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 - Both boxplots and histograms are used to visualize data, but they have different approaches and purposes.

A histogram displays the distribution of data by dividing it into intervals or bins and counting the number of observations that fall into each bin. The bins are typically equal in width and the height of each bar represents the frequency or proportion of observations that fall into that bin. A histogram can reveal the shape of the distribution, such as whether it is symmetric or skewed, the location of the center, and the spread of the data.

A boxplot, on the other hand, provides a compact summary of the distribution of data by displaying its median, quartiles, and any outliers. The box represents the middle 50% of the data, with the bottom and top edges corresponding to the first and third quartiles, respectively. The whiskers extend from the box to the smallest and largest observations that are not considered outliers. Outliers are shown as individual points beyond the whiskers. A boxplot can provide a quick way to compare multiple datasets, identify skewness or outliers, and get an overall sense of the spread and central tendency of the data.

In summary, a histogram shows the distribution of data by binning and counting, while a boxplot summarizes the distribution by displaying its key statistics and identifying outliers.

11. How to select metrics?

Answer – Selecting the right metrics is critical for any data analysis project. Here are some steps to follow when selecting metrics:

1. Define your objectives: Before selecting metrics, it is important to define what you want to achieve with your analysis. This will help you to identify which metrics are most relevant to your project.
2. Consider your audience: Think about who will be using the metrics and what information they need to make decisions. The metrics you choose should be meaningful and relevant to your audience.
3. Identify key performance indicators (KPIs): KPIs are the most important metrics for your business or project. They should align with your objectives and be measurable.
4. Choose metrics that are relevant and actionable: Metrics should be relevant to your objectives and provide insights that can be acted upon. Avoid choosing metrics that are difficult to measure or do not provide useful information.
5. Use a variety of metrics: Using a variety of metrics can provide a more complete picture of your data. For example, you might use both quantitative and qualitative metrics, or use different metrics to measure different aspects of your project.
6. Test and refine your metrics: Once you have selected your metrics, it is important to test them to ensure they are providing meaningful insights. Refine your metrics as needed to ensure they are accurate and useful.

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

Answer - To access the statistical significance of an insight, we need to perform statistical hypothesis testing. The following are the general steps involved in hypothesis testing:

1. State the null hypothesis (H0) and alternative hypothesis (Ha)
2. Select the appropriate statistical test based on the data type, sample size, and distribution.
3. Choose the level of significance (alpha), typically set at 0.05.
4. Collect the data and calculate the test statistic.
5. Determine the p-value associated with the test statistic.
6. Compare the p-value to the alpha level to decide whether to reject or fail to reject the null hypothesis.
7. Draw conclusions based on the results of the hypothesis test.

If the p-value is less than or equal to the alpha level, we reject the null hypothesis and accept the alternative hypothesis, suggesting that the observed effect is statistically significant. If the p-value is greater than the alpha level, we fail to reject the null hypothesis and conclude that there is no evidence to suggest a statistically significant effect.

13. **Give examples of data that does not have a Gaussian distribution, nor log-normal.**

Answer - There are several examples of data that do not have a Gaussian or log-normal distribution. Some of them are:

1. Power law distributions: These distributions have a heavy tail and are characterized by a small number of events that have a very high frequency. Examples include the distribution of income, the distribution of city sizes, and the distribution of website links.
2. Exponential distributions: These distributions describe the time between events that occur at a constant rate, such as radioactive decay or the arrival of customers at a store.
3. Poisson distributions: These distributions describe the probability of a certain number of events occurring in a fixed interval of time or space, such as the number of phone calls received by a call center in an hour.
4. Uniform distributions: These distributions are characterized by a constant probability density function and are often used to model random variables with equal probability of occurring over a given interval.
5. Bimodal distributions: These distributions have two peaks in their probability density function and are often observed in data that represents a mixture of two or more underlying populations. An example could be the height distribution of a group of people that includes both adults and children.

It is important to note that many real-world datasets do not follow a simple mathematical distribution, but instead exhibit complex patterns and structures that require more sophisticated statistical models and techniques for their analysis.

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

Answer - The median is often considered a better measure of central tendency than the mean when the data is skewed or has outliers that can significantly affect the mean.

For example, let's consider the salaries of employees in a company. Suppose the salaries of the employees are as follows:

$30,000, $35,000, $40,000, $45,000, $50,000, $100,000

The mean salary of the employees in this case would be:

(mean) = ($30,000 + $35,000 + $40,000 + $45,000 + $50,000 + $100,000) / 6
= $300,000 / 6
= $50,000

However, we can see that the salary of $100,000 is an outlier that significantly affects the mean. The median salary in this case would be the middle value of the sorted salaries, which is $42,500. Hence, the median provides a better representation of the typical salary of the employees in this case.

15. **What is the Likelihood?**

Answer - Likelihood is a term used in statistics that refers to the probability of observing a set of data given a specific statistical model or hypothesis. In other words, it measures how well a certain probability distribution model explains or fits a set of data. The likelihood function is used to estimate the parameters of the model or to compare different models, and it is often used in maximum likelihood estimation. The maximum likelihood estimate (MLE) is the value of the parameter that maximizes the likelihood function.