STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

Q-1 Bernoulli random variables take (only) the values 1 and 0.

A-1 True.

Q-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-2 Central Limit Theorem (CLT)

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

A-3 Modeling Bounded Count Data

Q-4 Point out the correct statement.

A-4 All of the mentioned.

Q-5 ______ random variables are used to model rates.

A-5 Poisson

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

A-6 False

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

A-7 Hypothesis

Q-8 Normalized data are centered at______and have units equal to standard deviations of the original data.

A-8 0

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

A-9 Outliers cannot conform to the regression relationship.

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

Q-1 What do you understand by the term Normal Distribution?

A-1 The normal distribution is the most widely known and used of all distributions. Because the normal distribution approximates many natural phenomena so well, it has developed into a standard of reference for many probability problems. Many things actually are normally distributed, or very close to it. For example, height and intelligence are approximately normally distributed measurement errors also often have a normal distribution. The normal distribution is easy to work with mathematically. In many practical cases, the methods developed using normal theory work quite well even when the distribution is not normal.
There is a very strong connection between the size of a sample N and the extent to which a sampling distribution approaches the normal form. Many sampling distributions based on large N can be approximated by the normal distribution even though the population distribution itself is definitely not normal.

Q-2 How do you handle missing data? What imputation techniques do you recommend?

A-2 Missing data can be dealt 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 decision for you. Your application will remove things in a list wise sequence most of the time. Depending on why and how much data is gone, list wise deletion may or may not be a good idea.
Another common strategy among those who pay attention 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.

The following are some of the most prevalent methods:

1. Mean imputation

2. Substitution

3. Hot deck imputation

4. Cold deck imputation

5. Regression imputation

6. Stochastic regression imputation

7. Interpolation and extrapolation

8. Single or Multiple Imputation

Q-3 What is A/B testing?

A-3 A/B testing (also known as **split testing** or **bucket 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 essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

Running an A/B test that directly compares a variation against a current experience lets you ask focused questions about changes to your website or app and then collect data about the impact of that change

Testing takes the guesswork out of website optimization and enables data-informed decisions that shift business conversations from "we think" to "we know." By measuring the impact that changes have on your metrics, you can ensure that every change produces positive results.

Q-4 Is mean imputation of missing data acceptable practice?

A-4 The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, 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.

Q-5 What is linear regression in statistics?

A-5 Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.
This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

Q-6 What are the various branches of statistics?

A-6 There are three real branches of statistics: data collection, descriptive statistics and inferential statistics.

1. Data Collection: Data collection is all about how the actual data is collected. For the most part, this needn't concern us too much in terms of the mathematics (we just work with what we are given), but there are significant issues to consider when actually collecting data. If you are collecting data, you need to be careful where you get it from. For example, suppose you want to conduct a poll on who people plan to vote for in an election. You can't realistically ask everyone in the whole country (the population), so you have to choose a representative sample of people.

2. Descriptive statistics: Descriptive statistics is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually (via graphs, charts, etc.) or numerically (via averages and so on). The basic aim of descriptive statistics is to 'present the data' in an understandable way. If you simply write down every piece of data, it means little to someone who sees it; it needs to be summarised. Imagine if, on the TV news, they listed on the screen the votes of every single person interviewed by a polling company; it would just be a huge list of parties, and you couldn't arrive at any meaningful conclusion.

3. Inferential statistics: Inferential statistics is the aspect that deals with making conclusions about the data. This is quite a wide area; essentially you are asking 'What is this data telling us, and what should we do?' for example, a number of cars are driving too fast). Note, though, that this may not be the case; everyone might be driving at a perfectly acceptable speed, and the accidents are down to something other than speed (a blind spot or a pothole, for example). This is inferential statistics: take the data you have and make an 'inference' or 'conclusion' from it. We shall see much more of this later when we discuss things such as hypothesis testing, where we test to see whether the data supports a belief that we have.