STATISTICS WORKSHEET- 6

1) Which of the following can be considered as random variable?

Answer: d) All of the mentioned

****
****

2) Which of the following random variable that take on only a countable number of possibilities?

Answer: a) Discrete

****
****

3) Which of the following function is associated with a continuous random variable?

Answer: a) pdf

****
****

4) The expected value or _______ of a random variable is the center of its distribution.

Answer: c) mean

****
****

5) Which of the following of a random variable is not a measure of spread?

Answer: c) empirical mean

****
****

6) The _________ of the Chi-squared distribution is twice the degrees of freedom.

Answer: a) variance

****
****

7) The beta distribution is the default prior for parameters between

Answer: c) 0 and 1

****
****

8) Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?

Answer: b) bootstrap

****
****

9) Data that summarize all observations in a category are called __________ data.

Answer: b) summarized

****
****

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

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

Answer: A histogram is a graphical representation of the distribution of a continuous variable. It is a bar graph-like representation of the data that separates it into different ranges or bins. The height of each bar represents the number of data points that fall within each bin.

A boxplot, on the other hand, is a graphical representation of the distribution of a continuous variable based on five summary statistics: minimum, first quartile, median, third quartile, and maximum. It is a box-and-whisker diagram that gives a visual indication of the data's median, the interquartile range (IQR), and the data's variability. The main difference between a boxplot and a histogram is that a histogram shows the distribution of the data using the frequency of data points within certain intervals, while a boxplot shows the distribution of the data based on summary statistics.

****
****

11) How to select metrics?

Answer: Selecting appropriate metrics is an important step in any machine learning or data analysis project. Here are some general guidelines to help you select the right metrics:

Start with the project objective: The metrics you choose should be aligned with the project objective. For example, if your objective is to maximize accuracy, then accuracy is the most appropriate metric.

Consider the business context: You should choose metrics that are relevant to the business context of your project. For example, if you are working on a fraud detection system, then precision may be more important than recall.

Understand the limitations of the data: Sometimes the data may have certain limitations that make it difficult to use certain metrics. For example, if the data is imbalanced, then accuracy may not be a suitable metric.

Evaluate multiple metrics: It is often useful to evaluate multiple metrics to get a more complete picture of model performance. For example, in addition to accuracy, you may want to evaluate precision, recall, and F1 score.

Consider the tradeoffs: Some metrics may be in conflict with each other, so you need to consider the tradeoffs. For example, increasing recall may decrease precision, so you need to find a balance that works for your project.

Choose metrics that are easy to interpret: Finally, it is important to choose metrics that are easy to interpret and communicate to stakeholders. A metric like accuracy is easy to understand and explain, whereas more complex metrics may be difficult to explain to non-technical stakeholders.

****
****

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

Answer: To assess the statistical significance of an insight, it is important to use appropriate statistical tests and methods. This involves setting up a hypothesis, collecting and analyzing data, and using statistical tests to determine if the results are significant or not.

The level of significance is typically set at a p-value of 0.05 or lower, which means that there is a 5% or lower chance that the results are due to chance. It is also important to consider the effect size, which measures the practical significance of the result.

Additionally, it is important to consider the context and potential confounding factors that could impact the results. This includes considering the study design, sample size, and any potential biases or limitations.

****
****

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

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

Power law distributed data: Power law distributions are characterized by a small number of very large events or values and a large number of very small events or values. Examples include the distribution of the sizes of earthquakes or the popularity of websites.

Exponential distributed data: Exponential distributions are characterized by a rapid decay in the frequency of events as the value of the event increases. Examples include the time between successive requests to a website or the time between the failure of components in a system.

Bimodal distributed data: Bimodal distributions have two peaks in the distribution, indicating two distinct modes or clusters of data. Examples include the height of males and females in a population or the distribution of income in a population.

Poisson distributed data: Poisson distributions are used to model the number of occurrences of an event in a fixed interval of time or space. Examples include the number of cars passing through a toll booth in an hour or the number of mutations in a DNA sequence.

Uniform distributed data: Uniform distributions are characterized by all values in the distribution having an equal probability of occurring. Examples include the results of rolling a fair die or selecting a random number between 1 and 10.

These are just a few examples of data that do not have a Gaussian or log-normal distribution, and there are many other types of distributions that can occur in different contexts.

****
****

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

Answer: The median is a better measure than the mean in situations where there are extreme values or outliers that may heavily influence the mean.

For example, let's say we want to determine the typical income of a group of individuals. We have the following income data: 30,000,35,000, 40,000,45,000, 50,000,1,000,000.

The mean income for this group is calculated by summing all the incomes and dividing by the number of individuals in the group:

Mean income = (30,000 + 35,000 + 40,000 + 45,000 + 50,000 + 1,000,000) / 6 = $183,333.33

However, the income of $1,000,000 is an outlier and not representative of the typical income of this group. In this case, the median income would be a better measure of the typical income because it is not affected by extreme values.

The median income is calculated by arranging the incomes in order and selecting the middle value:

Median income = $40,000

Therefore, in this example, the median income of $40,000 is better measure than the mean income of $183,333.33 because it is not influenced by the extreme value of $1,000,000.

****
****

15) What is the Likelihood?

Answer: In statistics, likelihood refers to the probability of observing a set of data or sample, given a particular set of model parameters. The likelihood function is the function that describes this probability, and it is often used in maximum likelihood estimation.

To calculate the likelihood, we start with a statistical model that represents the underlying probability distribution of the data. This model includes one or more parameters, such as the mean or variance, that need to be estimated from the data.

The likelihood function is then defined as the probability of observing the data, given the values of the model parameters. It is a function of the model parameters and the observed data, and it can be written as:

Likelihood = P(data | model parameters)

The likelihood function is used in maximum likelihood estimation to find the values of the model parameters that maximize the probability of observing the data. This is done by calculating the likelihood function for different values of the parameters and selecting the values that give the highest likelihood.

It is important to note that the likelihood function is not a probability distribution itself. It is simply a function that describes the probability of observing the data, given a particular set of model parameters.
****
****