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

Ans. 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

Ans. 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

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

Ans. d) All of the mentioned

5. _____ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

Ans. c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False

Ans. 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

Ans. 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

Ans. 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

Ans. 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?

Ans.
Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.

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

Ans.

When dealing with missing data, data scientists can use two primary methods to solve the error: imputation or the removal of data.

The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

The other option is to remove data. When dealing with data that is missing at random, related data can be deleted to reduce bias.

Imputation techniques:
- Mean,Median and Mode
- Time-Series Specific Methods
- Substitution
- Hot deck imputation
- Cold deck imputation
- Regression imputation
- Stochastic regression imputation
- interpolation and extrapolation

12. What is A/B testing?

Ans.
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.

13. Is mean imputation of missing data acceptable practice?

Ans. No, mean imputation of missing data is not an acceptable practice due to the following reasons:
- Mean imputation does not preserve the relationships among variables.
- Mean imputation leads to an underestimate of standard errors.

14. What is linear regression in statistics?

Ans. 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?

Ans.
Two branches,descriptive statistics and inferential statistics, comprise the field of statistics.

- Descriptive Statistics : The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

  Examples: The average age of citizens who voted for the winning candidate in the last presidential election, the variation in the weight of 100 boxes of cereal selected from a factory's production line.

- Inferential Statistics : The branch of statistics that analyzes sample data to draw conclusions about a population.

  Examples: A survey that sampled 2,001 full-or part-time workers ages 50 to 70, conducted by the American Association of Retired Persons (AARP), discovered that 70% of those polled planned to work past the traditional mid-60s retirement age. By using methods discussed in Section 6.4, this statistic could be used to draw conclusions about the population of all workers ages 50 to 70.