STATISTICS WORKSHEET

ANS.1- A(True)

ANS.2-A(Central limit theorm)

ANS.3-B(Modeling bounded count data)

ANS.4-D(All of the mentioned)

ANS.5-C(Poisson)

ANS.6-B(False)

ANS.7-B(Hypothesis)

ANS.8-A(0)

ANS.9-C(Cannot conform to regression relationship)

ANS.10-NORMAL DISTRIBUTION:A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

Height is one simple example of something that follows a normal distribution pattern: Most people are of average height, the numbers of people that are taller and shorter than average are fairly equal and a very small (and still roughly equivalent) number of people are either extremely tall or extremely short.

A graphical representation of a normal distribution is sometimes called a bell curve because of its flared shape. The precise shape can vary according to the distribution of the population but the peak is always in the middle and the curve is always symmetrical. In a normal distribution, the MEAN, MEDIAN, MODE are all the same.

ANS11.- **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.

Best techniques to handle missing data

Use deletion methods to eliminate missing data

The deletion methods only work for certain datasets where participants have missing fields. There are several deleting methods – two common ones include Listwise Deletion and Pairwise Deletion. It means deleting any participants or data entries with missing values. This method is particularly advantageous to samples where there is a large volume of data because values can be deleted without significantly distorting readings. Alternatively, data scientists can fill out the missing values by contacting the participants in question. The problem with this method is that it may not be practical for large datasets. Furthermore, some corporations obtain their information from third-party sources, which only makes it unlikely that organisations can fill out the gaps manually. Pairwise deletion is the process of eliminating information when a particular data point, vital for testing, is missing. Pairwise deletion saves more data compared to likewise deletion because the former only deletes entries where variables were necessary for testing, while the latter deletes entire entries if any data is missing, regardless of its importance.

Use regression analysis to systematically eliminate data

Regression is useful for handling missing data because it can be used to predict the null value using other information from the dataset. There are several methods of regression analysis, like Stochastic regression. Regression methods can be successful in finding the missing data, but this largely depends on how well connected the remaining data is. Of course, the one drawback with regression analysis is that it requires significant computing power, which could be a problem if data scientists are dealing with a large dataset.

Data scientists can use data imputation techniques

Data scientists use two data imputation techniques to handle missing data: Average imputation and common-point imputation. Average imputation uses the average value of the responses from other data entries to fill out missing values. However, a word of caution when using this method – it can artificially reduce the variability of the dataset. Common-point imputation, on the other hand, is when the data scientists utilise the middle point or the most commonly chosen value. For example, on a five-point scale, the substitute value will be 3. Something to keep in mind when utilising this method is the three types of middle values: mean, median and mode, which is valid for numerical data (it should be noted that for non-numerical data only the median and mean are relevant).

Keeping things under control

Missing data is a sad fact of life when it comes to data analytics. We cannot avoid situations like these entirely because there are several remedial steps data scientists need to take to make sure it doesn’t adversely affect the analytics process. While these methods are helpful, they are not foolproof because they are contentious, meaning, their effectiveness depends heavily on circumstances. The best option available to data scientists is to work with powerful, processing tools that can make the data capturing and analysis process significantly easier.

ANS.12-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.

ANS14.- LINEAR REGRESSION: IN STATISTICS, LINEAR REGRESSION IS A LINEAR APPROACH FOR MODELLING THE RELATIONSHIP BETWEEN A SCALAR RESPONSE AND ONE OR MORE EXPLANATORY VARIABLES(ALSO KNOWN AS DEPENDENT AND INDEPENDENT VARIABLES. THE CASE OF ONE EXPLANATORY VARIABLE IS CALLED SIMPLE LINEAR REGRESSION; FOR MORE THAN ONE, THE PROCESS IS CALLED MULTIPLE LINEAR REGRESSION. THIS TERM IS DISTINCT FROM MULTIVARIATE LINEAR REGRESSION,WHERE MULTIPLE CORRELATED DEPENDENT VARIABLES ARE PREDICTED, RATHER THAN A SINGLE SCALAR VARIABLE.

ANS.15- BRANCHES OF STATISTICS ARE:
DESCREPTIVE AND INFERENTIAL STATISTICS

DESCRIPTIVE STATISTICS:deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biasis that are so easy to creep into thE Expriment.

Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

INFERENTIAL STATISTICS:
Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.

While drawing conclusions, one needs to be very careful so as not to draw the wrong or biased conclusions. Even though this appears like a science, there are ways in which one can manipulate studies and results through various means. For example, data dredging is increasingly becoming a problem as computers hold loads of information and it is easy, either intentionally or unintentionally, to use the wrong inferential methods.

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good **scientific methodology** needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.