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 : 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 : 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 : 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 are dependent
   c) The square of a standard normal random variable follows what is called chi-squared distribution
   d) All of the mentioned

   Answer : All of the mentioned

5. _______random variables are used to model rates.

   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned

   Answer : Poisson

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

   a) True
   b) False

   Answer : 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 : 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 : 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 : 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?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing?
13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

Q 10 : What do you understand by the term Normal Distribution?

Answer : The normal distribution, also known as the Gaussian or standard normal distribution, is the probability distribution that plots all of its values in a symmetrical fashion, and most of the results are situated around the probability's mean. Values are equally likely to plot either above or below the mean.
The normal distribution is a probability distribution that (roughly) describes many common datasets in the real world. It is the most common type of distribution, and it arises naturally in statistics through random sampling techniques.

Despite the different shapes, all forms of the normal distribution have the following characteristic properties.

- They're all symmetric. The normal distribution cannot model skewed distributions.
  The mean, median, and mode are all equal.
- Half of the population is less than the mean and half is greater than the mean.
- The Empirical Rule allows you to determine the proportion of values that fall within certain distances from the mean.

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

Answer : 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.

Imputation techniques for the handling of missing data is as follows ;

- Mean or Median Imputation. When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. ...
- Multivariate Imputation by Chained Equations (MICE) MICE assumes that the missing data are Missing at Random (MAR). ...
- Random Forest.
- Imputation Using k-NN
- Stochastic regression imputation
- Extrapolation and Interpolation
- Hot-Deck imputation

Q 12 : What is A/B testing?

Answer: A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drive business metrics.

Essentially, A/B testing eliminates all the guesswork out of website optimization and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to 'control' or the original testing variable. Whereas B refers to 'variation' or a new version of the original testing variable.

Q 13 : What is linear regression in statistics?

Answer : 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 14: What are the various branches of statistics?

Answer: The 2 Branches of Descriptive Statistics & Inferential 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.

Q 15: Is mean imputation of missing data acceptable practice?

Answer :
Bad practice in general, If just estimating means: mean imputation preserves the mean of the observed data.

Leads to an underestimate of the standard deviation

Distorts relationships between variables by "pulling" estimates of the correlation toward zero