Statistics Worksheet 1

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

Solution: 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

Solution: 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

Solution: 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

Solution: d) All of the mentioned

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

Solution: c) Poisson

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

Solution: 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

Solution: 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

Solution: 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

Solution: 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 is a distribution that plots all of its values in a symmetrical fashion and most of the results are situated around the mean.
                                         OR
 Normal Distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

11. How do you handle missing data? What imputation techniques do you recommend?
Answer) Missing data can be handled in many ways. Ignoring the data is a commonly used method.   Choosing to make no decision otherwise, indicates that your programme will make a decision for you. Another common method is imputation.
Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values.

12. What is A/B testing?
Answer) A/B testing is a method to compare the 2 versions of a variable to find out which performs better in a controlled environment.

13. Is mean imputation of missing data acceptable practice?
Answer) Mean imputation of data is considered as a bad practice as it ignores feature correlation. Also it 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) Linear regression attempts to model the relationship between 2 variables by fitting a linear equation to observed data.

15. What are the various branches of statistics?

Answer) There are 3 branches of statistics: Data collection, Descriptive Statistics and Inferential statistics.

- Data Collection: it is all about how actual data is collected.
- Descriptive Statistics: it is the part of statistics that deals with presenting the data we have.
- Inferential Statistics: it is the aspect that deals with making conclusions about the data.