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.

a) True

b) False

Ans. 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?

a) Central Limit Theorem

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

Ans.A

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) Modeling contingency tables

d) All of the mentioned

Ans.C

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 distribution

d) All of the mentioned

Ans.d

5. ______ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Ans.C

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

a) True

b) False

Ans.B

7. 1. 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

8. 4. 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

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 mentione

Ans.C

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

Ans.

The normal distribution is the most known and used distribution if we compare it to all other distributions. Because it approximates many natural phenomena. It has developed ino a standard of reference for many problems.

Normal distributions have key characteristics that are easy to spot in graphs.
Such as:- The mean, median and mode are exactly the same.

The distribution is symmetric about the mean—half the values fall below the mean and half above the mean. The distribution can be described by two values: the mean and the standard deviation.

Many things are actually normally distributed or very nearer to it.

Eg. Height and intelligence are approximately normally distributed , measurements error also often have the normal distributions.

A normal distribution is one in which the values are evenly distributed both above and below the mean. A population has a precisely normal distribution if the mean, mode, and median are all equal.

It is easy to work with mathematically.

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

Ans.

Missing data can be handle with in a variety of ways. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programme will make the decision for us.

Our application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone, listwise deletion may or may not be a good idea.

Another common strategy among those who pay attention while data is missing is imputation. Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values.

There are many imputation technique is available for us. Among all I recommend to use MEAN & MEDIAN imputation.

Mean imputation is a method in which the mean of the observed values for each variable is computed and the missing values for that variable are imputed by this mean.

“Mean” will replace missing values using the mean in each column. It is preferred if data is numeric and not skewed.

As well as

“Median” will replace missing values using the median in each column. It is preferred if data is numeric and skewed.

12. What is A/B testing?

Ans.

In simple words A/B testing is a method of comparing two versions of a webpage or app against each other to determine which one performs better.

A/B testing is basically statistical hypothesis testing, or we can say statistical inference. It is an analytical method for making decisions that estimates population parameters based on sample statistics.

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.

A/B Testing is a widely used concept in most industries nowadays.

13. Is mean imputation of missing data acceptable practice?

Ans.

It is actually both acceptable and nonacceptable practice.its completely depend upon the situetions.

Because it 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.

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, the model is less accurate and the confidence interval is narrower.

Mean imputation does not preserve the relationships among variables.

Outliers data points will have a significant impact on the mean,Hence in such cases, it is not recommended to use the mean for replacing the missing values.

So we accept it as good and bad also according to the case or situetions.

14. What is linear regression in statistics?

Ans.

In Simple words linear regression is a regression that estimates the relationship between one independent variable and one dependent variable using a straight line.

Linear regression shows the relationship between one or more predictor variable(s) and one outcome variable.

It is commonly used for predictive analysis and modeling.

It helps to predict the effects of the independent variable on the dependent one

Lets take an Example.

We can say that age and height can be described using a linear regression model. As we know a person's height increases as its age increases, So we can call a linear relationship between age & height

15. What are the various branches of statistics?

Ans.

Statistics is a method of interpreting, analysing and summarising the data.

Statistics have majorly categorised into two types,such as:-

(i). Descriptive statistics

(ii). Inferential statistics

(i). Descriptive Statistics:-

Descriptive statistics deals with the collection of data, its presentation in various forms, such as tables, graphs and diagrams and finding averages and other measures which would describe the data.

In this type of statistics, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean and standard deviation.

It is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorised into four different categories.

The frequency measurement displays the number of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data. Central tendencies are the mean, median and mode. And the measure of position describes the percentile and quartile ranks.

For example:

Industrial statistics, population statistics, trade statistics, etc.

Businessmen make use of descriptive statistics in presenting their annual reports, final accounts, and bank statements.

(ii). Inferential Statistics:-

In simple words it deals with techniques used for the analysis of data, making estimates and drawing conclusions from limited information obtained through sampling and testing the reliability of the estimates.

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that goes beyond the available data or information.

For example:

Suppose we want to have an idea about the percentage of the illiterate population of our country. We take a sample from the population and find the proportion of illiterate individuals in the sample. With the help of probability, this sample proportion enables us to make some inferences about the population proportion.

This study belongs to inferential statistics.