STATISTICS WORKSHEET-1

Objective Questions

Q1.

Ans: a) True

Q2.

Ans: a) Central Limit Theorem

Q3.

Ans: b) Modelling bounded count data

Q4.

Ans: d) All of the mentioned

Q5.

Ans: b) Binomial

Q6.

Ans: b) False

Q7.

Ans: b) Hypothesis

Q8.

Ans: a) 0

Q9.

Ans: b) Outliers can be the result of spurious or real processes

Subjective Questions:

Q10.

Ans: Normal Distribution:

The Normal Distribution is defined by the probability density function for a continuous random variable in a system. Let us say, f(x) is the probability density function and X is the random variable. Hence, it defines a function which is integrated between the range or interval (x to x + dx), giving the probability of random variable X, by considering the values between x and x+dx.

$f(x) \geq 0 \; \forall \; x \in (-\infty, +\infty)$

And $-\infty \int +\infty \; f(x) = 1$

Normal Distribution Formula

The probability density function of normal or gaussian distribution is given by;

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

Where,

- x is the variable
- μ is the mean
- σ is the standard deviation

Q11.

Ans: Missing data and Imputation technique:

Missing Data can occur when no information is provided for one or more items or for a whole unit. Missing Data is a very big problem in real life scenario. Missing Data can also refer to as NA(Not Available) values in pandas. In DataFrame sometimes many datasets simply arrive with missing data, either because it exists and was not collected or it never existed.

For Example, suppose different user being surveyed may choose not to share their income, some user may choose not to share the address in this way many datasets went missing.

When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. However, there can be multiple reasons why this may not be the most feasible option:

- There may not be enough observations with non-missing data to produce a reliable analysis.
- In predictive analytics, missing data can prevent the predictions for those observations which have missing data
- External factors may require specific observations to be part of the analysis

In such cases, we impute values for missing data. A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations. So rather we can say depending upon the nature of the missing data, we use different techniques to impute data.

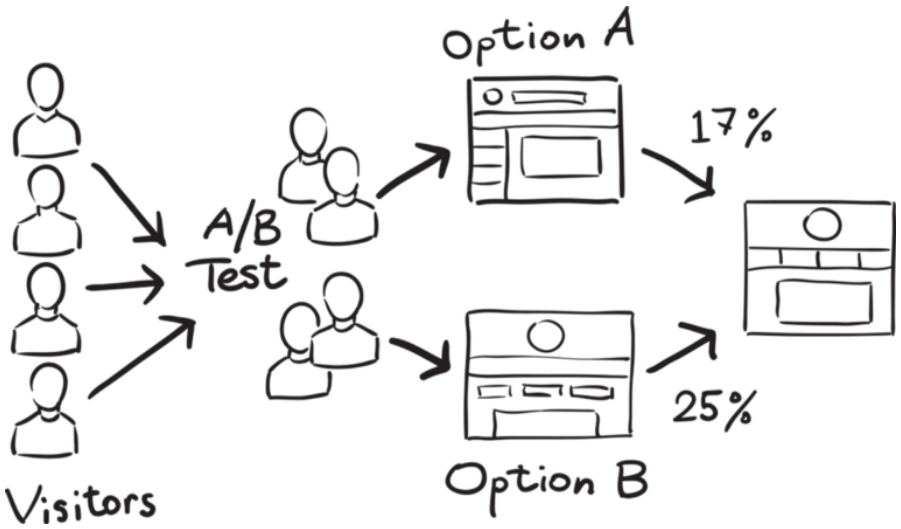
Q12.

Ans: A/B Testing:

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.



It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

Q13.

Ans: Mean imputation can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data.

Q14.

Ans: Linear Regression:

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome 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

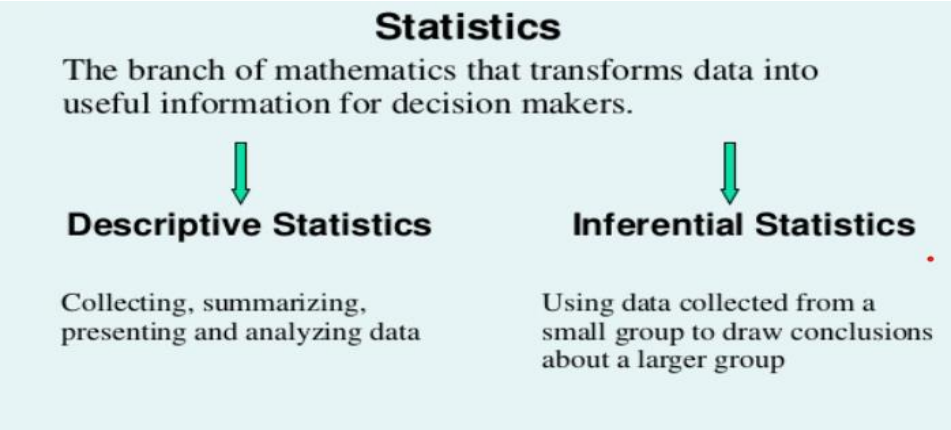
x = score on the independent variable.

Q15.

Ans:

Statistics is concerned with developing and studying different methods for collecting, analysing and presenting the empirical data.

The field of statistics is composed of two broad categories- Descriptive and inferential statistics. Both of them give us different insights about the data. One alone doesn't not help us much to understand the complete picture of our data but using both of them together gives us a powerful tool for description and prediction.



Descriptive Statistics:

It describes the important characteristics/ properties of the data using the measures the central tendency like mean/ median/mode and the measures of dispersion like range, standard deviation, variance etc.

Data can be summarized and represented in an accurate way using charts, tables and graphs.

Inferential Statistics:

It is about using data from sample and then making inferences about the larger population from which the sample is drawn. The goal of the inferential statistics is to draw conclusions from a sample and generalize them to the population. It determines the probability of the characteristics of the sample using probability theory. The most common methodologies used are hypothesis tests, Analysis of variance etc.

End of Document