STATISTICS WORKSHEET- 6

1. d) All of the mentioned
2. a) Discrete
3. a) pdf
4. c) mean
5. c) empirical mean
6. a) variance
7. c) 0 and 1
8. b) bootstrap
9. b) summarized
10. A histogram displays the distribution of data by dividing the data into a set of intervals (or "bins") along the range of the data, and then drawing a bar to represent the count (or proportion) of data points that fall within each bin. This allows us to see the shape of the distribution and identify features such as modes, outliers, and skewness.

    A box plot, on the other hand, provides a summary of the distribution by displaying a box that extends from the lower to upper quartile of the data, with a line inside the box representing the median value. The whiskers on the plot extend to the lowest and highest observed values within a certain range. Box plots are useful for quickly identifying the range, median, and skewness of a dataset, as well as potential outliers.
11. Selecting appropriate metrics is an important step in any data-driven project. The metrics you select should align with the goals of the project and provide a meaningful way to measure progress towards those goals. Here are some general steps to follow when selecting metrics:
    a. Define your goals: Clearly define the goals of the project and what you hope to achieve.
    b. Identify key performance indicators (KPIs): Determine the KPIs that will help you measure progress towards your goals. KPIs should be specific, measurable, and relevant to the project.
    c. Choose metrics: Select metrics that align with your KPIs and provide a clear picture of progress towards your goals. Make sure the metrics are objective and can be easily measured and tracked over time.
    d. Evaluate data availability: Ensure that the data required to calculate the metrics is available and can be accessed in a timely manner.
    e. Monitor and adjust: Monitor the metrics over time and make adjustments as necessary to ensure that they continue to align with the goals of the project.

    Overall, it's important to keep in mind that the metrics you select will have a significant impact on the decisions that are made based on the data. It's therefore critical to take the time to carefully select and evaluate the metrics to ensure they accurately reflect progress towards the project goals.
12. Assessing the statistical significance of an insight generally involves performing a hypothesis test to determine whether the observed effect is likely to have occurred by chance or is a real effect. The steps involved in assessing the statistical significance of an insight can vary

depending on the nature of the data, the research question, and the hypothesis being tested, but some common steps include:

a. Formulate the null hypothesis: This is the hypothesis that there is no real effect, and that any observed difference is due to chance. The null hypothesis is typically denoted as H0.
b. Formulate the alternative hypothesis: This is the hypothesis that there is a real effect, and that any observed difference is not due to chance. The alternative hypothesis is typically denoted as Ha.
c. Choose an appropriate test statistic: This is a measure of the difference between the observed data and what would be expected under the null hypothesis. The choice of test statistic depends on the nature of the data and the research question.
d. Determine the significance level: This is the level of risk that is acceptable for rejecting the null hypothesis when it is actually true. The most commonly used significance level is 0.05, which means that there is a 5% risk of falsely rejecting the null hypothesis.
e. Calculate the p-value: This is the probability of observing a test statistic as extreme as the one obtained, assuming that the null hypothesis is true. If the p-value is less than the significance level, the null hypothesis is rejected in favor of the alternative hypothesis.
f. Interpret the results: If the null hypothesis is rejected, it means that the observed effect is statistically significant at the chosen significance level. If the null hypothesis is not rejected, it means that the observed effect is not statistically significant at the chosen significance level.

It is important to note that statistical significance does not necessarily imply practical significance. A statistically significant effect may be too small to be of practical importance, or may be affected by other factors not considered in the analysis. Therefore, it is important to consider the context of the research question and the practical implications of the findings when interpreting the results of a hypothesis test.

13. There are many types of data that do not follow a Gaussian (normal) distribution or a log-normal distribution. Here are a few examples:
    a. Poisson distribution: This distribution is commonly used to model count data, such as the number of arrivals per hour at a store, the number of defects in a manufacturing process, or the number of customers who make a purchase.
    b. Binomial distribution: This distribution is used to model the number of successes in a fixed number of independent trials, such as the number of people who click on an online advertisement or the number of people who respond to a survey.
    c. Exponential distribution: This distribution is used to model the time between events, such as the time between arrivals of customers at a store or the time between failures of a machine.
    d. Gamma distribution: This distribution is used to model data that are positively skewed and have a long tail, such as income or insurance claims.
    e. Weibull distribution: This distribution is commonly used to model the time to failure of a component or system, as it can capture both the early failures and the wear-out failures.

f. Pareto distribution: This distribution is used to model data that have a few very large values and many small values, such as income, population, or website traffic.

14. The median is often a better measure than the mean for data sets that contain outliers or extreme values, because the median is not affected by outliers in the same way as the mean.

    For example, suppose you are analyzing the salaries of employees at a company, and one of the executives earns an extremely high salary compared to the rest of the employees. In this case, the mean salary would be greatly influenced by the executive's high salary, and would not be a good representation of the typical salary at the company. However, the median salary would not be affected as much by the outlier, and would be a better representation of the typical salary

15. In statistics, likelihood refers to the probability of observing a set of data given a particular hypothesis or model. More formally, the likelihood is the probability of observing the data under a specific set of model parameters.
    For example, suppose we have a coin that we believe may be biased, but we don't know the probability of heads or tails. We might model the coin as having some unknown probability of heads, which we'll call "p". If we then flip the coin 10 times and observe 6 heads and 4 tails, we can use the likelihood to determine the probability of observing this data for various values of p. The likelihood of observing 6 heads and 4 tails given a value of p is given by the binomial distribution.
    In general, the likelihood is a key concept in statistical inference, as it provides a way to compare different models or hypotheses based on how well they fit the observed data. The maximum likelihood estimation (MLE) method, for instance, is a common approach to estimating the parameters of a statistical model by maximizing the likelihood function.