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.

a) True

b) False

Answer- 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?

a) Central Limit Theorem

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

Answer- a) Central Limit Theorem

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

a) Modelling event/time data

b) Modelling bounded count data

c) Modelling contingency tables

d) All of the mentioned

Answer- b) Modelling bounded count data

4. Point out the correct statement.

a) The exponent of a normally distributed random variables follows what is called the log- normal distribution

b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

c) The square of a standard normal random variable follows what is called chi-squared distribution

d) All of the mentioned

Answer- d) All of the mentioned

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer- c) Poisson

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

a) True

b) False

Answer- b) False

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

a) Probability

b) Hypothesis

c) Causal

d) None of the mentioned

Answer-b) Hypothesis

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

a) 0

b) 5

c) 1

d) 10

Answer- a) 0

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

a) Outliers can have varying degrees of influence

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

c) Outliers cannot conform to the regression relationship

d) None of the mentioned

Answer- 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?

Answer- A normal distribution resembles an asymmetric arrangement of most of the values around the mean, such that the curve so formed looks like a bell. It has two key parameters: the mean (μ) and the standard

deviation (σ). This probability method plays a crucial role in asset return calculation and risk management strategy decisions.

Normal distribution, also known as the Gaussian distribution.

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

Answer- Imputation is the process of replacing missing values with substituted data. It is done as a pre-processing step. There are many techniques for Imputation.

I recommend K-Nearest Neighbors and Trees as good imputation techniques.

K-Nearest Neighbors

When the training set is small or moderate in size, KK-nearest neighbors can be a quick and effective method for imputing missing values. This procedure identifies a sample with one or more missing values. Then it identifies the KK most similar samples in the training data .
Often, however, data sets contain both numeric and categorical predictors. This metric uses a separate specialized distance metric for both the qualitative and quantitative predictors.
Once the KK neighbors are found, their values are used to impute the missing data. The mode is used to impute qualitative predictors and the average or median is used to impute quantitative predictors. KK can be a tunable parameter, but values around 5–10 are a sensible default.

Trees

Tree-based models are a reasonable choice for an imputation technique since a tree can be constructed in the presence of other missing data. In addition, trees generally have good accuracy and will not extrapolate values beyond the bounds of the training data. While a single tree could be used as an imputation technique, it is known to produce results that have low bias but high variance. Ensembles of trees, however, provide a low-variance alternative.

12. What is A/B testing?
Answer- A/B testing is, a scientific approach of experimentation when one or more content factors in digital communication (web, email, social, etc.) is changed deliberately, in order to observe the effects and outcomes for a predetermined period of time. Results are then analysed, reviewed, and interpreted to make a final decision with the highest yielding results.

A/B testing allows any organization to be more data-driven and strategic about their digital communications. It removes the guesswork from decision making and lets the data decide the path forward. Instead of spending valuable meeting time debating what colour the button should be? split testing helps facilitate the conversation to focus more on the data, rather than opinion or emotion

13. Is mean imputation of missing data acceptable practice?

Answer- Mean imputation is the practice of replacing null values in a data set with the mean of the data. Mean imputation is generally bad practice because it does not take into account feature correlation

The **drawbacks of mean imputation**:

1. Mean substitution leads to **bias in multivariate estimates** such as correlation or regression coefficients. Values that are imputed by a variable's mean have, in general, a correlation of zero with other variables. Relationships between variables are therefore biased toward zero.
2. **Standard errors and variance** of imputed variables are biased. For instance, let's assume that we would like to calculate the standard error of a mean estimation of an imputed variable. Since all

imputed values are exactly the mean of our variable, we would be too sure about the correctness of our mean estimate. In other words, confidence interval around the point estimation of our mean would be too narrow.

3. Even the **sample mean of your variable is biased**. Assume that you want to estimate the mean of a population's income and people with high income are less likely to respond; Your estimate of the mean income would be biased downwards.

14. What is linear regression in statistics?

Answer- 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, and x = score on the independent variable.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

15. What are the various branches of statistics?

Answer- 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 for the 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 biases that are so easy to creep into the experiment.

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.