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 value 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) Modeling event/time data B) Modeling bounded count data
   C) Modeling contingency tables D) All of the mentioned

Answer: (B) = Modeling 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 the 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 variable are used to model rates.**
   A) Empirical B) Binomial
   C) Poisson D) All of the mentioned

Answer: (C) = Poisson.

6. **Usually replacing the standard error by its estimated value does change the CLT.**
   A) True B) False

Answer: (B) = False.

7. **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. **Normalized data are centered at _____ and have units equal to standard deviation 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 degree 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.

Q10 to Q15 are subjective answer type questions, Answer them in your own words briefly:

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

Answer:

Normal Distribution is a Probability Distribution of the data on graph which looks like a bell shaped curve and is symmetric about the mean, which shows that the data near mean value are more frequent on occurrence than data far from the mean.
In Normal Distribution, Mean is always assumed as 0 and the standard Deviation is 1 also it has a 0 skew and kurtosis of 3. Normal Distribution is always Symmetrical but not all symmetrical Distributions are Normal Distribution.

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

Answer:

Understanding the nature of missing data is critical in determining what treatments can be applied to overcome the lack of data. Data can be missing in the following ways:

1. Missing Completely At Random (MCAR):
2. Missing At Random (MAR):
3. Not Missing At Random (NMAR):

Imputation techniques that can be used to handle the missing data are as follows:

1. Mean or Median imputation:
   A common technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean or median can result in loss of variation in data and it is better to use imputations. Depending upon the nature of the missing data

2. Multivariate Imputation by Chained Equations (MICE):
MICE assumes that the missing data are missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, Bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.

3. Random Forest:
Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple decision trees to estimate missing values and outputs OOB imputation error estimates.

12. 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. Let's say if we want to increase the sale of a product. Here, either we can use random experiments, or we can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools. In the above scenario, we may divide the products into two parts – A and B. Here A will remain unchanged while we make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, we try to decide which is performing better. It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The **population** refers to all the customers buying your product, while the **sample** refers to the number of customers that participated in the test.

13. Is mean imputation of missing data acceptable practice?

Answer:

Mean imputation can be useful in cases where the number of missing observations is low. However, for large number of missing values, using mean imputation can result in loss of variation in data. Mean imputation does not preserve the relationships among the variables imputing the mean preserves the mean of the observed data. So if the data is missing completely at random the estimate of the Mean remains unbiased, which is a good thing. Mean imputation leads to an underestimates of standard errors. You get the same mean from mean-imputed data that you would have gotten without the imputations. And yes, there are circumstances where that mean is unbiased. Even so, the standard error of that mean will be too small. Because the imputations are themselves estimates, there is some error associated with them. But your statistical software doesn't know that. It treats it as real data. Ultimately,

because your standard errors are too low, so are your p-values. Now you're making Type 1 errors without realizing it. Which is not good.

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.

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 student of statistics.

Descriptive 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. 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 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.

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.