STATISTICS WORKSHEET-4

Q1to Q15 are descriptive types. Answer in brief.

1.What is central limit theorem and why is it important?
The CLT is a statistical theory that states that- if you take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from population will be roughly equal to the population.
It is important because it allows us to safely assume that the sample distribution of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assume a normal distribution.

2. What is sampling? How many sampling methods do you know?
Sampling is a method that allows us to get information about the population based on statistics from a subset of the population(sample), without having to investigate every individual.
There are two types of sampling methods: Probability sampling and Non-probability sampling.

3. What is the difference between type1 and typeII error?
A type-I error(false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population; a type-II error(false-negative) occurs if the investigator fails to reject a null hypothesis the is actually false in the population.

4. What do you understand by the term Normal distribution?
Normal Distribution/Gaussian Distribution is a continuous probability distribution. It has a bell-shaped curve that is symmetrical from mean point to both halves of the curve.

5. What is correlation and covariance in statistics?
Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency. Correlation is a statistical measure that indicates how strongly two variables are related. The value of covariance lies in the range -∞ and +∞.

6. Differentiate between univariate ,Biavariate,and multivariate analysis.
Univariant statistics summarize only one variable at a time.
Bivariant statistics compare two variables.

7. What do you understand by sensitivity and how would you calculate it?
Sensitivity is a measure of how well a machine learning model can detect positive instances. It is also known as the true positive rate(TPR) or recall. Sensitivity is used to evaluate model performance because it allows us to see how many positive instances the model was able to correctly identify.
Sensitivity = TP/(TP+FN)

8. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?
It is a process to draw inferences or some conclusion about the overall population or data by conducting some statistical tests on a sample. The same inferences are drawn for different machine learning models through T-test.
Null-Hypothesis(H0): It is regarding the assumption that there is no anomaly pattern or believing according to the assumption made.
Alternate-Hypothesis(H1): Contrary to null hypothesis, it shows that observation is the result of real effect.

9. What is quantitative data and qualitative data?
Quantitative data are measures of values or counts and are expressed as numbers.It is about numeric variables(e.g., how many;how much; or how often). Quanlitative data are measures of ‘types’ and may be represented by a name, symbol, or a number code.

10. How to calculate range and interquartile range?

IQR describes the middle 50% of values ordered from lowest to highest. To find the interquartile range (IQR), first find the median (middle value) of lower and upper half of the data. These values are quartile 1(Q1) and quartile 3(Q3). The IQR is difference between Q3 and Q1.

11. What do you understand by bell curve distribution?

A bell curve is a common type of distribution for a variable, also known as the normal distribution. The term “bell curve” originates from the fact the graph used to depict a normal distribution consists of a symmetrical bell-shaped curve.

12. Mention one method to find outliers.

One of the methods to find outliers is by using box-plot.

The boxplot contains the upper and lower quartiles, so the box basically spans the Inter-Quartile Range (IQR). One of the main reasons why box plots are used is to detect outliers in the data. Since the box plot spans the IQR, it detects the data points that lies outside this range. These data points are nothing but outliers.

13. What is p-value in hypothesis testing?

p-values are used in hypothesis testing to help decide whether to reject the null hypothesis or not.

14. What is the Binomial Probability Formula?

The binomial distribution is used to obtain the probability of success on a single trial denoted by p. The binomial distribution assumes that p is fixed for all trials.

15. Explain ANOVA and its applications.

Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.