The answers are highlighted in green

STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.
a) True
b) False

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

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

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

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

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

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

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

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 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?

Answer:

Missing data appear when no value is available in one or more variables of an individual.

A. Deletions. Pairwise Deletion. Listwise Deletion/ Dropping rows. Dropping complete columns.
B. Basic Imputation Techniques. Imputation with a constant value. Imputation using the statistics (mean, median, mode)
C. K-Nearest Neighbor Imputation.

12. What is A/B testing?

Answer:

It is a way to compare the two versions of a variable to find out which performs better in a controlled environment. For example: You need to improve the number of hits on your website. You may approach the situation with two scenarios 'A' and 'B'. In Scenario 'A' remains unchanged and in scenario 'B' you've a Statistical approach. Now on the basis of responses from groups 'A' and 'B', we try to decide which is performing better. It is hypothetical testing method for making decisions that estimate population parameters based on sample statistics.

For A/B testing we need to formulate two hypothesis:

a. *Null Hypothesis (Ho)*: It that states that sample observations result purely from chance. The null hypothesis states that there is no difference between the control and variant groups. It states the default position to be tested or the situation as it is now, i.e. the status quo. Here our Ho is " there is no difference in the conversion rate in number of hits on website A and B".
b. *Alternate Hypothesis (Ha):* It is opposite of Null Hypothesis. Our Alternative Hypothesis will be 'The conversion rate of 'B' is higher than that of 'A''

After formulating the hypothesis, we proceed to dividing by random sampling into control group and test(variate) group. Control group receive approach 'A' while variate group receives approach 'B'.

We further perform test on our groups and evaluate the results.

▫ If p-value > alpha: Fail to reject the null hypothesis (i.e. not significant result).

▫ If p-value <= alpha: Reject the null hypothesis (i.e. significant result).

It may result in two types of error:

a. **Type I error (False Positive)**: We reject the null hypothesis when it is true. That is, we accept the variant B when it is not performing better than A

b. **Type II error (False Negative)**: We failed to reject the null hypothesis when it is false. It means we conclude variant B is not good when it performs better than A.

13. Is mean imputation of missing data acceptable practice?

Answer:

Mean imputation is not acceptable due to the following reasons:

✦ Mean Imputation ignore feature correlations: It distorts relationships between variables by "pulling" estimates of the correlation toward zero. This may lead to a bias.

✦ Mean reduces a variance of the data: A smaller variance leads to the narrower confidence interval in the probability distribution.

✦ Mean Imputation Leads to An Underestimate of Standard Errors: which leads to smaller p-value

14. What is linear regression in statistics?

Answer:

Linear Regression is a supervised machine learning algorithm which is derived from Statistical model of Linear Regression. It is the most commonly used predictive analysis, to find the modelling approach between a continuous dependent variable and a set of predictors which are independent.

15. What are the various branches of statistics?

Answer:

There are two major branches of statistics:

a. Descriptive Statistics: It deals with collection and statistical summarization of data. Statistical summarization means calculating important characteristics of data which may include finding Central tendencies (Mean, median, mode), Variance, Standard deviation, Covariance, Correlation etc.
b. Inferential Statistics: As the name suggests, inferential statistics draw right conclusions from descriptive statistics. We can draw samples from the population data and perform statistical analysis on sample data and later make the inferences from the population data. It determines the probability of the characteristics of the sample using probability theory. It includes hypothesis testing, Student's T- test, ANOVA test, Chi Square test etc.