STATISTICS-- WORKSHEET- 6

ANSWERS

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

ANS:- 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

ANS:- a) Discrete

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

a) pdf

b) pmv

c) pmf

d) all of the mentioned

ANS:- a) pdf

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

a) mode

b) median

c) mean

d) bayesian inference

ANS:- c) mean

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

ANS:- c) empirical mean

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

ANS:- b) standard deviation

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

ANS:- c) 0 and 1

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

ANS:- b) bootstrap

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

a) frequency

b) summarized

c) raw

d) none of the mentioned

ANS:- b) summarized

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

ANS:- **Histograms and box plots** are graphical representations for the frequency of numeric data values. They aim to describe the data and explore the central tendency and variability before using advanced statistical analysis techniques.

Both histograms and box plots allow to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points.

Both histograms and box plots are used to explore and present the data in an easy and understandable manner.

Histograms are preferred to determine the underlying probability distribution of a data where as Box plots, on the other hand, are more useful when comparing

between several data sets. They are less detailed than histograms and take up less space.

In univariate case, box plot do provide some information that the histogram does not provide .That is , it typically provides the median ,25th and 75th percentile (min ,max)that is not an outlier and explicitly separates the points that are considered as outliers.

A histogram is preferable over a box plot is when there is very little variance among the observed frequencies.

11. How to select metrics?

ANS:-The metrics are chosen on terms of nature of the problem. Classification , regression and unsupervised learning all have different metrics. Also, based on the problem given, we decide if we want specificity or sensitivity also where and how the results would be applied in real world.

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

ANS:- **Statistical significance** can be assessed using hypothesis testing. The null hypothesis and alternate hypothesis would be stated first. Second, calculate the p-value, which is the likelihood of getting the test's observed findings .If the null hypothesis is true then Finally, we would select the threshold of significance (alpha) and reject the null hypothesis and if the p-value is smaller than the alpha — in other words, **the result is statistically significant**.

There are multiple test we can perform based on thee nature of problem and features of dataset.

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

ANS:-The kind of distribution that does not have any Gaussian distribution nor log normal are the skewed distribution , discrete distributions and binomial distribution.

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

ANS:- The **mean** of a dataset represents the average value of the dataset and the **median** represents the middle value of a dataset. Median is calculated by arranging all of the observations in a dataset from smallest to largest and then identifying the middle value.

There are many way to find out outliers, in those cases ,if we use mean these are drastically affected by outliers. It is best to use the median when the distribution is either skewed or there are outliers present.

When a distribution is skewed, the median does a better job of describing the center of the distribution than the mean.

15. What is the Likelihood?

ANS:-A likelihood function takes the data set as a given and represents the likeliness of the different parameters for your distribution. The likelihood function gives us an idea of how well the data summarizes these parameters.