Statistics Worksheet-1

Q1. Bernoulli random variables take (only) the values 1 and 0.

Answer: a) True

Q2. 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?

Answer: a) Central Limit Theorem

Q3. Which of the following is incorrect with respect to use of Poisson distribution?

Answer: b) Modeling bounded count data

Q4. Point out the correct statement.

Answer: d) All of the mentioned

Q5. _____ random variables are used to model rates.

Answer: c) Poisson

Q6.Usually replacing the standard error by its estimated value does change the CLT.

Answer: b) False

Q7. Which of the following testing is concerned with making decisions using data?

Answer: b) Hypothesis

Q8. Normalized data are centred at _____ and have units equal to standard deviations of the original data.

Answer: a) 0

Q9. Which of the following statement is incorrect with respect to outliers?

Answer: c) Outliers cannot conform to the regression relationship

Q10. What do you understand by the term Normal Distribution?

Answer: The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables. It has a bell shaped curve. The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely. The normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it accurately describes the distribution of values for many natural phenomena. Characteristics that are the sum of many independent processes frequently follow normal distributions. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution.

Q11. *How do you handle missing data? What imputation techniques do you recommend?*

Answer: I will use average imputation technique. 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.

Q12. *What is A/B testing?*

Answer: 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.

Q13. *Is mean imputation of missing data acceptable practice?*

Answer: 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.

Q14. *What is linear regression in statistics?*

Answer: Linear regression is an attempt to model the relationship between two variables by fitting a linear equation to observed data, where one variable is considered to be an explanatory variable and the other as a dependent variable.

Q15. *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 student of statistics.

Descriptive Statistics: It 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: It 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.