STATISTICS WORKSHEET-1

Name: Darshan M.Katgeri

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

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

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

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

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

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

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

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

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

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?

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

A distribution of continuous random variable with a single peak, bell-shaped curve. The mean lies at the center of the distribution, and the curve is symmetrical about a vertical line erected at the mean. The two tails extend indefinitely, never touching the horizontal line. Most of the real life situations and phenomena may be modelled using Normal distribution.

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

Missing data can be handled in various ways. If the dataset consists of considerable observations then omission of complete row/column may be one way to handle missing value or missing data.

If dataset does not contain sufficient number of observations then omission may not be the ideal solution, in such situations the missing data must be replaced by the mean, mode or median of the corresponding feature.

Generally, missing data is replaced with mean.

In sklearn SimpleImputer can be used to deal with missing data.

12. What is A/B testing?
Answer.12:

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 drives 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.

The version that moves your business metric(s) in the positive direction is known as the 'winner.' Implementing the changes of this winning variation on your tested page(s) / element(s) can help optimize your website and increase business ROI.

A/B testing is one of the components of the overarching process of Conversion Rate Optimization (CRO), using which you can gather both qualitative and quantitative user insights. You can further use this collected data to understand user behavior, engagement rate, pain points, and even satisfaction with website features, including new features, revamped page sections, etc.

13. Is mean imputation of missing data acceptable practice?
Answer.14:

Yes, mean imputation of missing data acceptable practice.

14. What is linear regression in statistics?

Answer.14:

Conceptually, linear regression is a technique to find the relationship between input variable and output or target. This relationship is more of cause and effect relationship. The input is the independent variable and the output is the dependent variable. The main focus of linear regression is to find the best fit line or curve(in case of non-linear regression), it involves finding the slope of the line and the y-intercept.

Simple linear regression: In case of Simple linear regression there is only one independent variable.

$$y = \beta X + c \text{.....(i)}$$

Multiple linear regression: Multiple linear regression has more than one independent variable.

$$y = \beta_o + \beta_1 X_2 + \beta_2 X_2 + \beta_3 X_3 + c \text{.....(ii)}$$

Polynomial regression: This is non-linear type of regression.

$$y = \beta_o + \beta_1 X_1^2 + \beta_2 X_2^2 + c \text{.....(iii)}$$

15. What are the various branches of statistics?

Answer.15:

Statistics is broadly classified into two categories, namely,

i. Descriptive Statistics and
ii. Inferential statistics

Descriptive statistics basically involves graphs, tables and charts which make understanding the data easy.

Inferential statistics deals with arriving at educated decisions using information obtained from data. It involves calculating probabilities, understanding distributions and various metrics, and formation of decision trees etc.