STATISTICS

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

Answer: 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

Answer: 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

Answer: 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

Answer: 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

Answer: 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

Answer: a) variance

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

Answer: 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

Answer: 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

Answer: b) summarized

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

Answer: A histogram is a type of bar chart that graphically displays the frequencies of a data set. A histogram plots the frequency on the Y-axis and variable to be measured on the X-axis.

A boxplot is a chart that graphically represents the five most important descriptive values for a data set. These values include the minimum value, the first quartile, the median, the third quartile, and the maximum value.

A histogram is preferable over a box plot is when there is very little variance among the observed frequencies. A box plot allows to compare multiple data sets better than histograms as they are less detailed and take up less space.

11. How to select metrics?

Answer: The metrics used in Classification problem are:

Confusion matrix

Type I Error

Type II Error

Accuracy

Recall

Precision

Specificity

F1 Score

ROC Curve-AUC Score

PR Curve

Generally Accuracy, F1 score and ROC Curve-AUC Score are chosen best metrics for the classification problem.

The metrics used in Regression problem are:

Mean Squared Error

Root Mean Squared Error

Mean Absolute Error

R-Squared

Generally R-Squared is chosen best metrics for the regression problem.

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

Answer:

==> Creating a null hypothesis.

==> Creating an alternative hypothesis.

==> Determining the significance level.

==> Deciding on the type of test we use.

==> Performing a power analysis to find out the sample size.

==> Calculating the standard deviation.

==> Using the standard error formula.

==> Determining the t-score.

==> Finding the degrees of freedom.

==> Using a t-table.

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

Answer:

Distribution	Type Data	Examples
Lognormal	Continuous	Cycle or lead time data
Weibull	Continuous	Mean time-to-failure data, time to repair and material strength
Exponential	Continuous	Constant failure rate conditions of products
Poisson	Discrete	Number of events in a specific time period
Binomial	Discrete	Proportion or number of defectives

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

Answer: Mean is sensitive to outliers.

For example, we have the following data: 1,2,3,4,5

Mean = (1+2+3+4+5)/5 = 15/5 = 3

Median = Middle value = 3

Here the mean and median are same. When an outlier is added to the same data.

Data : 1,2,3,4,5,100

Mean = (1+2+3+4+5+100)/6 = 19.16

Median = Middle value = Mean of middle values =(3+4)/2 = 3.5

So, we can say that median is a better measure because median is not much effected to the added outlier.

15. What is the Likelihood?

Answer: Likelihood refers to finding the best distribution of the data given a particular value of some feature or some situation in the data. Likelihood function measures the goodness of fit of a statistical model to a sample of data for given values of the unknown parameters.