Statistics Worksheet - 6 Ans

1. d) All of the mentioned
2. a) Discrete
3. a) pdf
4. c) mean
5. c) empirical mean
6. d) none of the mentioned
7. c) 0 and 1
8. b) bootstrap
9. a) frequency

10. A histogram is a graph that shows the distribution of a numerical variable. The data is divided into intervals, or "bins," and the height of each bar represents the number of data points that fall within that bin. It can help identify the distribution's shape, any outliers, and other important features.

A boxplot, on the other hand, is a visual representation of the data distribution based on five summary statistics: minimum, first quartile, median, third quartile, and maximum. The box represents the interquartile range (IQR), and the whiskers extend to the highest and lowest values that are not considered outliers. It is helpful in identifying any outliers and provides a quick summary of the data distribution.

11. Metrics are selected based on the problem being solved and the desired outcome. In machine learning, the choice of metric depends on the task being performed. For example, in classification tasks, accuracy is a commonly used metric. In contrast, mean squared error (MSE) or mean absolute error (MAE) are often used in regression tasks. When selecting a metric, it is important to choose one that aligns with the business goal and objectives.

12. We do hypothesis testing to assess the statistical significance of insight. The first step is to define the null hypothesis, which states that there is no difference or relationship between the variables of interest. Then we choose an appropriate test statistic and calculate its value based on the sample data. We also determine the p-value, which is the probability of observing a test

statistic as extreme as the one we calculated, assuming the null hypothesis is true. Suppose the p-value is below a pre-defined significance level (usually 0.05). In that case, we reject the null hypothesis and conclude that there is a statistically significant difference or relationship. Suppose the p-value is above the significance level. In that case, we fail to reject the null hypothesis and conclude that there is insufficient evidence to support the insight.

13. Some examples of data that do not have a Gaussian distribution or log-normal distribution include:

a) Binomial distribution: This is a discrete distribution that models the number of successes in a fixed number of independent trials, where each trial has the same probability of success. For example, the number of heads in 10-coin flips follows a binomial distribution.
b) Poisson distribution: This is another discrete distribution that models the number of events that occur in a fixed interval of time or space, given a known average rate. For example, the number of arrivals at a bus stop in an hour could follow a Poisson distribution.
c) Poisson distribution: This is another discrete distribution that models the number of events that occur in a fixed interval of time or space, given a known average rate. For example, the number of arrivals at a bus stop in an hour could follow a Poisson distribution.

14. There are many examples where median is a better measure than mean. For example, salary package in placement of the students as students being placed at significant higher packages are low but those packages impact the mean greatly, while median is the middle value when the data is sorted in ascending or descending order and is clear indicator of typical salary package in placements through the institute.

15. The likelihood is a term used in statistics to describe the probability of observing the data given the model or parameters. It determines the best parameters or model that explains the observed data.