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.a) frequency

10.A histogram is a graph showing the distribution of a numerical variable. It is constructed by dividing the data into intervals, called bins, and counting the number of observations that fall into each bin. The bins are usually represented as bars that touch each other.

A box plot, also known as a box-and-whisker plot, is a graph that summarizes the distribution of a numerical variable. It shows the five-number summary of the data: the minimum value, the first quartile (Q1), the median, the third quartile (Q3), and the maximum value. A box plot consists of a rectangular box that spans from Q1 to Q3, with a vertical line at the median. The whiskers extend from the box to the minimum and maximum values that are not outliers, which are represented as points.

The main difference between a histogram and a box plot is that a histogram shows the distribution of the data in a continuous manner, while a box plot shows the distribution of the data in a summary manner. A histogram is useful for exploring the shape of the distribution, including the presence of peaks, gaps, and outliers. A box plot is useful for comparing the distributions of different groups or variables, including the location, spread, and skewness of the data, as well as the presence of outliers.

11.Selecting metrics is an important step in data analysis and depends on the goals of the analysis. Here are some general steps to follow when selecting metrics:

Define the problem: Start by clearly defining the problem you are trying to solve or the goal you are trying to achieve. This will help you identify the key metrics you need to track.

Identify relevant data sources: Identify the data sources that are relevant to the problem you are trying to solve. This may include data from internal systems, third-party data providers, or publicly available data.

Brainstorm potential metrics: Brainstorm a list of potential metrics that could be relevant to the problem you are trying to solve. These might include measures of customer engagement, conversion rates, user retention, or revenue growth.

Evaluate metrics: Evaluate each potential metric based on how well it aligns with the problem you are trying to solve. Consider factors such as the availability and quality of data, how easily the metric can be measured, and how relevant it is to your business objectives.

Prioritize metrics: Once you have identified the most relevant metrics, prioritize them based on their importance to the problem you are trying to solve. This will help you focus your analysis on the most critical areas.

Monitor and adjust: Finally, regularly monitor your metrics and adjust them as needed based on changes in the business environment or shifts in your goals and priorities.

12.To assess the statistical significance of an insight, you need to perform statistical hypothesis testing. Here are the basic steps to perform a hypothesis test:

State the null hypothesis (H0): This is the hypothesis that there is no significant difference between the two groups or variables being compared.

State the alternative hypothesis (Ha): This is the hypothesis that there is a significant difference between the two groups or variables being compared.

Choose a significance level (alpha): This is the threshold probability at which you will reject the null hypothesis. Common choices for alpha are 0.05, 0.01, and 0.001.

Select an appropriate test statistic: This is the statistic that will be used to determine the probability of obtaining the observed data under the null hypothesis.

Calculate the p-value: This is the probability of obtaining a test statistic as extreme as the one observed, assuming that the null hypothesis is true.

Compare the p-value to the significance level: If the p-value is less than the significance level, then you reject the null hypothesis and conclude that there is evidence to support the alternative hypothesis. If the p-value is greater than the significance level, then you fail to reject the null hypothesis.

Interpret the results: If the null hypothesis is rejected, then the insight is considered statistically significant. If the null hypothesis is not rejected, then the insight is not considered statistically significant.

It is important to note that statistical significance does not necessarily imply practical significance or importance. Practical significance refers to the real-world implications of the insight and should also be considered when interpreting the results.

13.There are many examples of data that do not have a Gaussian or log-normal distribution. Here are some examples:

Poisson distributed data: Poisson distribution is a discrete probability distribution that is commonly used to model count data. Examples include the number of calls to a call center, the number of defects in a manufacturing process, or the number of accidents on a highway.

Exponential distributed data: Exponential distribution is a continuous probability distribution that is commonly used to model waiting times. Examples include the time between two consecutive earthquakes, the time between arrivals at a service center, or the time until a light bulb fails.

Power law distributed data: Power law distribution is a type of heavy-tailed distribution that is commonly used to model phenomena that exhibit scale invariance. Examples include the distribution of city sizes, the distribution of income, or the distribution of word frequencies in a text.

Bimodal distributed data: Bimodal distribution is a type of distribution that has two peaks or modes. Examples include the distribution of height in humans (which often has a mode around 5-6 feet and another mode around 6-7 feet), or the distribution of test scores in a bimodal grading system.

Bernoulli distributed data: Bernoulli distribution is a discrete probability distribution that is commonly used to model binary data. Examples include the outcome of a coin flip (heads or tails), the outcome of a vote (yes or no), or the presence or absence of a certain gene in a population.

14.The median is a better measure than the mean when the distribution is skewed or has outliers. For example, consider a dataset of salaries in a company where the majority of employees earn a modest salary, but a few high-ranking executives earn much higher salaries. In this case, the mean salary would be significantly influenced by the salaries of the executives, while the median salary would provide a better representation of the typical salary earned by employees. Therefore, in this case, the median would be a better measure than the mean.

15.Likelihood is a concept in statistics that measures the plausibility of a particular set of parameter values, given some observed data. It is often used in statistical inference to estimate parameters of a probability distribution based on the observed data. The likelihood function is a function of the parameters of the model and is defined as the probability of observing the data given the values of the parameters. The maximum likelihood estimate is the value of the parameter that maximizes the likelihood function, and is often used as the best estimate of the true value of the parameter. The likelihood can be used to compare different models and to perform hypothesis testing.