Is mean imputation of missing data acceptable practice?

Answer:

The process of replacing null values in a data collection with the data's mean is known as mean imputation.
Mean imputation is typically considered terrible practice since it ignores feature correlation.
Mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance.

14. What is linear regression in statistics?

Answer:

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. The variable you are using to predict the other variable's value is called the independent variable.

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the
Formula:
y = c + b*x,
Where y = estimated dependent variable score,
c = constant,
b = regression coefficient, and
x = score on the independent variable.

Naming the Variables:
There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand.
The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are:

(1) Determining the strength of predictors.
(2) Forecasting an effect.
(3) Trend forecasting.

15. What are the various branches of statistics?

Answer:
There are majorly two different types of statistics:

1. **Descriptive statistics**

3. **Inferential statistics**

Descriptive Statistics:

Descriptive Statistics is used to summarise the data through a given set of observations and illustrations. The given data set is summarized from a sample of the population using some of the formulae such as computation of the mean or standard deviation of the given observation. For instance, descriptive statistics are used for the data collection of the number of persons in a city that uses either television or the internet.

This type of statistics deals with the organization and representation of data. It also describes the given collection of data in the form of tables, graphs, and summary measures.

Descriptive statistics can also be divided into four different categories, which are as follows:

Measure of frequency: Indicator of the number of times a particular data occurs.

Measure of dispersion: Range, variance, and standard deviation are measures of dispersion of the data.

Measure of central tendency: Mean, median, and mode are the measure of central tendency of the given data distribution.

Measure of position: The measure of position describes the percentile and quartile ranks.

Inferential Statistics:

Inferential statistics is basically used to make interpretations of the meaning held by descriptive statistics. The data collection and analysis are followed by the interpretation of the collected facts which is done by using inferential statistics. The summarised data is subjected to statistical analysis to illustrate its meaning further. To summarise, inferential statistics are used to draw conclusions from the data distribution. For instance, inferential statistics can be used in deriving estimates from hypothetical research.

This type of statistics basically allows the person to make use of the data collected. It simulates in making decisions and predictions or even inferences from a specified population. It grants us permission to give statements that goes beyond the available data or information.

Inferential Statistics is dependent on the following random variations:

Observational errors
Sampling variation