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. 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?

The normal (or Gaussian) distribution is one particular kind of a bell shaped curve.

It is unimodal (that is, there is one peak"), symmetric (that is you can flip it around its midpoint) and its mean, median and mode are all equal.
However, it is only one such distribution - others meet all those conditions and are not normal.

Many things are approximately normally distributed, for example the heights of adult human females or males, IQ, etc.

11. How do you handle missing data? What imputation techniques do you recommend?
Below are the most common ways of handling missing data

Zero Replacement: Here, you replace the missing value with zero irrespective of everything.

Min or Max Replacement: Replace the missing value with the minimum or maximum value of a feature.

Mean/ Median/ Mode Replacement: Replace missing value with mean or median or most frequent feature value.

Also, one can replace the value of the missing cell with the previous cell's value. This kind of technique is popular while inputting time series data. For example, if the price of an instrument is missing on the i-th day, it makes sense to replace it with the (i-1)-th day's price.

12. What is A/B testing?
A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better.

A/B testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.

13. Is mean imputation of missing data acceptable practice?
Missing data is usually not acceptable but that depends a lot on the type of data you're mentioning.
Two examples:
Sensorial data (IoT) usually tends to be very tolerant to this due to the number of rows that a single machine produces in seconds. If you miss a sensor value you will get it again 100ms after.

In the opposite, financial data is intolerant to failures, some companies don't accept reconciliation differences at all (example: Banking). You can't say to a client "Oops i missed a row on a transfer and now your account is missing 5 dollars."

14. What is linear regression in statistics?

Linear Regression is a supervised machine learning algorithm, which uses labelled data to predict outcomes.
It estimates the relationship between independent and dependent variables using a line (the plane in case of more than one independent variable).
If there is one independent and one target variable (Dependent variable), it's a case of Simple linear regression.
If more than one independent variable against the target variable in prediction, It is a case of Multiple Linear Regression.
It takes numerical and continuous data, the data should be linear to get high accuracy through linear regression.

15. What are the various branches of statistics?

Statistics:
Statistics is a study of presentation, analysis, collection, interpretation and organization of data
There are two main branches of statistics
- Inferential Statistic.
- Descriptive Statistic.

Inferential Statistics:

Inferential statistics used to make inference and describe about the population. These stats are more useful when it's not easy or possible to examine each member of the population.

Descriptive Statistics:

Descriptive statistics are use to get a brief summary of data. You can have the summary of data in numerical or graphical form.