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?

Answer: A boxplot displays summary statistics of a set of data (minimum, first quartile, median, third quartile, and maximum) while a histogram displays the distribution of a set of data by dividing it into bins and showing the frequency of data in each bin as bars.

11. How to select metrics?

Answer: To select metrics, consider the following steps:
1. Define the objectives of your project or system.
2. Determine the key factors that affect the objectives.
3. Choose metrics that measure the key factors.
4. Ensure that the metrics are accurate, relevant, and actionable.
5. Decide on the frequency of measurement.
6. Choose metrics that are easily understandable by stakeholders.
7. Consider the cost of measurement and availability of data.
8. Regularly evaluate and update the metrics if necessary.

12. How do you assess the statistical significance of an insight?

Answer: To assess the statistical significance of an insight, follow these steps:
1. Define the null hypothesis and the alternative hypothesis.
2. Choose a significance level, typically 0.05 or 0.01.
3. Determine the appropriate statistical test (e.g. t-test, ANOVA, chi-squared test) based on the type of data and hypotheses.
4. Calculate the test statistic and the p-value.
5. Compare the p-value to the significance level.
6. Draw a conclusion: reject or fail to reject the null hypothesis.
7. Consider the practical significance of the results and the effect size.
8. Validate the findings with additional tests if necessary.

13. Give examples of data that doesnot have a Gaussian distribution, nor log-normal.

Answer: Examples of data that do not have a Gaussian (normal) distribution or a log-normal distribution include:
1. Bernoulli distribution - models binary data with two outcomes (e.g. success or failure)
2. Poisson distribution - models count data (e.g. number of events in a given time interval)
3. Exponential distribution - models the time between events in a Poisson process
4. Weibull distribution - models the time to failure in reliability engineering
5. Gamma distribution - models continuous non-negative data with a shape parameter (e.g. income, rainfall)
6. Pareto distribution - models the distribution of wealth and income inequality
7. Logistic distribution - models binary classification problems in statistics and machine learning
8. Uniform distribution - models data that is evenly distributed over a range of values.

14. Give an example where the median is a better measure than the mean.

Answer: The median is a better measure than the mean when the data is heavily skewed (has extreme values) or contains outliers.
Example: Income of a country

- Mean: The average income could be heavily influenced by a small number of very high-income individuals, leading to a mean that is much higher than most of the population earns.

- Median: The median income represents the income of the middle person in the country, which is a more representative measure of the typical income.

15. What is the Likelihood?

Answer: Likelihood is a statistical concept that measures the probability of observing a specific set of data given a set of model parameters. It is often used in maximum likelihood estimation (MLE), a method of estimating the parameters of a statistical model that maximizes the likelihood of the observed data.
The likelihood is a function that maps from the model parameters to a probability value. The value of the likelihood gives an idea of how well the model fits the data. Maximizing the likelihood means finding the parameters that make the observed data the most probable given the model.
In summary, the likelihood is a tool that assesses the goodness of fit between a statistical model and the data, and helps in estimating the parameters of the model that best fit the data.