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.

Ans: 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?

Ans: a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

Ans: b) Modeling bounded count data

4. Point out the correct statement.

Ans: d) All of the mentioned

5. _____ random variables are used to model rates.

Ans: c) Poisson

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

Ans: b) False

7. Which of the following testing is concerned with making decisions using data?

Ans: b) Hypothesis

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.

Ans: a) 0

9. Which of the following statement is incorrect with respect to outliers?

Ans: c) Outliers cannot conform to the regression relationship

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Ans: Normal distribution is also known as "Gaussian distribution". It is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. It is most widely known & used of all distributions. In a graph, we can notice that normal distribution will appear as a bell curve shape.

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

Ans: Missing data is an inevitable part of the process. It takes lot of resources, time and energy into making sure the data set is as accurate as possible for data scientists. Sometimes, data sets come up short, no matter how many times data scientists clean and prepare it. The best way to handle such situations is to develop contingency plans to minimize the damage.

Missing data is a huge problem for data analysis because it distorts findings. It's difficult to be fully confident in the insights when you know that some entries are missing values. According to data scientists, there are three types of missing data. These are Missing Completely at Random (MCAR) – when data is completely missing at random across the dataset with no discernable pattern. There is also Missing At Random (MAR) – when data is not missing randomly, but only within sub-samples of data. Finally, there is Not Missing at Random (NMAR), when there is a noticeable trend in the way data is missing.

There are different techniques to handle missing data like deletion methods to eliminate missing data, regression analysis to systematically eliminate data, data imputation techniques or by Keeping things under control

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 utilize 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 utilizing 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).

12. What is A/B testing?

Ans: A/B testing is a type of split testing and is commonly used to drive improvements to specific variables or elements by measuring user or audience engagement. The approach is commonly used to optimize marketing campaigns or digital assets like websites. In A/B testing a specific variable is altered such as a title, image, or element layout.

A sample of the audience is shown the control version and the altered version in a 50/50 split. Half traffic will interact with the original version, the other half will interact with the newer version. Engagement or the completion of a defined goal is the metric that is compared between the versions after a set period of time. A/B testing is performed in tandem for a specific period of time, and the audience sample for both variations is randomized.

An example would be an A/V test for a marketing campaign. An element such as banner size, button color, header image or title text is altered. A control or original version and the newly altered version are then deployed in tandem to an audience sample over a set period of time.

The resulting engagement metrics identify which version the user prefers, so the campaign can be refined and improved for best effect. The preferred version is then implemented, with the aim of

improving engagement with future users. The metric is usually the conversion rate of a defined business goal, such as the portion of users that click on a button or make a purchase.

A/B testing is an important method of optimization of products or digital assets like websites. Web designers may use A/B testing to refine web page layouts, as two versions of a webpage can be tested against user engagement. Experiments are regularly performed so that improvements can be continuously implemented. Because the users are randomized, A/B testing can be used to evaluate decisions to avoid bias or assumptions. The deciding factor is user preference in a dynamic environment. A/B testing works best with audience segmentation, so that user insights can be refined even more.

A/B testing can be used to:

- Refine marketing campaign messaging and design.
- Improve conversion rates through enhancements to user experience.
- Continuously optimize assets like web pages by considering user engagement.

13. Is mean imputation of missing data acceptable practice?

Ans: Yes, according to me that's true, 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.

Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

14. What is linear regression in statistics?

Ans: Linear regression is a basic and commonly used type of predictive analysis. It examines two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? and

(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, and x = score on the independent variable.
Three major uses for regression analysis are:

(1) Determining the strength of predictors: The regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, or age and income.

(2) Forecasting an effect: it can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or

more independent variables. A typical question is, "how much additional sales income do I get for each additional $1000 spent on marketing?"

(3) Trend forecasting: Third, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. A typical question is, "what will the price of gold be in 6 months?"

Types of Linear Regression:

- Simple linear regression: 1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)
- Multiple linear regression: 1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)
- Logistic regression: 1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)
- Ordinal regression: 1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)
- Multinomial regression: 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)
- Discriminant analysis: 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

15. What are the various branches of statistics?

Ans: There are three branches of statistics: Data collection, Descriptive statistics and Inferential statistics. But the two main branches of statistics are Descriptive statistics and Inferential statistics. Both of these are employed in scientific analysis of data and both are equally important.

1)Descriptive Statistics:

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. The statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment. 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.

2)Inferential Statistics:

Inferential statistics 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.