Statistics

Worksheet 4

1) What is central limit theorem and why is it important?

Ans : The central limit theorem (CLT) says that as sample sizes grow higher, the distribution of sample means approaches a normal distribution, independent of the population's distribution. For the CLT to hold, sample sizes of 30 or more are frequently regarded as sufficient.

Central limit theorem is important as in the real world and practical datasets, the samples size is sufficiently larger, thus we can infer the distribution of the sample set will be a Gaussian/normal distribution.

2) What is sampling? How many sampling methods do you know?

Ans : Sampling is a process of selecting, manipulating and analyzing a subset out of the entire population. The samples are selected such that they represent the population. we analyze and find patterns data on the representative sample subset and validate those finding on the entire population using hypothesis testing

Sampling in statistical study is of two types i.e. – probability sampling and non-probability sampling.

Probability sampling includes simple random sampling, systematic, stratified and cluster sampling. Non-probabilistic sampling method includes Convenience and Judgmental.

3) What is the difference between type I and type II error?

Ans : Type 1 error, is the error caused by rejecting a null hypothesis when it is true. Type II error is the error that occurs when the null hypothesis is accepted when it is not true. Type 1 error is the conclusion for false positives while type 2 error concludes false negative.

4) What do you understand by the term Normal distribution?

Ans : Normal distribution is a type of probability density function in a shape of a bell curve. It is also known as a Gaussian distribution curve. For an ideal normally distributed data mean, median and the mode all lie on the same point, that is the peak of the bell curve. For a normally distributed feature 68.26% of the data lies in the 1st standard deviation, 95.44% of the data lies in the 2nd standard deviation area and 99.73% of data lies within 3 standard deviation of the feature.

5) What is correlation and covariance in statistics?

Ans : Covariance is a measure of how much two random variables vary together whereas Correlation is a statistical measure that indicates how strongly two variables are related.

Covariance involve the relationship between two variables or data sets and value lies between -infinity and +infinity and correlation involve the relationship between multiple variables and value lies between -1 and +1.

Covariance is a measure of the joint variability of two random variables. Correlation is obtained by dividing the covariance of the two variables by the product of their standard deviations.

6) Differentiate between univariate , Biavariate, and multivariate analysis ?

Ans : Univariate analysis is done using a single feature from the dataset, bivariate analysis is performed using 2 feature whereas multi-feature analysis is performed using more than 2 variables. Plots used for visualizing univariate analysis are count plots, histograms, density curves, distribution plots etc. Plots used for visualizing bivariate analysis are bar plots, scatter plots, joint plots, strip plots etc. Multivariate analysis plots are mode by adding hued data as a indication to the bivariate plots.

7) What do you understand by sensitivity and how would you calculate it?

Ans : Sensitivity informs us about the proportion of actual positive cases that have been predicted as positive by our model. It is also knows as the true positive rate. It is also known as recall.

8) What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?

Ans : Hypothesis testing is the process used to find the strength of evidence from the sample and provides a framework for making determinations related to the population. This sample is selected using one of the various sampling methods, probabilistic or non-probabilistic. H0 is the notation for null hypothesis whereas H1 is the notation for alternate hypothesis.

For a two tailed test, the null hypothesis (H0) should be rejected when the test value is in either of two critical regions on either side of the distribution of the test value and vice versa for alternate hypothesis.

9) What is quantitative data and qualitative data?

Ans : Quantitative data is one that can be counted, measured, and expressed using numbers. while Qualitative data is descriptive and conceptual. Qualitative data can be categorized based on traits and characteristics.

10) How to calculate range and interquartile range?

Ans : Range is calculated by : highest value – lowest value
IQR is calculated by: upper quartile (Q3) – lower quartile (Q1)

11) What do you understand by bell curve distribution ?

Ans : A bell curve distribution represents the normal/ Gaussian distribution.

12) Mention one method to find outliers ?

Ans : One of the best methods to find outliers is Z Score.

13) What is p-value in hypothesis testing?

Ans : The P-value method is used in Hypothesis Testing to check the significance of the given Null Hypothesis. Then, deciding to reject or support it is based upon the specified significance level or threshold.
A P-value is calculated in this method which is a test statistic. This statistic can give us the probability of finding a value (Sample Mean) that is as far away as the population mean.
The P in P-value stands for Probability.
Based on that probability and a significance level, we Reject or Fail to Reject the Null Hypothesis.
Generally, the lower the p-value, the higher the chances are for Rejecting the Null Hypothesis and vice versa.

14) What is the Binomial Probability Formula ?

Ans : The binomial probability formula can be used to calculate the probability of success for binomial distributions.

Binomial Probability formula: P(X) = (n! / (n-X)! X!) * (p)^X * (q)^n-X

Where X is the total number of successes.
p is the probability caused by success of an individual trail
q is the probability caused by failure of an individual train (q = 1-p)
n Is the number of trials .

15) Explain ANOVA and it's applications ?

Ans : Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.
There are two main types of ANOVA: one-way (or unidirectional) and two-way. There also variations of ANOVA.
Applications of ANOVA :

- Understanding the impact of different catalysts on chemical reaction rates
- Understanding the performance, quality or speed of manufacturing processes based on number of cells or steps they're divided into
- Comparing the gas mileage of different vehicles, or the same vehicle under different fuel types, or road types.