STATISTICS WORKSHEET-1

Response sheet

1) A.

2) A.

3) B.

4) C.

5) C.

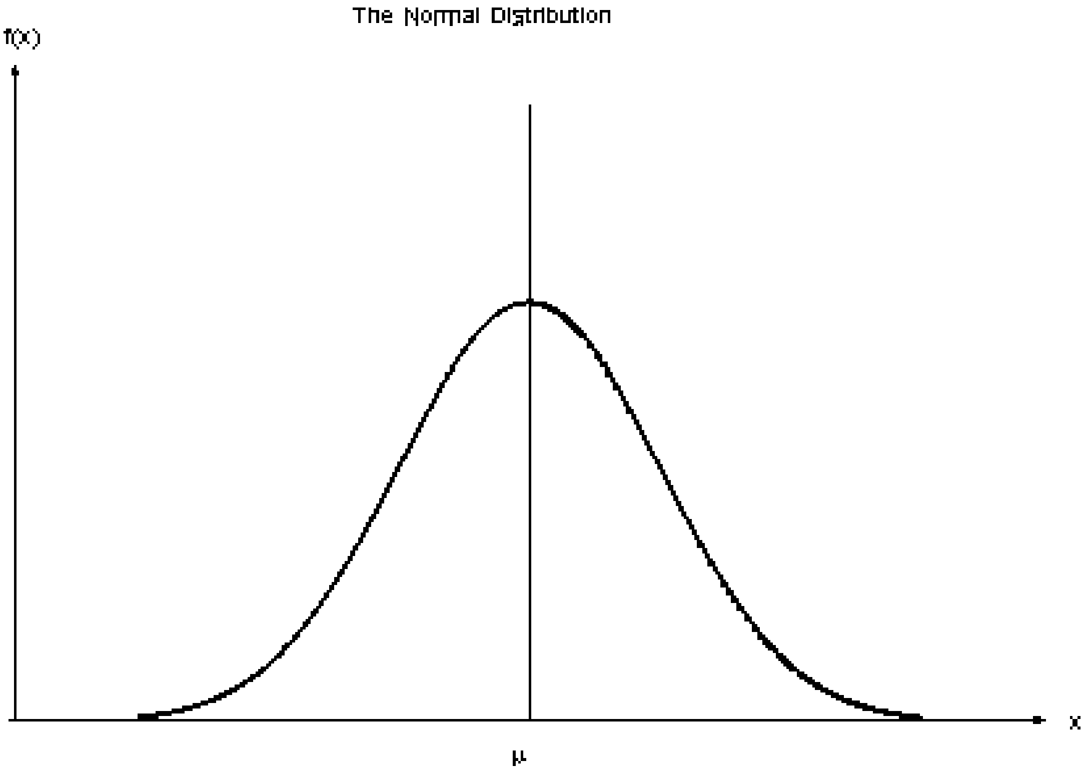
6) A.

7) B.

8) A.

9) ----

10) A common random continuous distribution in statistics is the Normal distribution. In shape it is bell shaped so it is also called bell-curve distribution. It is a continuous probability distribution relating to the mean and standard deviation.



Probability distribution function for a normal random variable x

Probability Distribution for a Normal Random Variable x

Probability density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

μ = Population mean of the normal random variable x

σ = Population standard deviation

P(x<a) is obtained from a table of normal probabilities.

The normal distribution plays an important role in inferential statistics.

Noraml distribution with μ =0 and σ =1 is called Standard Noraml Distribution. It is denoted by z.

11) When the data is aggregated over long time stretches from disparate sources then its the common problem of missing data. The reliable machine learning model demands for careful handling of missing data.

There are three missing value imputation techniques-

MEAN

MEDIAN

MODE

One of the main problem is to decide which technique to use to get the most effective value for the missing value.

The goal is to find out which is a better measure of the central tendency of data and us that value for replacing missing values appropriately.

The common technique to decide which technique is to use is plot such as **box plot and distribution plot.**

If the data distribution is symmetric then we can use **mean value for** imputing missing values.

But if the data distribution is skewed then we have to choose either **MODE or MEDIAN**.

If data distribution is skewed and data is neumerical then we can choose either **MODE or MEDIAN**.

But if the data distribution is skewed and the data is neumerical or categorical the it is the best practise to use **MODE** value or most frequent value will replace the missing value of the entire feature column.

12)

13) Mean imputation is so simple technique to use and it’s a popular solution to missing data,despite its drawbacks,mainly because it’s easy.

But that doesn’t make it a good solution,and it may not help you find strong relationship with strong parameter estimates .

The main problem with using mean imputation is-

A) Imputing the mean preserves the mean of the observed data. So if he data are missing completely at random, the estimate of the mean remains unbiased.But it will biased the standard error. The values it will put in place of missing values does affect the correlation.

So mean imputation is good at preserving unbiased estimates for the mean isn’t so good for unbiased estimates of relationships.

B) You get the same mean from mean imputed data that you would have gotten without the imputations. And there are circumstances where that mean is unbiased. Even so, the standard errror of that mean will be too small.

Because the imputations are themselves estimates, there is some error associated with them. But your statistical software doesn’t know that. It treats it as real data.

Ultimately standard errors are too low, so as p values. It will cause Type 1 errors without realizing it .

So at the end we can say that mean imputation of missing data is not acceptable

14) Linear regression is a statistical modeling technique used to describe a continuous response variable as a function of one or more predictor variables. It can help you understand and predict the behavior of complex systems or analyze experimental, financial, and biological data.

Linear regression techniques are used to create a linear model. The model describes the relationship between a dependent variable y (also called the response) as a function of one or more independent variables x (called the predictors). The general equation for a linear regression model is:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i$$

Types of Linear Regression --

Simple Linear Regression- Models using only one predictor.

Multiple Linear Regression- Models using multiple predictors.

Multivariate Linear Regression- Models using multiple response variable.

Multivariate Multiple Linear Regression- Models using multiple predictors for multiple response varaibles.

Applications of linear regression

Linear regressions have some properties that make them very interesting for the following applications :

1) Prediction or forecasting – Use a regression model to build a forecast model for a specific data set. From the model, you can use regression to predict response values where only the predictors are known.

2) Strength of the regression – Use a regression model to determine if there is a relationship between a variable and a predictor, and how strong this relationship is.

15) The **statistics** Is a branch of mathematics, which corresponds to the collection, analysis, interpretation, presentation and organization of data (set of values of qualitative or quantitative variable). This discipline seeks to explain the relationships and dependencies of a phenomenon (physical or natural).

The statistic is divided into two major areas: ***Descriptive statistics*** **And** ***Inferential statistics***

1- Descriptive Statistics --

The **Descriptive statistics** Is the branch of statistics that describes or summarizes quantitatively (characteristics) a collection of a collection of information.

That is, descriptive statistics is responsible for summarizing a statistical sample (set of data obtained from a population) Rather than learning about population Which represents the sample.

Some of the measures commonly used in descriptive statistics to describe a set of data are the measures of central tendency and the Measures of variability or dispersion .

As for measures of central tendency, measures such as half , the median and the fashion . While in the measures of variability the Variance , the Kurtosis , etc.

2- Inferential Statistics --

The **Inferential statistics** Is distinguished from descriptive statistics mainly by the use of inference and induction.

That is, this branch of statistics seeks to deduce properties from a population Studied, that is, not only collects and summarizes the data, but also seeks to explain certain properties or characteristics from the data obtained.

In this sense, inferential statistics implies obtaining the correct conclusions from a statistical analysis performed using descriptive statistics.

For this reason, many of the experiments in social sciences involve a group of population As well as inferences and generalizations can be determined as the population In general behaves.

Inferential statistics are divided into:

Parametric Statistics

It comprises the statistical procedures based on the distribution of the actual data, which are determined by a finite number of parameters (number that summarizes the amount of data derived from a statistical variable).

Non-parametric statistics

This branch of inferential statistics comprises procedures applied in tests and statistical models in which their distribution does not fit the so-called parametric criteria. As the data studied are those that define its distribution, it can not be defined previously.

3-Mathematical Statistics --

This consists of a previous scale in the study of statistics, in which they use the theory of probability (branch of mathematics that studies the Random phenomena) And other branches of mathematics.

Mathematical statistics consists of obtaining information from the data and uses mathematical techniques such as: Mathematical analysis, linear algebra, stochastic analysis, differential equations, etc. Thus, mathematical statistics has been influenced by applied statistics.