WORKSHEET

STATISTICS WORKSHEET- 6

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

Ans.1 d) All of the mentioned

Ans2. a) Discrete

Ans.3 a) pdf

Ans.4 c) mean

Ans.5 c) empirical mean

Ans.6 a) variance

Ans.7 c) 0 and 1

Ans.8 b) bootstrap

Ans.9 b) summarized

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

10. What is the difference between a boxplot and histogram?

Ans. A histogram is a graphical representation of the distribution of numerical data, where the data is divided into a set of intervals or bins and the count or frequency of observations falling in each bin is represented by the height of a bar. On the other hand, a boxplot is a graphical representation of the distribution of numerical data through their quartiles, where the box represents the interquartile range, the whiskers extend to the minimum and maximum values within a certain distance from the box, and individual data points beyond the whiskers are shown as dots. In summary, while a histogram provides a detailed view of the distribution of data, a boxplot gives a simplified summary of the distribution and highlights any outliers.

11. How to select metrics?

Ans. Selecting metrics involves a few key steps:

a) Define goals and objectives: Before selecting metrics, it's important to define the goals and objectives of your project or business. What are you trying to achieve and what is most important to measure in order to assess progress towards those goals?

b) Identify key performance indicators (KPIs): KPIs are metrics that are critical to the success of your project or business. They should be specific, measurable, and directly related to your goals and objectives.

c) Consider data availability and reliability: Metrics should be based on data that is easily accessible and reliable. If the necessary data is not available, it may be necessary to adjust your goals or find new metrics.

d) Prioritize metrics: Focus on a small set of metrics that are most important to your goals and objectives. Too many metrics can be overwhelming and make it difficult to assess progress.

e) Monitor and evaluate metrics: Regularly review and analyze your metrics to evaluate progress and identify areas for improvement. Use metrics to guide decision-making and adjust goals and strategies as necessary.

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

Ans. To assess the statistical significance of an insight, you typically need to conduct a hypothesis test. Here are the basic steps:

State your null hypothesis and alternative hypothesis. The null hypothesis is the default assumption that there is no difference or no effect, while the alternative hypothesis is the hypothesis that you are testing, such as the hypothesis that there is a difference or an effect.

Choose an appropriate test statistic and determine its distribution under the null hypothesis. The choice of test statistic depends on the type of data and the hypothesis being tested. The distribution of the test statistic under the null hypothesis is used to determine the probability of observing the test statistic or a more extreme value if the null hypothesis is true.

Determine the level of significance (alpha). This is the probability threshold that you set to reject the null hypothesis. A common value for alpha is 0.05.

Calculate the p-value. This is the probability of observing the test statistic or a more extreme value if the null hypothesis is true. If the p-value is less than the level of significance (i.e., p-value < alpha), then you can reject the null hypothesis and conclude that the alternative hypothesis is supported.

Interpret the results. If the null hypothesis is rejected, you can conclude that there is statistical evidence to support the alternative hypothesis. However, it's important to also consider the effect size and practical significance of the result, as statistical significance does not necessarily imply practical significance.

Note that the specific details of conducting a hypothesis test depend on the type of data and the specific test being used, and it's important to carefully follow the appropriate procedures and assumptions for each test.

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

Ans. Here are a few examples of data that do not have a Gaussian distribution or log-normal distribution:

a) Pareto distribution: This distribution is commonly used to model phenomena where a small number of observations account for a large proportion of the total, such as wealth distribution, city populations, or internet traffic. The distribution is characterized by a heavy tail and does not have a finite variance.

b) Weibull distribution: This distribution is commonly used to model phenomena where the failure rate changes over time, such as the lifetime of electronic components, the time until a machine breaks down, or the time until a customer churns. The distribution is characterized by a decreasing or increasing hazard function and can take on a wide range of shapes.

c) Poisson distribution: This distribution is commonly used to model phenomena where events occur randomly and independently over time or space, such as the number of customers arriving at a store, the number of calls received by a call center, or the number of defects in a production process. The distribution is characterized by a single parameter that represents the average rate of occurrence.

d) Exponential distribution: This distribution is commonly used to model phenomena where events occur continuously and independently over time, such as the time between radioactive decays, the time until a machine fails, or the time until a customer completes a transaction. The distribution is characterized by a single parameter that represents the average time until the event occurs.

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

Ans. the median is often a better measure than the mean when dealing with skewed data or data that has outliers. In such cases, the median is less sensitive to extreme values than the mean, and provides a more representative measure of the central tendency of the data.

15. What is the Likelihood?

Ans. Likelihood is a concept in statistics that refers to the probability of observing a set of data given a specific value of one or more unknown parameters in a statistical model. The likelihood function is a function of the unknown parameters, and is used to estimate or infer the values of these parameters based on the observed data.

In other words, the likelihood function measures how well a set of parameter values explain or fit the observed data. The goal in statistical inference is often to find the values of the parameters that maximize the likelihood function, or to compare the likelihoods of different parameter values or models.

Note that likelihood is different from probability, in that probability refers to the chance of observing a particular set of data given a fixed set of parameters, while likelihood refers to the chance of observing a particular set of parameters given a fixed set of data.