STATISTICS WORKSHEET

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

Correct 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

Correct 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

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

Correct answer – d) All of the mentioned

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

Correct answer – c) Poisson

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

Correct 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

Correct answer – b) Hypothesis

8. Normalized data are centered at______and have units equal to standard deviations of the original data.
a) 0
b) 5
c) 1
d) 10

Correct answer – 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

Correct answer – 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?

Answer – Normal distribution also known as guardian distribution is the most important probability distribution in Statistics for independent random variable, the normal distribution describes how the values of variables are distributed.
Example – Height Data, Age Data, Sand Collection Data, etc

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

Answer – Data that is not stored or present for some variable in the datasets are Missing data, Missing data can bias the result of ML models or can reduce the accuracy of models There are Two ways to handle the missing data which are:

1) Deleting the Missing Data.
2) Imputing the Missing Data.

The KNN Imputer by sci kit -learn is a widely used method to impute missing values, It is widely being observed as a replacement for traditional imputation techniques. K- Nearest Neighbors work in a way that it identify the neighboring points through a measure of distance and the missing values can be estimated using completed value of neighboring observation.

12. What is A/B testing?

Answer – A/B testing also known as bucket testing or split run testing is a user experienced research methodology, A/B testing is a way to compare two version of a single variable typically by testing a subject response to variant A against variant B and determining which of the two variant is more effective.

A/B test are useful for understanding user engagements and satisfaction of online features like a new feature or products. A/B testing make user experience more successful.

13. Is mean imputation of missing data acceptable practice?

Answer – 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.

14. What is linear regression in statistics?

Answer – Ans: Linear regression is statistical way of measuring the relationship between variables such as if we take time and cost example: if the time increases so does the cost increases, It is used for future prediction also.

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.

Three major uses for regression analysis are
(1) determining the strength of predictors,
(2) forecasting an effect, and
(3) trend forecasting.

15. What are the various branches of statistics

Answer – The two main branches of statistics are:
1) Descriptive Statistics – Through graphs or tables, or numerical calculations, descriptive statistics uses the data to provide descriptions of the population.
2) Inferential Statistics – Based on the data sample taken from the population, inferential statistics makes the predictions and inferences.
