WORKSHEET SET 1 – STATISTICS

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

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) Modelling event/time data b) Modelling bounded count data c) Modelling 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?

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?

ANSWERS:

1. a) True

2. a) Central Limit Theorem

3. b) Modelling bounded count data

4. d) All of the mentioned

5. c) Poisson

6. b) False

7. b) Hypothesis

8. a) 0

9. c) Outliers cannot conform to the regression relationship

10. The normal distribution is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the centre is a mirror image of the left side. The area under the normal distribution curve represents probability and the total area under the curve sums to one. In normally distributed data, there is a constant proportion of data points lying under the curve between the mean and a specific number of standard deviations from the mean. Thus, for a normal distribution, almost all values lie within 3 standard deviations of the mean.

11. To Work With Nan/Missing Values:

1.drop method(for low number of nan values) -pandas

2.replace method(for high number of nan values)-pandas

-mean,median=numerical values

-mode =categorical values

3.simple imputer-sk learn-(mean,median,most frequent)

12. A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.And thus helps us to statistically prove which hypothesis is true using the conversion rate.

13. If the data are missing completely at random, the estimate of the mean remains unbiased.Thus,imputing the mean preserves the mean of the observed data.But,when the data is skewed then outliers data points will have a significant impact on the mean and hence, in such cases, it is not recommended to use the mean for replacing the missing values.

14. Linear regression is a basic and commonly used type of predictive analysis.This 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.DESCRIPTIVE STATISTICS-collecting and summarizing the data

1.Measure of Central Tendency-Mean,Median,Mode

2.Measure of Dispersion-Range,Standard Deviation,Variance

INFERENTIAL STATISTICS-analysing the data to draw a conclusion

1.Hypothesis Testing-chi-square test,anova test,correlation test,t-test etc…