# STATISTICS WORKSHEET-4 SOLUTIONS

1. The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed. Sufficiently large samples are term as sample set over 30 samples. Therefore, as a sample size increases, the sample mean and standard deviation will be closer in value to the population mean μ and standard deviation σ.
   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. Sampling is a process of selecting, manipulate and analyze a subset of the entire population. The samples are selected such that they represent the population. As a data scientist we analyze and find patterns data on the representative sample subset and validate those finding on the entire population using hypothesis testing. Sampling allows the data scientist to work on a smaller data set representing the entire population. Sampling methods can be classified into 2 types, probabilistic sampling methods and non-probabilistic types. Some of the sampling techniques I know are from probability sampling are simple random sampling, systematic, stratified and cluster sampling. From non-probabilistic sampling method, Convenience and Judgmental are known to me.

3. 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. Type 1 error is also called as the significance of the test. Type 2 error is also called as the beta error. As a result of type 1 error, then we might end up believing that the hypothesis works even when it doesn't. Whereas, as result of type 2 error, we might end up believing that the hypothesis works even when it doesn't.

4. 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 $1^{st}$ standard deviation, 95.44% of the data lies in the $2^{nd}$ standard deviation area and 99.73% of data lies within 3 standard deviation of the feature.

5. Covariance is a measure of the joint variability of two random variables. Correlation it is obtained by dividing the covariance of the two variables by the product of their standard deviations. The covariance values of the variable can lie anywhere between –inf to +inf whereas

the values of correlation are between -1 to +1. Also correlation is a unit-free measure whereas covariance is not a unit-free mesure.

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

8. Hypothesis testing is the process used to evaluate 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. Quantitative data can be counted, measured, and expressed using numbers. Qualitative data is descriptive and conceptual. Qualitative data can be categorized based on traits and characteristics.

10. Range is calculated by : highest value – lowest value

    IQR is calculated by: upper quartile (Q3) – lower quartile (Q1)

11. A bell curve distribution represents the normal/ Gaussian distribution.

12. Two of the many methods to find outliers are Z-score and IQR

13. The P value or calculated probability is the estimated probability of rejecting the null hypothesis (H0) of a study question when that hypothesis is true. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

14. 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. 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.