STATISTICS WORKSHEET- 6

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following can be considered as random variable?
   a) The outcome from the roll of a die
   b) The outcome of flip of a coin
   c) The outcome of exam
   d) All of the mentioned
2. Which of the following random variable that take on only a countable number of possibilities?
   a) Discrete
   b) Non Discrete
   c) Continuous
   d) All of the mentioned
3. Which of the following function is associated with a continuous random variable?
   a) pdf
   b) pmv
   c) pmf
   d) all of the mentioned
4. The expected value or_____of a random variable is the center of its distribution.
   a) mode
   b) median
   c) mean
   d) bayesian inference
5. Which of the following of a random variable is not a measure of spread?
   a) variance
   b) standard deviation
   c) empirical mean
   d) all of the mentioned
6. The_____of the Chi-squared distribution is twice the degrees of freedom.
   a) variance
   b) standard deviation
   c) mode
   d) none of the mentioned
7. The beta distribution is the default prior for parameters between _____
   a) 0 and 10
   b) 1 and 2
   c) 0 and 1
   d) None of the mentioned
8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?
   a) baggyer
   b) bootstrap
   c) jacknife
   d) none of the mentioned

9. Data that summarize all observations in a category are called_____data.
   a) frequency
   b) summarized
   c) raw
   d) none of the mentioned

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What is the difference between a boxplot and histogram?
11. How to select metrics?
12. How do you assess the statistical significance of an insight?
13. Give examples of data that doesnot have a Gaussian distribution, nor log-normal.
14. Give an example where the median is a better measure than the mean.
15. What is the Likelihood?

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

A boxplot and histogram are both graphical representations of data distribution, but they have different ways of displaying the data.

A histogram is a bar graph-like representation of data that displays the frequency of values within a range of values called "bins." It shows the distribution of data over a continuous interval or certain ranges. Histograms are useful for showing the underlying frequency distribution of a continuous variable, such as age or income. The histogram shows the shape of the data distribution, including its center, spread, and any skewness or outliers.

A boxplot, also called a box-and-whisker plot, is a way to summarize the distribution of a dataset. It displays the median, quartiles, range, and any outliers in a compact manner. The box represents the interquartile range (IQR), which is the range of the middle 50% of the data. The whiskers extend from the box and indicate the range of the data, excluding any outliers. Boxplots are useful for comparing the distribution of multiple datasets or for identifying the presence of outliers.

In summary, histograms display the frequency distribution of a continuous variable by dividing it into intervals, while boxplots summarize the distribution of a dataset by displaying its quartiles, median, and any outliers.

11. How to select metrics?

Selecting the appropriate metrics for a machine learning project is crucial as it directly impacts the performance evaluation of the model. Here are some general guidelines to consider when selecting metrics:

Task-specific: Choose metrics that are relevant to the specific task of your machine learning project. For example, if you are working on a classification problem, you might want to choose metrics like accuracy, precision, recall, and F1 score.

Interpretability: Metrics should be easy to understand and interpret. It should be easy to communicate the metric and its value to stakeholders.

Robustness: Metrics should be robust to outliers and anomalies. Outliers should not overly influence the metric.

Scalability: Metrics should scale well with large datasets. If the dataset is too large, computing the metric should still be feasible.

Consistency: The selected metrics should be consistent with the business objectives and goals of the project. The metric should reflect the desired outcome of the model.

Comparative: The selected metric should allow for comparison between different models and approaches. For example, if you are comparing two models, it is essential to use the same metric to evaluate both models.

Trade-offs: Different metrics can have trade-offs. For example, improving accuracy might lead to a decrease in recall. It is essential to understand these trade-offs and choose a metric that best suits your requirements.

Overall, selecting the right metric(s) for your machine learning project is essential for evaluating the performance of your model and ensuring it meets the desired objectives.

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

To assess the statistical significance of an insight, one typically conducts hypothesis testing. Hypothesis testing involves comparing the observed result to what would be expected by chance, assuming a null hypothesis. The null hypothesis typically represents no difference or no effect, while the alternative hypothesis represents the hypothesis that the researcher is trying to test.

There are several steps involved in hypothesis testing:

1. Formulate the null and alternative hypotheses.
2. Choose a level of significance (alpha), which is the probability of rejecting the null hypothesis when it is actually true. A common value for alpha is 0.05.
3. Choose a test statistic that is appropriate for the data and the hypotheses being tested.
4. Determine the distribution of the test statistic under the null hypothesis.
5. Calculate the p-value, which is the probability of observing a test statistic as extreme as the one observed, assuming the null hypothesis is true.
6. Compare the p-value to the level of significance. If the p-value is less than the level of significance, reject the null hypothesis in favor of the alternative hypothesis. Otherwise, fail to reject the null hypothesis.
7. If the null hypothesis is rejected, this suggests that the observed result is statistically significant and unlikely to have occurred by chance. However, it is important to note that statistical significance does not necessarily imply practical or clinical significance, and further investigation may be necessary to determine the practical implications of the result.

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

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

Power law distributed data: Examples of such data are city populations, sizes of earthquakes, the distribution of wealth, and the frequency of words in natural language.

Exponential distributed data: Examples of such data are the lifetime of electronic components, the time between failures of a machine, or the time between calls to a call center.

Bimodal distributed data: Examples of such data are SAT scores, where students tend to score either high or low, but not in the middle, or the income distribution in some countries, where there is a large gap between the rich and the poor.

Poisson distributed data: Examples of such data are the number of phone calls received by a call center per hour, the number of defects per unit of production, or the number of accidents per day on a road.

Weibull distributed data: Examples of such data are the time to failure of a product, the time between repairs of a machine, or the time until an event occurs.

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

The median is a better measure than the mean in situations where the data is skewed or contains outliers. One such example is income data, where the distribution is often heavily skewed towards the high end. In this case, the mean can be significantly influenced by the presence of a few very high-income earners, while the median would be a more representative measure of the central tendency of the data. Another example is test scores, where the presence of a few very high or very low scores can skew the mean, while the median would provide a more accurate representation of the typical performance of students.

15. What is the Likelihood?

Likelihood refers to the probability of observing a set of data given a certain hypothesis or model. In other words, it measures how well the model or hypothesis fits the data. The likelihood function is a function that describes the probability of observing the data given the model parameters. The maximum likelihood estimate is the set of parameter values that maximizes the likelihood function and is often used to estimate the model parameters. The likelihood function plays a central role in statistical inference and is commonly used in techniques such as maximum likelihood estimation, Bayesian inference, and hypothesis testing.