STATISTICS WORKSHEET- 4

Q.1] D

Q.2] A

Q.3] A

Q.4] C

Q.5] C

Q.6] A

Q.7] C

Q.8] B

Q.9] A

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

ANS. = A boxplot and a histogram are both graphical representations of data, but they provide different types of information and are used for different purposes.

A boxplot (also known as a box-and-whisker plot) is a way to visualize the distribution of a dataset. It shows the minimum, first quartile (Q1), median, third quartile (Q3), and maximum values of the data, as well as any outliers. The box represents the interquartile range (IQR) between Q1 and Q3, which contains the middle 50% of the data. Boxplots are useful for identifying the center, spread, and skewness of a dataset, as well as for comparing the distribution of multiple datasets.

A histogram, on the other hand, is a way to visualize the frequency distribution of a dataset. It shows how often each value or range of values occurs in the data by plotting the number of observations in each bin (or range of values) on the y-axis and the values on the x-axis. Histograms are useful for understanding the shape of a dataset, identifying patterns and trends, and comparing the distribution of multiple datasets.

In summary, a boxplot is used to display the summary of the set of data having properties like minimum, first quartile, median, third quartile and maximum, whereas histogram displays the frequency distribution of a numeric variable.

Q.11] How to select metrics?

ANS. = Selecting metrics to evaluate a model or process can be a complex task and it depends on the specific problem and context. Here are some general guidelines to follow when selecting metrics:

1. Understand the problem and its objectives: Before selecting metrics, it's important to have a clear understanding of the problem you're trying to solve and the objectives you're trying to achieve.

2. Identify the key aspects of the problem: Identify the key aspects of the problem that you want to measure and focus on selecting metrics that align with those aspects.

3. Choose metrics that are relevant and meaningful: Select metrics that are relevant and meaningful to the problem and its objectives, not just metrics that are easy to collect or calculate.

4. Consider multiple metrics: Don't rely on just one metric to evaluate your model or process. Use multiple metrics to get a more complete picture of its performance.
5. Interpretation of the metrics: Also, consider the interpretability of the metrics, make sure that it is easy to understand and interpret the results.

6. Baseline Performance: Define a baseline performance, that allows to compare the results of your model with a simple or naive model.

7. Adapt to the context: Consider the specific context in which the problem is being addressed and choose metrics that are appropriate for that context.

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

ANS. = Assessing the statistical significance of an insight involves determining the probability that the observed results occurred by chance. This is typically done by performing a hypothesis test, which involves comparing the observed data to a null hypothesis, which is a statement about the population that is being studied.

The general steps for assessing the statistical significance of an insight are:

1. Define a null hypothesis: The null hypothesis is typically that there is no difference or relationship between the variables being analyzed.

2. Define an alternative hypothesis: The alternative hypothesis is typically that there is a difference or relationship between the variables being analyzed.

3. Determine a significance level: The significance level is the probability of rejecting the null hypothesis when it is true. Commonly used significance levels include 0.05 and 0.01.

4. Calculate a test statistic: Depending on the type of data and analysis, a test statistic such as a p-value, t-value, or z-score is calculated to determine the likelihood of the observed results occurring by chance.

5. Compare the test statistic to the significance level: If the test statistic is less than the significance level, the results are considered statistically significant and the null hypothesis is rejected.

6. Interpret the results: The statistical significance of an insight can be interpreted in terms of the effect size and the p-value. The effect size measures the magnitude of the difference or relationship found between variables. The p-value measures the likelihood that the results are due to chance.

It is important to keep in mind that statistical significance does not always imply practical significance. It's important to also consider the practical implications of the findings and the sample size used in the analysis.

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

ANS. = Examples of data that does not have a Gaussian (normal) distribution, nor log-normal include:

1. Categorical data (data that can be divided into categories, such as "red" or "blue")

2. Count data (data that represents the number of occurrences of something, such as the number of bugs in a field)

3. Binary data (data that can only take on two values, such as "yes" or "no")

4. Power-law distributed data (data that follows a power-law distribution, such as the frequency of words in a language)

5. Exponential distribution data (data that follows an exponential distribution, such as time between events in a Poisson process)

6. Pareto distribution data (data that follows a Pareto distribution, such as income distribution)

7. Skewed data (data that has a long tail on one side, such as income or wealth distribution)

8. Non-Parametric data (data that cannot be fit by a simple probability distribution, such as data from a chaotic system)

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

ANS. = An example where the median is a better measure than the mean is when the data has outliers or extreme values. Outliers are values that are significantly larger or smaller than the majority of the data.

For instance, consider a dataset of the prices of houses in a neighbourhood. The mean price would be influenced by a small number of extremely expensive houses, leading to a high mean price value, however, the median price would give a more accurate representation of the typical price of the majority of houses in the neighbourhood, which would be less affected by the extreme values. In this case, the median would be a better measure of central tendency than the mean.

Q.15] What is the Likelihood?

ANS. = Likelihood is a statistical concept that represents the probability of a set of observations given a set of parameters for a specific probability distribution. It is often used in statistical inference to estimate the parameters of a model.

For example, suppose you are trying to estimate the mean and standard deviation of a normal distribution that describes a set of observations. The likelihood function is the probability of the observations given the estimated mean and standard deviation. The maximum likelihood estimate (MLE) is the set of parameter values that maximize the likelihood function, and it is often used as an estimate of the true values of the parameters.

In short, likelihood is a way to measure how well a set of parameters fit a given set of data, it's a measure of goodness of fit. It's a function of the parameters of the model and can be used to find the best parameter values. The best parameter values are the values that make the observed data most probable.