NAME: Dipesh Ramesh Limaje

INTERNSHIP BATCH : 33

TOPIC: STATISTICS

SME : Mr. Shwetank Mishra

WORKSHEET NO : 6

Q1 Ans D All of the mentioned

Q2 Ans A Discrete

Q3 Ans A PDF

Q4 Ans C mean

Q5 Ans A variance

Q6 Ans A variance

Q7 Ans C 0 and 1

Q8 Ans B bootstrap

Q9 Ans B summarized

Q10 What is the difference between a boxplot and histogram?
Ans

Parameter	Histogram	BoxPlot
Purpose	The main purpose of a histogram is to show the frequency distribution of a set of continuous data	while a boxplot is used to display the summary statistics of a set of continuous data and to identify potential outliers.
Shape	A histogram is typically shaped like a bar graph, with bars of varying heights representing the frequency of data values within certain ranges or bins.	A boxplot is shaped like a box, with a line representing the median value, and two lines extending from either end of the box to represent the lower and upper quartiles.
Data Type	Histograms are used with continuous data	while boxplots can be used with either continuous or categorical data.
Outlier Identification	while histograms do not typically have this capability.	Boxplots are particularly useful for identifying outliers in the data
Distribution Information	Histograms can provide a lot of information about the distribution of the data, including skewness, kurtosis, and the shape of the distribution.	Boxplots provide a less detailed view of the distribution, but they are still useful for quickly identifying key features such as the median, quartiles, and range.
Bin Size	The size of the bins in a histogram can have a significant impact on the shape and interpretation of the histogram.	This is not a factor in boxplots, as the quartiles and median are calculated from the raw data, rather than being based on binned data.
Comparison between different datasets	Histograms can also be used for comparison, but this requires creating multiple histograms and comparing them, which can be time-consuming and less intuitive than comparing boxplots.	Boxplots are useful for comparing the summary statistics between different datasets, as they provide a compact representation of the data that can be easily compared between groups.

Q11] How to select metrics?

ANS

- Define the objective: The first step is to clearly define the objective of the study. This will help determine the type of metrics that are relevant and important to measure.

- Determine the population: The next step is to determine the population that you are interested in studying. This will help you select the appropriate metrics and determine the sample size needed to get accurate results.

- Choose relevant metrics: Based on the objective and population, select relevant metrics that best represent the information you are trying to obtain. For example, if your objective is to determine the average height of a population, the mean would be an appropriate metric.

- Consider measurement units: Consider the units of measurement for the metrics you have chosen. This will ensure that you are using comparable units and will make it easier to compare results.

- Evaluate the reliability and validity of metrics: Ensure that the metrics you have selected are reliable and valid. Reliable metrics produce consistent results, while valid metrics measure what they are supposed to measure.

- Test the metrics: Test the metrics you have chosen with a sample of the population. This will help you determine the accuracy and precision of your results and make any necessary adjustments.

- Review and revise: Finally, review and revise your metrics as needed. This will ensure that you are using the best metrics for your study and will help you achieve accurate results.

Q12] How do you assess the statistical significance of an insight?

ANS

- Assessing the statistical significance of an insight involves determining whether the results of a statistical test are likely due to chance or whether they are real and meaningful differences. To assess the statistical significance of an insight, you can use a hypothesis test.

- State the null hypothesis: The null hypothesis states that there is no difference between the populations or groups being compared.

- State the alternative hypothesis: The alternative hypothesis states that there is a difference between the populations or groups being compared.

- Choose a test statistic: Choose a test statistic that is appropriate for the data and the hypotheses being tested. Common test statistics include t-tests, ANOVA, chi-squared tests, and regression analysis.

- Calculate the test statistic: Calculate the test statistic based on the sample data.

- Determine the p-value: The p-value is the probability of observing a test statistic as extreme or more extreme than the one calculated, assuming the null hypothesis is true. If the p-value is less than a predetermined significance level (usually 0.05), the null hypothesis is rejected and the alternative hypothesis is accepted.

- Interpret the results: If the p-value is less than the significance level, the results are considered statistically significant, and the insight is considered meaningful and not likely due to chance. If the p-value is greater than the significance level, the results are considered not statistically significant, and the insight is considered likely due to chance.

Q13] Give examples of data that doesnot have a Gaussian distribution, nor log-normal.
ANS

There are many examples of data that do not have a Gaussian (normal) distribution or a log-normal distribution. Some common examples include:

- Exponential distribution: The exponential distribution is often used to model the time between events in a Poisson process, such as the time between customers arriving at a store.

- Weibull distribution: The Weibull distribution is commonly used in reliability engineering to model the time to failure of a product.

- Pareto distribution: The Pareto distribution is often used to model the distribution of wealth or income. It has a long right tail, which means that a small percentage of the population has a large proportion of the wealth or income.

- Poisson distribution: The Poisson distribution is used to model the number of events that occur in a given time period or in a given area, such as the number of customers arriving at a store or the number of calls received at a call center.

- Cauchy distribution: The Cauchy distribution is a heavy-tailed distribution that is commonly used in finance to model the returns of assets.

- Logistic distribution: The logistic distribution is often used in modeling the growth of populations, particularly in the early stages of growth.

Q14] Give an example where the median is a better measure than the mean.
ANS The median is often a better measure than the mean in situations where the data is highly skewed or contains outliers. For example:

- Income data: Income data is often highly skewed, with a few individuals having much higher incomes than the rest of the population. The mean income in this case would be heavily influenced by these high-income individuals, while the median income would give a better representation of the typical income.

- Housing prices: Housing prices can vary widely in different neighborhoods, with a few high-priced homes distorting the mean value. In this case, the median housing price would give a better representation of the typical housing price in a neighborhood.

- Pollution levels: In the case of pollution levels, outliers such as large industrial facilities can greatly influence the mean value. The median pollution level would give a better representation of the typical pollution level in a region.

Q15] What is the Likelihood?
ANS In statistics, the likelihood is a measure of how well a statistical model fits a set of observed data. It is a function that assigns a probability to each possible set of parameter values for a given model, based on how well the model predicts the observed data.

The likelihood is used to estimate the parameters of a statistical model that best fit the data. The idea is to find the values of the parameters that maximize the likelihood, which corresponds to finding the best fitting model for the data. This is known as maximum likelihood estimation.

For example, if we have a normal distribution with unknown mean and standard deviation, we can use maximum likelihood estimation to find the values of the mean and standard deviation that maximize the likelihood of observing the data we have. These estimates will give us the most likely values of the mean and standard deviation that generated the data.

The likelihood can also be used to perform hypothesis tests, such as testing the goodness of fit of a model or comparing the fit of two or more models. The likelihood is a valuable tool for statistical modeling and inference, as it provides a way to assess the fit of models to data and to estimate the parameters of models that best describe the data.