STATISTICS WORKSHEET- 6

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following can be considered as random variable?
a) The outcome from the roll of a die
b) The outcome of flip of a coin
c) The outcome of exam
d) All of the mentioned

2. Which of the following random variable that take on only a countable number of possibilities?
a) Discrete
b) Non Discrete
c) Continuous
d) All of the mentioned

3. Which of the following function is associated with a continuous random variable?
a) pdf
b) pmv
c) pmf
d) all of the mentioned

4. The expected value or _______ of a random variable is the center of its distribution.
a) mode
b) median
c) mean
d) bayesian inference

5. Which of the following of a random variable is not a measure of spread?
a) variance
b) standard deviation
c) empirical mean
d) all of the mentioned

6. The _________ of the Chi-squared distribution is twice the degrees of freedom.
a) variance
b) standard deviation
c) mode
d) none of the mentioned

7. The beta distribution is the default prior for parameters between ____________
a) 0 and 10
b) 1 and 2
c) 0 and 1
d) None of the mentioned

8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?
a) baggyer
b) bootstrap
c) jacknife
d) none of the mentioned

9. Data that summarize all observations in a category are called ___________ data.

a) frequency
b) summarized
c) raw
d) none of the mentioned

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What is the difference between a boxplot and histogram?
11. How to select metrics?
12. How do you assess the statistical significance of an insight?
13. Give examples of data that doesnot have a Gaussian distribution, nor log-normal.
14. Give an example where the median is a better measure than the mean.
15. What is the Likelihood?

Q10 Ans: A box plot and histogram are both graphical representations of data in statistics, but they differ in how they display the data and what type of information they convey.
A histogram is a chart that shows the distribution of a set of continuous data by dividing it into intervals or "bins" and plotting the number or frequency of observations that fall within each bin. It is a way to visually represent the shape and spread of the data, and can be used to identify patterns or outliers.

Q11 Ans: Selecting appropriate metrics for a machine learning model depends on the problem you are trying to solve and the type of data you are working with. Here are some steps that can help guide the selection process:
Define the problem:
Understand the data:
Identify performance metrics:
Consider business objectives

Q12 Ans: Define the null hypothesis:
Choose a significance level:
Select an appropriate statistical test:
Calculate the p-value:
Interpret the results: If the result is statistically significant, then we reject the null hypothesis and conclude that there is a significant difference between the groups or variables being compared. If the result is not statistically significant, then we fail to reject the null hypothesis and conclude that there is no significant difference.

Q13 Ans: Here are a few examples of data that do not have a Gaussian distribution, nor log-normal:
Pareto distribution:
Poisson distribution:
Beta distribution:
Power law distribution:

Q14 Ans: Consider the income of a group of people. The majority of people in the group might have a moderate income, but a small number of people might have extremely high incomes. In this case, the distribution of income would be skewed to the right, with the tail of the distribution on the right side. If we were to use the mean to measure the central tendency of the income data, the extremely high incomes would significantly raise the mean, making it appear as if the majority of people had higher incomes than they actually did.

Q15 Ans: Likelihood is a statistical concept that measures the probability of observing a given set of data, assuming that a particular model or hypothesis is true. It is the probability of the data given the model or hypothesis, rather than the probability of the model or hypothesis given the data, which is the basis of Bayesian inference.