Statistics Worksheet-4

1. The CLT is a statistical theory that states that - if we take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from that population will be roughly equal to the population mean.

   Importance:
   - The CLT gives us a certain distribution over our estimations. We can utilize this to pose an inquiry about the probability of an estimate that we make.
   - The CLT performs a significant part in statistical inference. It depicts precisely how much an increase in sample size diminishes sampling error, which tells us about the precision or margin of error for estimates of statistics, for example, percentages, from samples.

2. Sampling is a method that allows us to get information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual

   2 types of sampling methods:

   - **Probability sampling method**

It uses randomization to select sample members. You know the probability of each potential member's inclusion in the sample.

- **Non-Probability sampling method**
  It uses non-random techniques (i.e. the judgment of the researcher). You can't calculate the odds of any particular item, person or thing being included in your sample.

3. Difference between type1 and typeII error :
   - Type I error refers to non-acceptance of hypothesis which ought to be accepted. While Type II error is the acceptance of hypothesis which ought to be rejected.
   - Type I error refers to False Positive while Type II error refers to False Negative.

4. Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

   In graphical form, the normal distribution appears as a "bell curve".

5. **Correlation:**
   Correlation is a statistical measure that indicates how strongly two variables are related.

   Covariance:
   Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency.

6. Difference between univariate ,Bivariate,and multivariate analysis :
   - Univariate statistics summarize only one variable at a time.
   - Bivariate statistics compare two variables.
   - Multivariate statistics compare more than two variables.

7. Sensitivity is the metric that evaluates a model's ability to predict true positives of each available category.

   Sensitivity = True Positives / (True Positives + False Negatives)

8. The process of hypothesis testing is to draw inferences or some conclusion about the overall population or data by conducting some statistical tests on a sample.

H0(Null Hypothesis) - The complement of the alternative hypothesis
H1(Alternative Hypothesis) - The hypothesis that we are interested in proving

When using a two-tailed test, regardless of the direction of the relationship you hypothesize, you are testing for the possibility of the relationship in both directions.
Our null hypothesis is that the mean is equal to x. A two-tailed test will test both if the mean is significantly greater than x and if the mean significantly less than x.

9. Qualitative data are descriptive and conceptual whereas, quantitative data is expressed in numbers that can be measured or counted.

10. Range = Maximum Value–Minimum Value

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

11. A bell curve is a graph depicting the normal distribution, which has a shape reminiscent of a bell. The top of the curve shows the mean, mode, and median of the data collected. Its standard deviation depicts the bell curve's relative width around the mean.

12. Z-score method for outlier detection:
    If the z score of a data point is more than 3, it indicates that the data point is quite different from the other data points. Such a data point can be an outlier.

13. In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct.
    The p-value serves as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected.
    A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

14. Binomial probability refers to the probability of exactly x successes on n repeated trials in an experiment which has two possible outcomes (commonly called a binomial experiment).

If the probability of success on an individual trial is p , then the binomial probability is **nCx · px · (1−p)n−x**

15. ANOVA (Analysis of Variance) helps us to complete our job of selecting the best features.

    Analysis of Variance is a statistical method, used to check the means of two or more groups that are significantly different from each other.

    Application of ANOVA :

    So in ANOVA, we will compare Between-group variability to Within-group variability. ANOVA uses F-test to check if there is any significant difference between the groups. If there is no significant difference between the groups that all variances are equal, the result of ANOVA's F-ratio will be close to 1.