STATISTICS WORKSHEET-1

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

The curve has bell shaped in the middle. Hence half of data is on both left and right side. The mean, mode and median are all equal / same. As per my understanding it is used on continuous data related problems. About 68% of data falls within 1 standard deviation, 95% of data in 2 standard deviation and 99.7% in 3 standard deviations.

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

We can go ahead with Mean and Mode values as replacement for missing values. If the missing data is continuous go with mean and if its categorical go ahead with Mode.

Eg: if the missing data is gender, then we simply cannot take mean as it makes no sense. Mode could be the best option as it outputs the maximum occurring value in the list.

We can also go with backfill and forward fill using pandas, if need be.

12. **What is A/B testing?**

The term itself A/B testing is quite common among the web developers as its generally used by them. This has 2 approaches and the best approach / method is selected upon analysis. In statistics Hypothesis testing is also similar as we go with certain assumption at the beginning and its either true or false and we can identify Type 1 and Type 2 errors if any.

13. **Is mean imputation of missing data acceptable practice?**

Yes, Mean imputation of missing data is one of the acceptable methods if the dataset / problem we are working on is continuous in nature.
Eg: Height, weight,
But it should not be used for categorical data.
Eg: if a column has either True or False, we can't take mean of that column. Although we get statistics details of any dataset using .describe() , we need to use other techiques.

14. **What is linear regression in statistics?**

In regression analysis, linear regression is an approach used to analyze relation between dependent variable and a single independent variable.
Eg: age vs salary.

In any profession, age / experience of a person plays and important role determining what the salary could be. Along with other factors.

Age is an independent variable. Your experience will increase gradually and is continuous in nature.

However, the salary is higher for those having more experience compared to a fresher. Hence salary either increases or decreases based on age. (considering only age as a feature)

15. **What are the various branches of statistics?**

There are 2 kinds / branches in statistics.

1) Descriptive statistics
2) Inferential statistics

By definition itself, descriptive statistics / analysis is used on quantities and inferential statistics are used to draw conclusions ie. on factors like quality which cannot be measured directly.

Under Differential statistics we have 2 types:

1) Central tendency measures :
   a) Mean : gives the average of the given data
   b) Median : gives the middle most value of the given data. If the list has odd numbers then its takes the average of the middle 2 values.
   c) Mode: gives the most occurring value in the list.

2) Variability measures : This consists of quartiles (eg: box plot), standard deviation and variance.

Under Inferential statistics, we have:

1) Chi Square test
2) ANOVA test
3) ANCOVA test