STATISTICS WORKSHEET- 6

1. d

2. a

3. a

4. c

5. c

6. a

7. c

8. b

9. a

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

A boxplot and a histogram are both graphical representations of data, but they display different types of information.

A boxplot, also known as a box-and-whisker plot, is a way of summarizing a set of data through five number summary, it shows the minimum value, first quartile, median, third quartile, and maximum value of a dataset. It also shows any outliers present in the data.

A histogram is a graphical representation of the distribution of a dataset. It shows the frequency of the different values that occur in a dataset. The x-axis represents the data values, and the y-axis represents the frequency of those values. It is used to give an idea about the frequency and distribution of the data.

So, the main difference between a boxplot and a histogram is that a boxplot shows a summary of the distribution of a dataset, while a histogram shows the frequency of the different values in the dataset.

11. **How to select metrics?**

Understand the problem statement and the business objectives of the model. This will help us determine the metrics that are most relevant to the problem.

Define the goals of the model, such as accuracy, precision, recall, F1-score, etc.

Choose the appropriate metrics based on the problem and goals. For example, accuracy is commonly used for classification problems, while mean squared error is commonly used for regression problems.

Consider using multiple metrics to evaluate the model. This will give us a more complete understanding of the model's performance.

Monitor the metrics over time and make sure the model is not overfitting or underfitting, this can be done by using train and test dataset.

Communicate the results clearly, showing the metrics and their values and explaining the context of the results.

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

We basically consider following facts -

The null hypothesis is typically that there is no real effect, while the alternative hypothesis is that there is a real effect.

Typically, a significance level of 0.05 or 0.01 is used, which corresponds to a 5% or 1% chance of the results being due to chance.

Calculate a test statistic: This is a measure of how far the observed data is from what would be expected under the null hypothesis.

Calculate 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.

Compare the p-value with the significance level: If the p-value is less than the significance level, the null hypothesis is rejected and the insight is considered statistically significant.

It's important to keep in mind that statistical significance does not always imply practical significance, so it's important to also evaluate the size of the effect and whether it is meaningful in the context of the problem.

13. **Give examples of data that does not have a Gaussian distribution, nor log-normal.**

There are many types of data that do not have a Gaussian (normal) distribution or a log-normal distribution. Here are a few examples-

Bernoulli distribution: This distribution is used for binary data, where the outcome is either a success or a failure. It is a discrete probability distribution and the probability of success is represented by a single parameter.

Poisson distribution: This distribution is used for modeling rare events, such as the number of customers arriving at a store. It is a discrete probability distribution and the parameter is the average number of events per unit of time or space.

Exponential distribution: This distribution is used to model the time between events in a Poisson process. It is a continuous probability distribution and the parameter is the rate of events.

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

The median is a better measure than the mean in cases where the data is heavily skewed or has outliers, i.e. when the data is not symmetric or has outliers, the median is a better measure of central tendency than the mean. **For example**, consider a dataset of the salaries of a group of employees in a company. If the company has a small number of highly paid executives, the mean salary will be significantly higher than the median salary, which will give a misleading idea of the typical salary in the company. In this case, the median salary would be a better measure of central tendency as it is not affected by the high salaries of the executives and would give a more accurate representation of the typical salary in the company.

15. **What is the Likelihood?**

In statistics, the likelihood is a function that expresses the probability of observing a particular set of data given a set of model parameters. It is a measure of how well the model fits the data, and it is used to estimate the parameters of the model.